Optimizing Metagenome Analysis for Early Detection of Colorectal Cancer: Benchmarking Bioinformatics Approaches and Advancing Cross-Cohort Prediction

Yangyang Sun [1, $], Xiaoqian Fan [2, $], Shunyao Wu [1], Zhenming Wu [1], Wenjie Zhu [1], Hao Gao [1], Jieqi Xing [1], Jin Zhao [1], Xiaoquan Su [1, *]

[1] College of Computer Science and Technology, Qingdao University, Qingdao, Shandong, China, 266071

[2] Shouguang Hospital of Traditional Chinese Medicine, Weifang, Shandong, China, 262700

[$] Contributed equally, co-first authors

[*]Corresponding author, email: suxq@qdu.edu.cn

Abstract

Colorectal cancer (CRC) continues to be a major global public health challenge. Extensive research has underscored the critical role of the gut microbiome for predictive diagnostics of CRC. However, the variability in analytical methods and microbial features across previous studies has significantly impacted detection performance, hindering further research and application. In this study, we conducted a systematic analysis of over 2,000 gut metagenomic samples from 15 globally-sampled public and in-house cohorts. By benchmarking 1,080 analytical combinations across multiple analytical steps, we established an optimal bioinformatics workflow for metagenome-based CRC detection. Despite the substantial heterogeneity of gut microbiomes across multiple regions and cohorts, our workflow demonstrated the robust performance, achieving an AUROC of 0.83 by identifying consistent microbial dynamic patterns associated with CRC. However, early-stage prediction of CRC, particularly at the precancerous adenomas (ADA) stage, remains challenging due to the instability of microbial signatures across cohorts. To address this, we developed an instance-based transfer learning strategy, Meta-iTL, which improved ADA detection in a newly recruited cohort, increasing the AUROC from 0.58 to 0.77 using models trained on existing cohorts. This study not only provides a comprehensive bioinformatics guild for metagenomic data processing and modeling, but also advances the development and application of non-invasive approaches for the early screening and prevention of CRC.

Keywords: Gut metagenome; Colorectal cancer; Adenomas; Machine learning; Transfer learning.

Introduction

Colorectal cancer (CRC) is one of the most prevalent cancers globally and remains a major cause

of cancer-related mortality. According to the American Cancer Society, approximately 106,590 new CRC cases and 53,010 deaths are expected in 2024 alone [1]. A large number of CRC cases are diagnosed at advanced stages, resulting in poor prognoses and contributing significantly to high mortality rates. Early screening is therefore critical for improving patient outcomes [2, 3]. Traditional screening methods, such as colonoscopy [4], are effective for early detection and can reduce CRC incidence and mortality. However, challenges associated with convenience, risks, and the high costs of invasive procedures remain pressing concerns [4-8]. Consequently, researchers have been actively exploring alternative, non-invasive diagnostic approaches.

As a non-invasive stool-based approach, gut microbiome holds significant potential and has become widely utilized in the detection and study of related diseases [9-11]. Studies have demonstrated that the gut microbiome plays a pivotal role in the occurrence and progression of CRC [12], with microbial markers [13-15] identified for predicting disease susceptibility [16], progression [17], prognosis [18, 19], and treatment response [20, 21]. Gut microbiome analysis typically relies on metagenomic sequencing and bioinformatics workflows that include *i.* sequence profiling, *ii.* confounding factor analysis, *iii.* biomarker screening, and *iv.* predictive model construction [22-24]. However, as datasets grow larger and analytical complexities increase, significant variability in results arises depending on the tools and parameters used [24-26]. For instance, the dynamic patterns of microbial signatures often differ across cohorts, feature types, marker selection methods, and model training approaches, which hinder the reproducibility and practical application of microbiome-based screening [27-29]. These limitations underscore the urgent need for systematic, large-scale analyses of gut microbiomes across multiple cohorts and the development of optimized bioinformatics protocols for CRC detection.

On the other hand, it is estimated that 80% of CRC cases arise following the adenoma-carcinoma sequence, which involves the progressive accumulation of mutations in a period of 10 -15 years on average [30, 31]. The early screening of CRC at precancerous-stage adenoma (ADA) has brought the 5-year relative survival rate to around 90% [32], significantly facilitating early clinical decision making, alleviating the incidence of CRC, and reducing economic burden [2, 3]. However, research on the association between the microbiome and ADA remains limited. Few studies have examined microbial changes in ADA samples [3, 13, 22, 33, 34] , and most only focused on single cohorts or specific populations with small sample sizes, restricting the generalizability of microbiome-ADA linkage on a broader population [22, 34]. For example, Thomas et al. [22] conducted a meta-analysis of metagenomic data for ADA samples but reported poor performance in distinguishing ADA from healthy controls (e.g., area under the receiver operating characteristic curve (AUROC) < 0.6). Thus, accurate detection of ADA across diverse cohorts remains a major challenge.

In this study, we performed a systematic and comprehensive analysis of over 2,000 gut

metagenomes from diverse regions worldwide, aiming to quantify the impact of various factors and analytical methods. Although substantial regional and cohort-based diversity in gut microbiomes, we revealed consistent dynamic patterns in key microbial features under CRC. By benchmarking 1,080 testing combinations of multiple analytical steps across 15 different cohorts, we established an optimized bioinformatics workflow for CRC identification, encompassing procedures from microbiome profiling to predictive modeling. This workflow achieved a mean AUROC of 0.83 in multi-cohort cross-validation, inter-cohort validation and leave-one-out-of-the-dataset (LODO) validation, effectively mitigating the effects of technical and geographical variations. Its robustness and generalizability were further confirmed by two independent cohorts with mean AUROC up to 0.82.

Despite these advancements, challenges remain in the cross-cohort prediction of CRC risk at the precancerous ADA stage due to cohort-specific variability. Here we developed Meta-iTL, an instance-based transfer learning (TL) modeling strategy. By leveraging knowledge transfer between samples to be tested (i.e., target domain) and existing data repository (i.e., source domain), Meta-iTL overcomes challenges limited by the cohort-specific effects and scarcity of ADA cases for training, improving cross-cohort applicability of data models. Notably, when applied to a newly produced cohort, Meta-iTL raised the AUROC from 0.58 to 0.77 by models trained from existing datasets. In summary, our study provides a reproducible bioinformatics framework for cross-cohort CRC and ADA detection. This approach advances the development and application of non-invasive early screening, offering significant potential for improving early diagnosis and prevention of CRC.

Results

Overview of CRC characterization using multi-cohort gut metagenomes

To address the challenges posed by data heterogeneity and bioinformatic variability in linking the gut microbiome to CRC, we conducted a systematic and comprehensive analysis of 2,053 human gut paired-end sequenced metagenomes (**Fig.1a**; **Table 1**). These samples were derived from 14 published cohorts and one in-house recruited cohort (i.e., CHN_WF in **Table 1**; refer to **Materials and methods** for details), spanning three health statuses: 735 samples from CRC patients, 349 from individuals with ADA, and 969 from healthy controls (CTR). Samples exhibited high geographical diversity that surveyed from ten countries/regions and were generated using different sampling procedures, storage conditions, and DNA extraction protocols [35, 36] (**Table S1**). Importantly, all samples were collected prior to any therapeutic interventions to avoid treatment-related biases [21, 37, 38]. Raw metagenomic sequences were filtered by quality control and host removal, and then taxonomical and functional profiles were parsed by MetaPhlan4 [39] and HUMAnN3 [40], respectively, with consistent parameters setting across all datasets (**Fig. 1b**; refer to **Materials and methods** for details).

For our exploratory analysis of CRC, we designated 10 cohorts as the 'discovery' dataset (**Table 1a**), focusing on distinguishing samples with CRC from healthy controls. To comprehensively evaluate the impact of analytical options, we assessed the effects of five confounding factors, eight types of annotations, and three preprocessing strategies. Next, we compared seven kinds of widely used feature selection methods to identify microbial biomarkers that are most informative for CRC detection. These signatures were then used to evaluate the performance of five machine learning models, allowing us to establish an optimized bioinformatics workflow for CRC detection (**Fig. 1c**). To validate the robustness and applicability of the proposed workflow, we performed both within-cohort cross-validation and cross-cohort validation within the discovery dataset. Additionally, the detection model was applied to an independent 'validation' dataset (**Table 1b**) comprising two external cohorts.

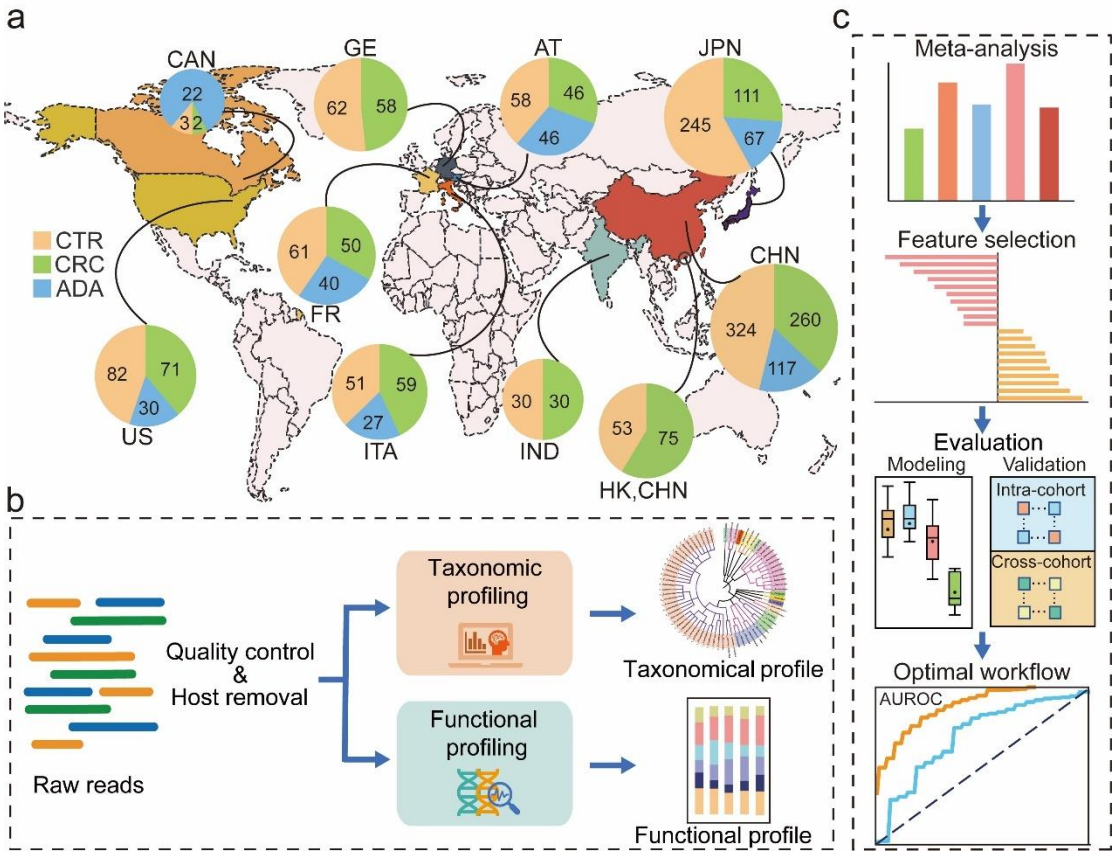

Fig. 1. Workflow for cross-cohort integrative analysis of gut metagenomics for CRC detection. (a) Overview of sample collection and distribution across cohorts. The world map was generated by the Python GeoPandas package (version 0.14.0) [41]. (b) Unified preprocessing and profiling of raw metagenomic sequences to ensure data consistency. (c) Analytical framework encompassing effect size estimation of contributing factors, feature selection, model construction, and performance evaluation.

Table 1. Metagenome cohorts and samples.

(a) CRC discovery dataset.

Cohort	# of CTR	# of CRC	# of ADA	Country/Region	Source
AT [13]	58	46	46	Austria	PRJEB7774
IND [42]	30	30	0	India	PRJNA531273 PRJNA397112
ITA [22]	51	59	27	Italy	SRP136711
JPN [15]	245	111	67	Japan	DRA008156
GE [35]	57	20	0	Germany	PRJEB27928
CHN_HK [43]	53	75	0	HK, China	PRJEB10878
CHN_SH [44]	85	80	0	China	PRJNA731589
FR [45]	66	88	40	France & Germany	ERP005534
US [46]	48	46	0	USA	PRJEB12449
CHN_WF (in-house produced)	48	48	54	China	PRJNA1198725

(b) CRC validation dataset

Cohort	# of CTR	# of CRC	# of ADA	Country/Region	Source
CHN_SH-2 [47]	44	32	23	China	PRJNA514108
CHN_SH-3 [48]	100	100	0	China	PRJNA763023

(c) Additional ADA dataset

Cohort	# of CTR	# of CRC	# of ADA	Country/Region	Source
US-2 [49]	28	0	26	USA & Canada	PRJNA389927
US-3 [50]	9	0	26	USA	PRJNA745329
CHN_SH-4 [51]	47	0	40	China	PRJNA706060

Impact of confounding factors on microbial composition and mitigation of cohort-related variability

Given the substantial biological and technical differences [35, 52] among 10 cohorts of the discovery dataset, we investigated the effects of confounding factors on microbial composition and distribution, including cohort-specific variables, BMI (body mass index), age, gender, and health status. By measuring the overall beta diversity across eight feature types (taxonomic profiles on class, order, family, genus, species, and SGB (species-level genome bins) [39] levels, functional profiles on KO（KEGG Orthology） [53] and uniref family [54] levels, we found that the cohort-related factor exhibited the largest effect size than others (**Fig. 2a** and **Fig. 2b**; **Fig. S1** and **Fig. S2**). Variance of each microbial component further validated that both cohort and disease status significantly influenced the microbial communities (**Fig. S3**).

To explore the influence of technical batch effects, we compared metagenomes sampled from the same city of China conducted by three studies (CHN_SH, CHN_SH-2, CHN_SH-3; **Table 1**). Despite sharing similar geographic origin and health status, significant differences in beta diversity persisted across these cohorts (PERMANOVA test, p-value < 0.01; **Fig. S4**). This indicates that technical batch effects [36, 55], in addition to regional differences [27], contributed substantially to microbial variability. Hence, we applied batch correction [56] and logarithmic transformation to the relative abundance of profiles (refer to **Materials and methods** for details), respectively. With batch correction, the influence of cohort factors was significantly reduced, which became lower than that of disease (**Fig. 2b**). However, batch correction may inadvertently introduce leakage of healthy status information, potentially impacting downstream analysis [57]. On the other hand, logarithmic transformation, while less effective at eliminating cohort effects, improved disease status discrimination as annotation resolution was refined (**Fig. 2b**; **Fig. S5**). This adjustment balanced the feasibility of detecting CRC across cohorts with the preservation of biological relevance.

Using the three preprocessing strategies (i.e., original, batch-corrected, and log-transformed profiles), we identified microbial features significantly associated with CRC as candidate biomarkers using MaAsLin2 (**Table 2**; refer to **Materials and methods** for details) [58]. Notably, as the profiling resolution increased (e.g., taxonomy annotation from class level to SGB level), the influence of disease status on biomarkers became more pronounced, surpassing the effect of cohort-related factors. Specifically, at the genus, species, and SGB levels, the PERMANOVA F-values for disease status grouping exceeded those of cohort effects (**Fig. 2c**). This finding highlights the importance of selecting high-resolution microbial signatures to minimize cohort and batch variability, thereby focusing on variations directly linked to CRC status across multiple cohorts. Building on this foundation, subsequent analyses evaluated and refined biomarker screening strategies to enhance the robustness of CRC detection across diverse datasets.

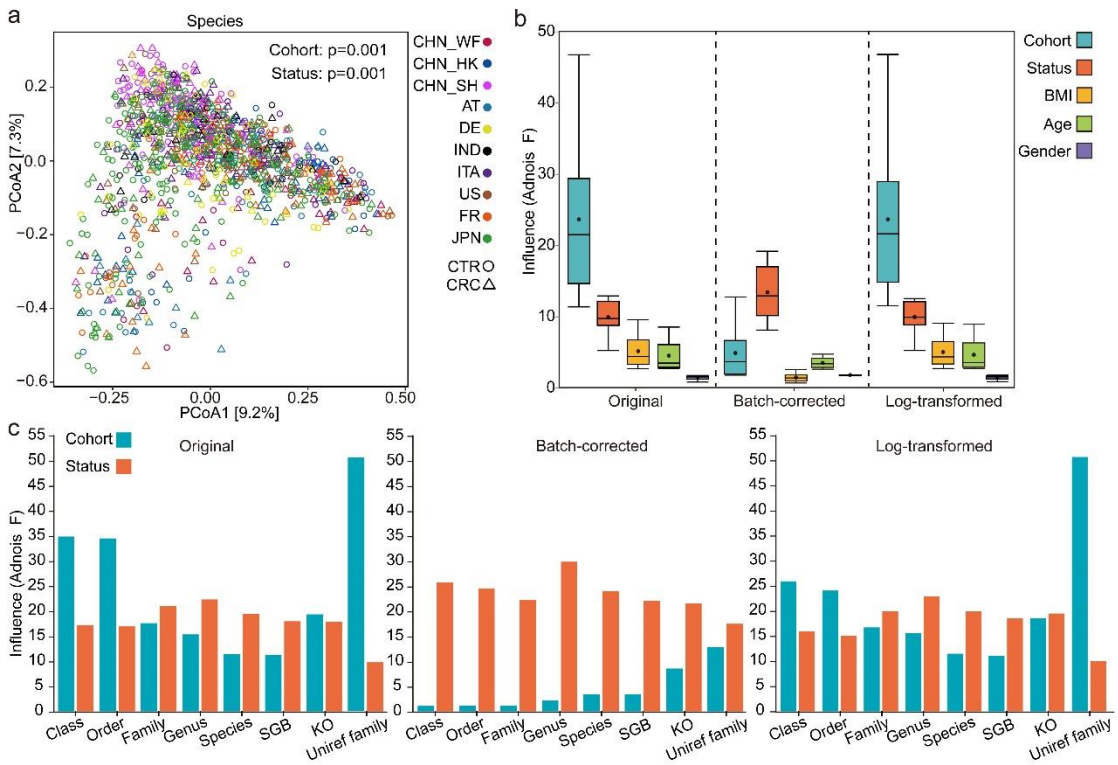

Fig. 2. Impacts of different confounding factors on gut microbial composition. (a) Beta diversity patterns of CRC and CTR samples in the discovery dataset, visualized using Bray-Curtis distances. Statistical significance was assessed by the PERMANOVA test for beta diversity with corresponding *p*-values reported. (b) Effect size dynamics of confounding factors across three preprocessing strategies applied to microbial profiles. Each boxplot represents PERMANOVA *F*-values derived from microbial features across eight feature types. (c) Detailed influences of cohort and health status on eight types of microbial features. Source data is available in the Dataset S1.

A synergistic feature selection strategy reveals consistent CRC-associated microbial biomarkers

Since selected microbial species demonstrated promising potential distinguishing healthy status across multiple cohorts (**Fig. 2c**), we compared seven feature selection tools using species-level microbiome profiles of the discovery dataset (refer to **Materials and methods** for details of tools). Guided by the principle that methods yielding feature sets with greater overlap are likely to provide more reliable results [25], we first evaluated the consistency of CRC biomarkers selected by each tool. Among these, ANCOM-II was the most conservative, with 62% of its selected features overlapping with those identified by other methods. In contrast, RFECV detected the largest number of unique features, at 33%. Then based on the candidate biomarkers screened by each method, we employed XGBoost modeling (refer to **Materials and methods** for details about the model) with five-fold cross-validation to assess CRC detection performance on the discovery dataset. LEfSe achieved the highest AUROC, followed by RFECV and Wilcoxon test (**Fig. 3a**).

We also repeated the above test on each individual cohort of the ten discovery cohorts respectively. As shown in **Fig. 3b,** features selected by different tools varied widely in their distribution within single cohorts, measured by normalized proportion (NP). RFECV (mean NP: 38.36%) and LEfSe (mean NP: 11.44%) parsed out the highest rate, while ANCOM-II (mean NP: 2.4%), T-test (mean NP: 3.64%) and MaAsLin (mean NP: 4.05%) obtained the lowest NP (**Fig. 3b**). Specifically, ANCOM-II, T-test and metagenomeSeq identified few markers in at least two cohorts. Such performance of these tools within individual cohorts aligns with their behavior when applied to the combined multi-cohort dataset (**Fig. 3a)**. The XGBoost-based cross-validation (**Fig. 3c**) demonstrated that microbial markers identified by LEfSe significantly improved detection performance, exceeding the baseline accuracy of unfiltered profiles in almost all cohorts. Similarly, MaAsLin2 and Wilcoxon test also contributed largely to the accuracy of the XGBoost model. Although RFECV identified the largest number of candidate markers, its predictive performance was suboptimal likely due to the inclusion of false-positive features. These trends remained consistent when applied to profiles preprocessed with logarithmic transformation and batch correction (**Fig. S6** and **Fig. S7**).

Considering the varying characteristics of the feature selection methods, here we proposed a synergistic strategy to optimize biomarker screening while minimizing the impact of cohort-related variability. Firstly, LEfSe, ANCOME-II, and MaAsLin2 were applied to each single cohort and the all-merged data to identify a union set of candidate markers associated with CRC. In this set, we observed that 72% candidates were shared by at least six individual cohorts (**Fig. S8**). Then this set was refined by the Max-Relevance and Min-Redundancy (mRMR) method [59] alongside an iterative feature elimination (IFE) [52] to finalize the optimal biomarkers for CRC (**Fig. 3d**). Among the refined biomarkers of the ten discovery cohorts, 78% had the same dynamics on relative abundance associated with CRC in more than six cohorts (**Fig. 3e**). These findings demonstrated that the microbial signatures identified through the synergistic approach show high consistency in their dynamic changes across health states in multiple heterogeneous cohorts, providing a solid foundation for the development of a reliable, cross-cohort CRC detection model.

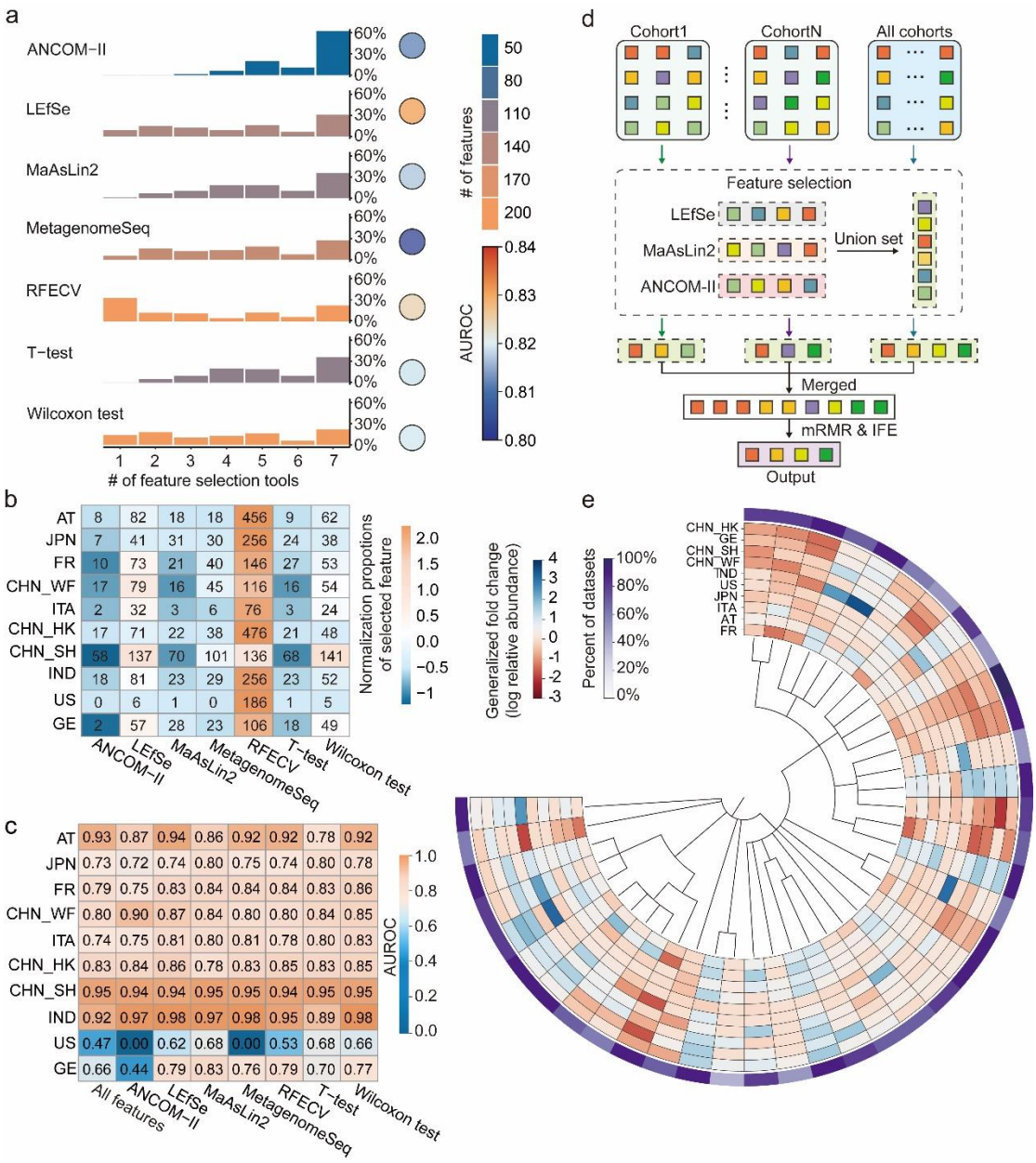

Fig. 3. Comparison of different feature selection tools. (a) Performance of different tools on all cohorts of the discovery dataset. For each method, the color of the bars indicates the number of identified features shared with others, and the height represents their proportion within a method. The colors of the dot chart illustrated the AUROC of the XGBoost-based cross-validation using features selected by each method. (b) Feature selection of different methods on each single cohort of the discovery dataset. Values are the number of identified microbial markers, and colors are their normalized proportion (NP) by scaled mean-centralization in each single cohort (each row). (c) Performance of XGBoost-based cross-validation in each single cohort. (d) Diagram of the synergistic feature selection strategy with mRMR and IFE. (e) The dynamic trends of CRC markers identified across all cohorts of the discovery dataset using the synergistic feature screening method. The heat map displays the generalized fold change (gFC) [35] values of these markers between cohorts, with red indicating enrichment in colorectal cancer (CRC) and blue indicating depletion in CRC. The outermost purple ring shows the proportion of cohorts exhibiting consistent directional changes for each microbial marker. Source data is available in the Dataset S1.

Optimal workflow for CRC detection across multiple cohorts

To quantify the impact of various analytical steps on CRC detection, we systematically compared three preprocessing strategies (original, log-transformed, and batch-corrected profiles), nine feature screening methods (LEfSe, MaAsLin2, ANCOM-II, T-test, metagenomeSeq, two-sided Wilcoxon test, RFECV, synergistic approach, and non-feature selection), and five commonly used machine learning/deep learning algorithms (supporting vector machine (SVM), multilayer perceptron (MLP), k-nearest neighbor (KNN), random forest (RF), and XGBoost; refer to **Materials and methods** for details about models). CRC data models were constructed using eight feature types, including taxonomic profiles at the class, order, family, genus, species, and SGB levels and functional profiles at the KO and uniref family levels. This comprehensive analysis across ten discovery cohorts generated a total of 1,080 testing combinations. For each combination, the discovery dataset was randomly divided into training (70%) and test (30%) sets with 100 times repeat. Performance was assessed using AUROC, F1-score, AUPRC, and MCC, with AUROC selected as the primary metric due to its strong correlation with other measures (**Fig. S9**).

Among the variables tested, the choice of machine learning algorithm had the most significant impact on CRC detection performance, followed by the type of features used (two-sided Kruskal test, p-value < 0.01; **Fig. 4a**). However, no significant differences were observed in AUROC values among the three preprocessing strategies (two-sided Kruskal test, p-value = 0.4, **Fig. 4b**; **Fig. S10**). Considering batch correction may introduce label leakage [57], while logarithmic transformation reduced batch-related variability (**Fig. 2c**) with slightly better CRC detection effect than other pre-processed data (**Fig. 4b**; **Fig. S10**), subsequent analyses were conducted based on log-transformed profiles.

Taxonomy profiles at the genus, species, and SGB levels outperformed other annotation types, yielding significantly higher AUROC values (two-sided Wilcoxon test, p-value < 0.01; **Fig. 4c**; **Fig. S11a**), in agreement with confounder analysis results (**Fig. 2c**). For feature screening, biomarkers identified using the synergistic method reported slightly better AUROC values compared to individual tools (**Fig. 4c**; **Fig. S11b**). Among machine learning models, XGBoost and RF significantly outperformed others in CRC prediction (two-sided Wilcoxon test, p-value < 0.01; **Fig. 4c**; **Fig. S11c**), likely due to their capacity for handling high-dimensional, complex microbial data by non-linear modeling and noise resistance.

Based on these findings, we propose an optimized bioinformatics workflow for microbiome-based CRC modeling (**Fig. 4d**):

i. Preprocessing. Annotate microbiome composition at the species level (with genus and SGB levels as optional alternatives) and apply log-transformation to relative abundance data to reduce skewness.

ii. Feature selection. Employ the synergistic feature approach (**Fig. 3d**) to select disease-associated biomarkers.

iii. Model construction. Construct and adjust the model using RF (XGBoost as an alternative; also refer to **Materials and methods** for details about the modeling).

This workflow demonstrated high performance, with five-fold cross-validation across the ten discovery cohorts achieving AUROC values of 0.86 by RF at the species level (0.85 for genus and SGB levels; **Fig. S12**).

Using biomarkers and profiles screened from all ten discovery cohorts, we further performed inter-cohort validation (training on a single cohort and testing on others) and leave-one-out-of-dataset validation (LODO; make a cohort the testing set and others the training set). Despite the limited sample size in individual cohorts, the average AUROC for inter-cohort validation using the RF model remained at a high level (**Fig. 4e**). We noticed that tests on the US cohort exhibited poor performance, mainly due to the fecal samples in this cohort were freeze-dried prior to DNA extraction and sequencing and stored at -80°C for over 25 years [46]. After excluding the US cohort as an outlier, the average AUROC of the inter-cohort validation reached 0.79, at the species levels, respectively. LODO validation further improved the AUROC values to 0.85 at the species level (**Fig. 4e**), emphasizing the importance of integrating multi-cohort data for enhanced performance.

The robustness of the workflow across multiple cohorts were further evaluated using two independent validation cohorts from China (CHN_SH-2 and CHN_SH-3; **Table 1b**). The CHN_SH-2 cohort comprised 44 CTR samples and 32 CRC patients, while the CHN-SH-3 cohort included 100 CTR and 100 CRC samples. Using species-level RF model trained from the ten discovery cohorts, the proposed workflow achieved AUROC values of 0.86 and 0.77 for the CHN_SH-2 and CHN_SH-3 cohorts, respectively (**Fig. 4f**). These results validated the universality of the proposed workflow across multiple cohorts, underscoring the efficacy of using intestinal microbial signatures for accurate and scalable CRC screening.

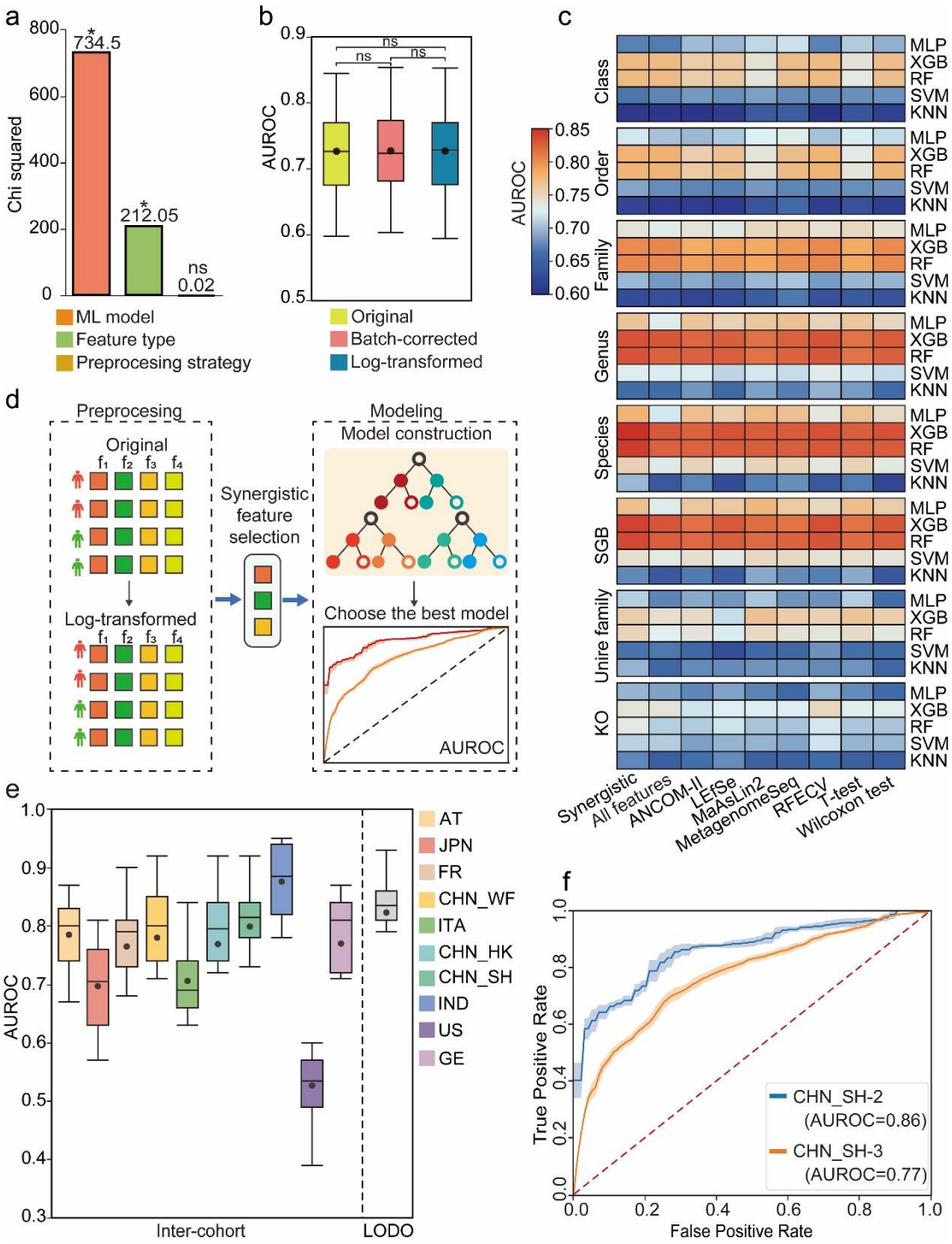

Fig. 4. Comprehensive benchmarking of the performance for CRC detection workflow. (a) Impact of factors on AUROC depicted by chi-square statistics from two-sided Kruskal test. (b) Difference of AUROC among preprocessing strategies measured by two-sided Wilcoxon test. (c) The heatmap displays 360 AUROC results by five machine learning, nine feature selection methods and eight feature types using logarithmically transformed profiles. Results of all 1,080 test combinations are available in the Dataset S1. (d) The scheme of the optimal bioinformatics workflow for CRC detection modeling. (e) Results of the inter-cohort validation and LODO using profiling on species levels. Each box of the inter-cohort validation summarizes the AUROC values of a single cohort tested by models trained from other

individual cohorts separately. The LODO box summarizes the AUROC value of a single cohort tested by models trained from all other cohorts. (f) ROCs of two independent validation cohorts using the model derived from the discovery dataset. Source data is available in the Dataset S1.

Transfer learning enables cross-cohort CRC prediction at precancerous stage

Following the optimized bioinformatics workflow, we performed five-fold cross-validation for precancerous adenoma (ADA) detection using random forests on nine cohorts containing ADA samples in all datasets (i.e., ADA vs. CTR; **Table 1**). Surprisingly, the AUROC for ADA detection was only 0.70 (**Fig. 5a**), which was significantly lower than the AUROC achieved for CRC detection (0.86). When tested on the in-house CHN_WF cohort, a model trained on all other cohorts yielded an even lower AUROC of 0.58 (**Fig. 5b**). Further analysis revealed that only 50% ADA-associated signatures exhibited the same dynamic patterns among cohorts (**Fig. 5c**). This high heterogeneity undermines the effectiveness of cross-cohort models, making within-cohort models the most reliable approach for ADA detection. However, such a strategy can also be constrained by the limited sample sizes within individual cohorts. In such cases, adopting well-validated models from other locations becomes a more practical option. However, the applicability and compatibility of the model need to be taken into consideration.

Here we propose Meta-iTL, an instance-based transfer learning (TL) approach designed to enhance cross-cohort ADA detection (**Fig. 5d**) Basically, Meta-iTL operates by randomly extracting a small subset of samples (e.g., 30%) from the target domain to capture its specific characteristics, and integrating this information into the source domain of a much larger dataset for model training (refer to **Materials and methods** for details). This approach effectively addresses challenges associated with inconsistent shifts in ADA-associated markers across cohorts and limited data availability within single cohorts. Using the CHN_WF cohort as the target domain and all other ADA cohorts as the source domain, we evaluated Meta-iTL for ADA detection. The proportion of CHN_WF samples for TL ranged from 20% to 60%, and 30% samples were randomly selected from the remaining for testing. Each iteration was repeated 50 times to minimize random bias, and both TL-based and non-TL models shared the exact same testing set. Additionally, we performed five-fold cross-validation within the CHN_WF cohort to make a baseline AUROC of 0.70.

As expected, the source domain model without TL performed poorly in detecting ADA within the CHN_WF cohort that was far below the baseline (AUROC=0.57; **Fig. 5e**). When only 20% of the CHN_WF cohort was used for TL, the model's accuracy did not improve due to insufficient target cohort information. However, as the proportion of fused CHN_WF samples increased to 30%, the TL model outperformed the non-TL model. Performance peaked at 40% target domain integration, achieving an AUROC of 0.77—surpassing the baseline AUROC of the within-cohort model.

Interestingly, AUROC declined with further increases in target domain sample proportions (e.g., 50% and 60%), likely due to overfitting or reduced influence of source domain data.

We further validated Meta-iTL by designating the CHN_SH-4 (**Table 1c**) cohort as the target domain. With 30% of target domain data, Meta-iTL achieved an average AUROC of 0.87 (**Fig. S13a**), outperforming the source domain model without TL. A Jensen-Shannon (JS) divergence analysis provided insights into Meta-iTL's efficacy (refer to **Materials and methods** for details). For ADA metagenomes, significant JS divergence was observed between the target and source domains, reflecting substantial differences in their data distributions. After applying Meta-iTL, the JS divergence was notably reduced, indicating that Meta-iTL successfully mitigates gaps caused by cohort-specific factors (**Fig. 5f)**. Therefore, by integrating a small subset of targeted samples with large-scale data, Meta-iTL holds great promise for the early detection and prevention of CRC by addressing the variability inherent in multi-cohort ADA microbiome datasets.

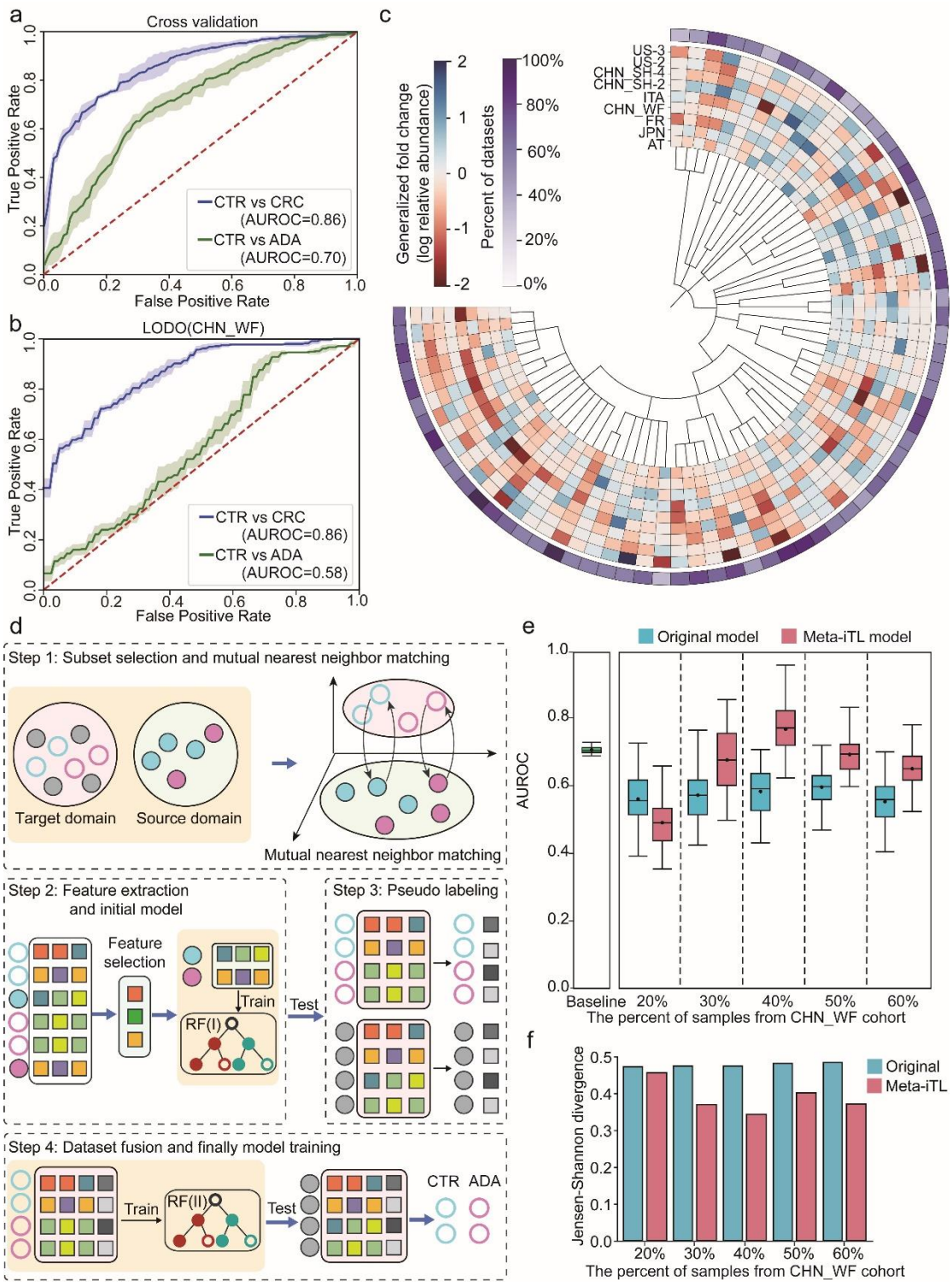

Fig. 5. Meta-iTL significantly improves the performance for ADA detection across cohorts. (a) ROC curves of CRC and ADA detection on multiple cohorts by five-fold cross validation. (b) ROC curves of CRC and ADA detection on CHN_WF cohort by other cohorts. (c) Dynamics trend of ADA markers identified from all cohorts of the discovery dataset by the synergistic feature screening method. (d) Diagram of the Meta-iTL for instance-based transfer learning. Refer to **Materials and methods** for details of each step. (e) Comparison of performance of ADA detection for target domain (CHN_WF cohort) using model trained by source domain (all other cohorts) before and after TL. The baseline is a

five-fold cross-validation within the target domain. (f) Jensen-Shannon (JS) values between the data before and after TL and the target domain. Source data is available in the Dataset S1.

Discussions

In this study, we systematically benchmarked bioinformatics approaches for metagenome-based colorectal cancer (CRC) detection using one of the largest and most diverse datasets of gut metagenomic samples, encompassing 2,053 samples from 15 cohorts across ten countries and regions. By evaluating 1,080 testing combinations of preprocessing steps, feature selection strategies, and machine learning algorithms, we provide actionable insights into how these analytical choices impact model performance. These methods were instrumental in mitigating biases introduced by confounding factors, such as sequencing batch effects, underscoring their importance in multi-cohort microbiome studies.

Despite the geographical and technical diversity inherent to metagenomic data across datasets, our synergistic feature selection strategy identified highly consistent dynamic patterns of gut microbial shifts associated with CRC. These findings enabled us to establish a robust workflow capable of achieving high predictive performance across diverse cohorts, highlighting its translational potential for clinical applications. By integrating data from multiple cohorts, our approach not only enhances generalizability but also ensures that identified biomarkers and models are resilient to data variability. This reproducible framework is readily extendable to other diseases, thereby maximizing the utility of existing large-scale microbiome datasets for diagnostic research and development.

However, substantial heterogeneity in microbial composition across cohorts remains a significant challenge, particularly for early-stage detection. This variability was especially pronounced for adenoma (ADA)-associated biomarkers, which exhibited greater cohort-specific effects compared to CRC biomarkers. Such inconsistencies hindered the accuracy of cross-cohort ADA detection models, posing a critical bottleneck in early CRC screening efforts. To address this, we developed Meta-iTL, an instance-based transfer learning (TL) approach that incorporates a small subset of target domain samples into models trained on large-scale source domain data. By effectively mitigating biases introduced by cohort-specific variability and addressing the scarcity of ADA cases, Meta-iTL significantly improved ADA detection accuracy, offering a practical and scalable solution for enhancing microbiome-based early screening tools in real-world settings.

Despite these advancements, certain limitations remain. First, as a multi-cohort study, it is challenging to completely eliminate biases introduced by sampling methodologies, experimental protocols, sequencing techniques, and other technical factors. Although we employed measures to minimize these effects, incomplete metadata for some samples and unknown confounding variables—such as diet, lifestyle, and host genetic factors—may still influence our findings. Further

investigation is needed to systematically evaluate and address these potential sources of variability. Second, while Meta-iTL demonstrated substantial gains in cross-cohort ADA detection, its performance was sensitive to the proportion of target domain data used for transfer learning. Excessive reliance on target domain data, as observed in our experiments, can lead to overfitting and diminish the contribution of source domain information. Future research should explore hybrid transfer learning frameworks that balance the integration of target and source domain data to optimize model performance further.

In conclusion, this study establishes a comprehensive roadmap for metagenome-based CRC early detection and prevention. By providing a standardized bioinformatics workflow and introducing advanced methodologies like Meta-iTL, we aim to bridge the gap between microbiome research and its clinical translation. These findings lay the foundation for developing non-invasive, scalable, and accurate microbiome-based screening tools for CRC and its precancerous stages. As microbiome research continues to advance, our study underscores the critical importance of harmonizing analytical approaches and leveraging innovative machine learning techniques to fully realize the diagnostic potential of the gut microbiome.

Materials and methods

Public dataset collection

We used PubMed to search for studies that published fecal shotgun metagenomic data of human colorectal cancer (CRC) patients, adenoma (ADA) patients, and healthy controls (CTR). Raw FASTQ files were downloaded using the SRA toolkit (V.2.9.1) from Sequence Read Archive (SRA) and European Nucleotide Archive (ENA) using identifiers listed in **Table 1**. All metadata for metagenomics were manually curated from the published papers (**Table S1)**.

In-house fecal sample collection and shotgun metagenomic sequencing

Subject recruitment and sample collection. The CHN_WF cohort (**Table 1a**) is an in-house produced dataset. Stool samples were collected from *n*=150 subjects at Shouguang Traditional Chinese Medicine Hospital, Shandong, China. All participants provided written informed consent. Based on pathology and colonoscopy results, subjects were classified into three groups: *i.* healthy subjects with negative colonoscopy results for tumors, adenomas, or other diseases (CTR; *n*=48); *ii.* subjects with colorectal adenoma (ADA; *n*=54); and *iii.* patients with newly diagnosed colorectal cancer (CRC; *n*=48). The clinical characteristics of the study participants are shown in Supplementary Data1. Fecal samples were collected using sterile specimen collectors and stored at -80°C prior to microbial analysis. This study was approved by the Medical Ethics Committee of Shouguang Traditional Chinese Medicine Hospital (approval no. szy2021-01).

Metagenomic sequencing. The DNA from stool samples was extracted using MagPure Stool DNA KF Kit B (MAGEN, Guangzhou, China) according to manual instructions. During the sequencing process, we utilized a short-insert library type and conducted sequencing using the DNBSEQ-T7 platform. The read length was configured to PE150 (paired-end, 150 base pairs each) to ensure comprehensive coverage and high resolution of genomic features. The quality of the raw fastq files was assessed using a Phred+64 encoding system, facilitating accurate measurement of base call quality across the dataset. Raw data were filtered using SOAPnuke [60], which involved removing contaminants, adapters, and low-quality sequences. The specific filtering parameters set in SOAPnuke were: "-n 0.01 -l 20 -q 0.4 --adaMis 3 --outQualSys 1 --minReadLen 150". A total number of approximately 1800 GB of high-quality clean sequence data was obtained.

Microbiome profiling

Sequence preprocessing and taxonomic and functional profiling. For all metagenomes, low-quality reads were removed using the Fastp [61] tool with parameters "-W 30 -M 20 -y". The pre-filtered reads were then aligned to the human genome (hg19) using Bowtie2 [62] with the parameters "--end-to-end, --no-mixed, --no-discordant, --no-unal, and –sensitive" for host removal. In the profiling step, taxonomic profiles were generated using MetaPhlAn v.4.0.6 (https://github.com/biobakery/MetaPhlAn) with the reference database mpa_vOct22_CHOCOPhlAnSGB_202212 and parameters: "-t rel_ab_w_read_stats and --ignore_archaea". Functional profiling was conducted using HUMAnN v.3.7 (https://github.com/biobakery/humann) with reference database UniRef90s (version 201901b) and default parameters. The resulting abundance profiles are further summarized into other gene groups, such as KOs (KEGG Orthologs), by using the command "human_regroup_table" with the parameter "--groups uniref90_ko".

Preprocessing of microbial features

Microbial taxonomic and functional profiles were first normalized to relative abundances to account for library size. To reduce data noise and enhance data quality, we dropped low-abundance features. At the taxonomic level, features with an average relative abundance of more than 1×10^{-5} in at least two discovery cohorts were retained, while features with a relative abundance of zero in 90% of the samples were removed, and the standard deviation of the selected features was ensured to be greater than zero. At the KEGG Orthology (KO) level, features with a mean relative abundance greater than 1×10^{-4} across at least eight discovery cohorts were retained, following the same criteria for prevalence and variability. Similarly, at the uniref family level, features with a mean relative abundance exceeding 1×10^{-5} in at least five discovery cohorts were selected, adhering to the same filtering thresholds. The number of features on each type for analysis was summarized in **Table S2**. We then further filtered the data and only retained samples containing at least 50 species for subsequent analysis.

Then the microbial features were preprocessed under three methods:

i. original profile with no further processing.

ii. log-transformation: all relative abundance values were transformed into logarithmic values;

iii. batch-correction: all relative abundance values were corrected by MMUPHin [56] R package (v.1.8.2) via the "adjust_batch" function to eliminate the batch effects, in which source study ID was set as the controlling factor.

Meta-analysis

The beta diversity was measured by Permutational multivariate variance analysis (PERMANOVA; also referred to as Adnois test) that was implemented by package vegan [63] with 999 times permutation. *P*-value less than 0.01 denoted statistical significance. Principal coordinates analysis (PCoA) was performed by the labdsv package [64] based on the Bray-Curtis distance matrix. All meta-analysis was conducted using R software (version 4.1.3) [65].

We also used an ANOVA-like [35] analysis to quantify the effect of potential confounding factors and disease status on each microbial component. The total variance within the abundance of a given microbial feature was compared to the variance explained by disease status and the variance explained by confounders, based on an approach similar to linear models, including CRC status and various confounders (such as age, BMI, country, gender, and cohort origin) as explanatory variables. Variance calculations were performed on a rank basis to account for the non-Gaussian distribution of the microbiome abundance data [35]. Continuous confounding variables were transformed into categorical variables either by quartile-based grouping or, in the case of BMI, by standard thresholds into normal weight (BMI ≤ 25), overweight (BMI 25–30), and obese (BMI > 30).

Plotting was performed using R (version 4.1.3), Python (version 3.9), CNSknowall (https://www.cnsknowall.com/), and Chiplot [66] (https://www.chiplot.online).

Features selection

The details of seven feature selection tools are summarized in **Table 2**.

Table 2. Feature selection methods.

Method	Version / Implementation	Distribution	Criteria
ANCOM-II [67]	ANCOM-II function (version 2.1) in R	Non-parametric	W-score>0.8
MetagenomeSeq [68]	metagenomeSeq (version 1.36.0) package in R	Zero-inflated (log-) Normal	*p*-value <0.01
T-test	Stats (version 4.1.3) package in R	Normal (default)	*p*-value <0.01

Wilcoxon test	Stats (version 4.1.3) package in R	Normal (default)	*p*-value <0.01
MaAsLin2 [58]	MaAsLin2 (version 1.8.0) package in R	Normal (default)	*p*-value <0.01
RFECV [69]	RFECV function in the sklearn (version 1.3.2) library	step=10, cv=5	Highest accuracy
LEfSe [70]	Microeco package (version 1.4.0) in R	Non-parametric	*p*-value <0.05 & LDA>2

Machine learning models

Model construction. All machine learning-based (RF, XGBoost, MLP, KNN, and SVM) status detection was performed using the scikit-learn (version 1.3.2) package under Python 3.9. Among them, the neural network-based approach of MLP was implemented by PyTorch (version 1.11.0).

Hyper-parameter adjustment. In addition, for each ML model, the hyper-parameters were toned by GridSearchCV [71] (version 1.20.3) of scikit-learn.

Model evaluation. Performance of model was evaluated using metrics including AUROC (area under the receiver operating characteristic curve), F1-score (*Eq. 1*), AUPRC (area under the precision-recall curve), and MCC (Matthews correlation coefficient; *Eq. 2*). Correlations among these metrics were measured by Pearson correlation coefficient.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{2}$$

Instance-based transfer learning for cross-cohort ADA detection

i. Subset selection and matching by a mutual nearest neighbor set algorithm: Randomly select a subset (e.g., 30% samples) T' from the target domain T. For each sample t'_i in T', we search for its top k (e.g., k=10) nearest neighbors from source domain S by Euclidean distance, denoting as set $KNN_{Euclidean}(S, t'_i)$. Similarly, for each sample s_i in the source domain S, we also search its $KNN_{Euclidean}(T', s_i)$ set from T'. Here we define a sample pair $s_i \in S$ and $t'_j \in T'$ as pre-matched neighbor samples $N_{Euclidean}(s_i, t'_j)$ if and only if under *Eq. 3.*

$$s_i \in KNN_{Euclidean}(S, t'_j) \text{ and } t'_j \in KNN_{Euclidean}(T', s_i) \tag{3}$$

Then such pairing relation was re-confirmed by Cosine distance to determine the final matches $N_{Cosine}(s_i, t'_j)$. Finally, according to *Eq. 4*, we take the intersection of $N_{Euclidean}(si, t'j)$ and $N_{Cosine}(s_i,$

t'_j), thereby forming the mutual nearest neighbor set (MNN).

$$MNN(s_i, t_j^{'}) = N_{Euclidean}(s_i, t_j^{'}) \cap N_{Cosine}(s_i, t_j^{'}) \quad (4)$$

ii. Feature extraction and initial model: Utilize matched sample pairs *MNN* (s_i, t'_j) from the source and target domains to identify ADA-relevant features through synergistic feature selection (**Fig. 3d**). Subsequently, construct an initial Random Forest (RF) model using *MNN* (s_i) from the source domain.

iii. Pseudo labeling: Apply the initial model to parse out the detected status (*i.e.,* assign pseudo-labels) of each sample in target subset *T'*, forming a new set as *T''*.

iv. Dataset fusion: Combining *T''* with pseudo labels with the target domain *T'* to create an enhanced dataset *E (i.e., E = T'' + T')*.

v. Final model training: Train the final TL model on the enhanced dataset *E* using the optimal bioinformatics workflow (**Fig. 4c**).

Jensen-Shannon divergence analysis

To evaluate the distributional similarity between the source and target domains before and after transfer learning, we first normalize the feature values of each sample (*Eq. 5*).

$$P_i = \frac{X_i}{\sum_j X_{ij}} \quad (5)$$

Here, X_i represents the feature values of the *i*-th sample, and X_{ij} represents the value of the *j*-th feature for the *i*-th sample. Subsequently, we employ the Jensen-Shannon (JS) divergence (*Eq. 6*) to assess the similarity between the distributions of the two datasets. Each sample in the datasets was represented as a probability distribution over the feature space.

$$D_{JS}(P_i \parallel Q_j) = \frac{1}{2} D_{KL}(P_i \parallel M) + \frac{1}{2} D_{KL}(Q_j \parallel M) \quad (6)$$

Here, P_i and Q_j represent the probability distributions of the *i*-th sample from the target domain and the *j*-th sample from the source domain, respectively. where $M = (P_i + Q_j)/2$ is the mean distribution and $D_{KL}(P \parallel M)$ is the Kullback-Leibler (KL) divergence (*Eq. 7)*, which forms the basis of the JS divergence, quantifies the information loss when approximating one distribution *P* with another *Q*.

$$D_{KL}(P \parallel Q) = \sum_i P(k) \log\left(\frac{P(k)}{Q(k)}\right) \quad (7)$$

Here, k represent the features in the probability distribution of a single sample, such that *P(k)* and *Q*(*k*) denote the probabilities at the *k*-th feature for the respective samples.

The JS divergence is symmetric, satisfying $D_{JS}(P_i \parallel Q_j) = D_{JS}(Q_j \parallel P_i)$ and ranges between 0 and 1, where smaller values indicate greater similarity between the distributions. Thus, smaller JS divergence values suggest higher similarity between the sample distributions.

Data availability

The in-house produced (CHN_WF cohort) dataset has been deposited at the Sequence Read Archive (SRA) with project ID PRJNA1198725. The public datasets are available at SRA and ENA via identifiers listed in **Table 1**.

Code availability

Source codes and scripts for data analysis are available at the GitHub repository at https://github.com/qdu-bioinfo/CRC_project.

Author contribution

X.S. and X.F. conceived and designed the project. Y.S. and J.Z. designed the algorithms and workflow. X.F. collected the gut microbiome samples. Y.S., Z.W., J.X. and S.W. collected the public datasets. Y.S., W.Z. and H.G. performed the data analysis. X.S. and Y.S. drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

X.S. acknowledges support of grant 2021YFF0704500 from National Key Research and Development Program of China, grant 32070086 from National Natural Science Foundation of China, Shandong Province Youth Entrepreneurial Talent Introduction and Training Program, and Shandong Province Taishan Scholars Youth Experts Program. All authors thank Mr. Yi Zhao from Qingdao University for the support of computing resources.

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

1. Siegel, R.L., Giaquinto, A.N. & Jemal, A. Cancer statistics, 2024. *CA: a cancer journal for clinicians* **74**, 12-49 (2024).
2. Wong, S.H. & Yu, J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nature Reviews Gastroenterology & Hepatology* **16**, 690-704 (2019).
3. Liang, J.Q. et al. A novel faecal Lachnoclostridium marker for the non-invasive diagnosis of colorectal adenoma and cancer. *Gut* **69**, 1248-1257 (2020).
4. Shaukat, A. & Levin, T.R. Current and future colorectal cancer screening strategies. *Nature reviews. Gastroenterology & hepatology* **19**, 521-531 (2022).
5. Ladabaum, U., Dominitz, J.A., Kahi, C. & Schoen, R.E. Strategies for colorectal cancer screening. *Gastroenterology* **158**, 418-432 (2020).
6. Chen, H. et al. Participation and yield of a population-based colorectal cancer screening programme in China. *Gut* **68**, 1450-1457 (2019).
7. Kaminski, M.F., Robertson, D.J., Senore, C. & Rex, D.K. Optimizing the quality of colorectal cancer screening worldwide. *Gastroenterology* **158**, 404-417 (2020).
8. Senore, C. et al. Performance of colorectal cancer screening in the European Union Member States: data from the second European screening report. *Gut* **68**, 1232-1244 (2019).
9. Ren, Z. et al. Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. *Gut* **68**, 1014-1023 (2019).
10. Jiao, N. et al. Suppressed hepatic bile acid signalling despite elevated production of primary and secondary bile acids in NAFLD. *Gut* **67**, 1881-1891 (2018).
11. Chen, D., Wu, J., Jin, D., Wang, B. & Cao, H. Fecal microbiota transplantation in cancer management: Current status and perspectives. *International journal of cancer* **145**, 2021-2031 (2019).
12. Chen, G. et al. Exploring the gut microbiome's role in colorectal cancer: Diagnostic and prognostic implications. *Frontiers in Immunology* **15**, 1431747 (2024).
13. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature Communications* **6**, 6528 (2015).
14. Yu, T. et al. Fusobacterium nucleatum promotes chemoresistance to colorectal cancer by modulating autophagy. *Cell* **170**, 548-563. e516 (2017).
15. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature Medicine* **25**, 968-976 (2019).
16. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology* **19**, 55-71 (2021).
17. Ternes, D. et al. The gut microbial metabolite formate exacerbates colorectal cancer progression. *Nature Metabolism* **4**, 458-475 (2022).
18. Villéger, R. et al. Microbial markers in colorectal cancer detection and/or prognosis. *World journal of gastroenterology* **24**, 2327 (2018).
19. Kim, S.H. & Lim, Y.J. The role of microbiome in colorectal carcinogenesis and its clinical potential as a target for cancer treatment. *Intestinal Research* **20**, 31-42 (2022).
20. McQuade, J.L., Daniel, C.R., Helmink, B.A. & Wargo, J.A. Modulating the microbiome to improve therapeutic response in cancer. *The Lancet Oncology* **20**, e77-e91 (2019).
21. Jin, S., Zhong, W., Li, B., Wang, K. & Lai, D. Multidimensional analysis of the impact of Gemmatimonas, Rhodothermus, and Sutterella on drug and treatment response in colorectal cancer. *Frontiers in Cellular and Infection Microbiology* **14**, 1457461 (2024).

22. Thomas, A.M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine* **25**, 667-678 (2019).
23. Kumar, B., Lorusso, E., Fosso, B. & Pesole, G. A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Frontiers in Microbiology* **15**, 1343572 (2024).
24. Marcos-Zambrano, L.J. et al. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology* **12**, 634511 (2021).
25. Nearing, J.T. et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications* **13**, 342 (2022).
26. Teixeira, M. et al. A review of machine learning methods for cancer characterization from microbiome data. *NPJ Precision Oncology* **8**, 123 (2024).
27. He, Y. et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine* **24**, 1532-1535 (2018).
28. Li, M. et al. Performance of gut Microbiome as an independent diagnostic tool for 20 diseases: cross-cohort validation of machine-learning classifiers. *Gut Microbes* **15**, 2205386 (2023).
29. Bang, S. et al. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. *Scientific Reports* **9**, 10189 (2019).
30. Fischer, J. et al. Clinical implications of the genetics of sporadic colorectal cancer. *ANZ journal of surgery* **89**, 1224-1229 (2019).
31. Alves Martins, B.A. et al. Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research. *Frontiers in Oncology* **9**, 1284 (2019).
32. Cianci, N., Cianci, G. & East, J.E. Colorectal cancer: prevention and early diagnosis. *Medicine* (2024).
33. Zackular, J.P., Rogers, M.A.M., Ruffin, M.T. & Schloss, P.D. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. *Cancer Prevention Research* **7**, 1112-1121 (2014).
34. Mo, Z. et al. Meta-analysis of 16S rRNA Microbial Data Identified Distinctive and Predictive Microbiota Dysbiosis in Colorectal Carcinoma Adjacent Tissue. *Msystems* **5**, 10-1028 (2020).
35. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine* **25**, 679-689 (2019).
36. Costea, P.I. et al. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology* **35**, 1069-1076 (2017).
37. Forslund, K. et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262-266 (2015).
38. Maier, L. et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **557**, 623-628 (2018).
39. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology* **41**, 1633-1644 (2023).
40. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, e65088 (2021).
41. Bossche, J.V.d. et al. geopandas/geopandas: v0. 14.0. *Zenodo* (2023).
42. Gupta, A. et al. Association of Flavonifractor plautii, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *MSystems* **4**, 10-1128 (2019).
43. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70-78 (2017).
44. Liu, N.-N. et al. Multi-kingdom microbiota analyses identify bacterial–fungal interactions and biomarkers of colorectal cancer across cohorts. *Nature Microbiology* **7**, 238-250 (2022).

45. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10**, 766 (2014).
46. Vogtmann, E. et al. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLOS ONE* **11**, e0155362 (2016).
47. Gao, R. et al. Integrated analysis of colorectal cancer reveals cross-cohort gut microbial signatures and associated serum metabolites. *Gastroenterology* **163**, 1024-1037. e1029 (2022).
48. Yang, Y. et al. Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nature Communications* **12**, 6757 (2021).
49. Hannigan, G.D., Duhaime, M.B., Ruffin IV, M.T., Koumpouras, C.C. & Schloss, P.D. Diagnostic potential and interactive dynamics of the colorectal cancer virome. *MBio* **9**, 10.1128/mbio. 02248-02218 (2018).
50. Avelar-Barragan, J. et al. Distinct colon mucosa microbiomes associated with tubular adenomas and serrated polyps. *npj Biofilms and Microbiomes* **8**, 69 (2022).
51. Gao, R. et al. Alterations, interactions, and diagnostic potential of gut bacteria and viruses in colorectal cancer. *Frontiers in cellular and infection microbiology* **11**, 657867 (2021).
52. Wu, Y. et al. Identification of microbial markers across populations in early detection of colorectal cancer. *Nature Communications* **12**, 3063 (2021).
53. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457-D462 (2016).
54. Suzek, B.E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015).
55. McKnight, D.T. et al. Methods for normalizing microbiome data: an ecological perspective. *Methods in Ecology and Evolution* **10**, 389-400 (2019).
56. Ma, S. MMUPHin: Meta-analysis methods with uniform pipeline for heterogeneity in microbiome studies. *R package version* **1** (2021).
57. Gihawi, A. et al. Major data analysis errors invalidate cancer microbiome findings. *MBio* **14**, e01607-01623 (2023).
58. Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. *PLOS Computational Biology* **17**, e1009442 (2021).
59. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* **27**, 1226-1238 (2005).
60. Chen, Y. et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, gix120 (2018).
61. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
62. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).
63. Oksanen, J. et al. Package 'vegan'. *Community ecology package, version* **2**, 1-295 (2013).
64. Roberts, D.W. labdsv: Ordination and multivariate analysis for ecology. *R package version* **1** (2007).
65. Team, R.C. R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria* (2013).
66. Xie, J. et al. Tree Visualization By One Table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. *Nucleic Acids Research* **51**, W587-W592 (2023).
67. Kaul, A., Mandal, S., Davidov, O. & Peddada, S.D. Analysis of microbiome data in the presence of excess zeros. *Frontiers in Microbiology* **8**, 2114 (2017).
68. Paulson, J.N., Stine, O.C., Bravo, H.C. & Pop, M. Differential abundance analysis for microbial marker-gene

surveys. *Nature Methods*, 1200-1202 (2013).

69. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389-422 (2002).

70. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biology* **12**, 1-18 (2011).

71. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *Journal of machine learning research* **13** (2012).