Methodology for Microbiome Meta-Analyses with a Focus on Colorectal Cancer

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Jakob Wirbel
aus Neuwied

Tübingen 2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	12.05.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Daniel Huson
2. Berichterstatter:	Prof. Dr. Stephan Ossowski

Table of Contents

Summary

The human microbiome has been recognized as an important cornerstone of human physiology and immunity, situated at the interplay between health and disease. Consequently, comparative metagenomic studies have identified potential microbial biomarkers for common diseases with initial promising results for colorectal cancer, amongst others. Assessing biomarker robustness and generalization across populations, however, is complicated by widespread technical heterogeneity and biological confounding, which is further compounded by a lack of standardized methodology for statistical analyses.

In my doctoral research, I aimed to develop and evaluate methodology for the statistical and machine learning analysis of clinical metagenomic data, with a special focus on colorectal cancer.

In the first part, I developed a simulation framework for the benchmarking of differential abundance testing methods based on implanting signals into real data, enabling more realistic benchmarks than previous efforts. Most methods failed to control the false discovery rate, especially under confounded conditions, but the Wilcoxon test and linear models as well as their confounder-corrected varieties showed best performance in this benchmark.

The second part describes the SIAMCAT R package as a user-friendly toolbox which provides machine learning workflows for the analysis of metagenomic data. The publication includes an example for how SIAMCAT can detect confounding and illustrates common machine learning pitfalls.

The third section describes a colorectal cancer meta-analysis, which was able to establish robust, globally predictive, and disease-specific taxonomic and functional microbial biomarkers for colorectal cancer based on eight available shotgun metagenomic datasets from three different continents. More recent analyses, extending the original results and including different data types, have identified bacteria consistently and reliably associated with colorectal cancer, representing the starting point for future mechanistic studies.

Going beyond colorectal cancer, I explored the cross-study application of microbiome-based machine learning models in a meta-analysis encompassing various diseases. I uncovered substantial challenges for the naive transfer of models across

datasets and proposed a strategy to address those based on augmentation with external controls.

The outcome of my doctoral research therefore consists of empirical recommendations for differential abundance testing and machine learning model transfer in microbiome data, a software package for statistical and machine learning workflows, and a set of globally predictive microbial biomarkers for colorectal cancer.

Zusammenfassung

Das menschliche Mikrobiom wird zunehmend als Eckpfeiler für die humane Physiologie erkannte, insbesondere bei der Entwicklung von Krankheiten. Verschiedene vergleichende metagenomische Studien versuchten daher, potenzielle mikrobielle Biomarker für häufige Krankheiten zu finden, mit ersten vielversprechenden Ergebnissen unter anderem für Darmkrebs. Die Identifikation von robusten und allgemein prädiktiven Biomarkers wird aber durch technische Heterogenität und biologische Störfaktoren erschwert. Eine weitere Komplikation ergibt sich durch den Mangel an standardisierter Methodik für statistische Analysen.

Das zentrale Ziel meiner Doktorarbeit war die Entwicklung und Bewertung von Methodik für statistische Analysen und maschinelles Lernen im Rahmen von klinischen metagenomischen Studien, mit besonderem Augenmerk auf Darmkrebs.

Im ersten Teil entwickelte ich einen Ansatz für die realistische Simulation von metagenomischen Daten durch die Implantierung von Signalen in reale Daten. Die meisten Methoden verzeichneten erhöhte Falscherkennungsraten, insbesondere dann, wenn auch Störfaktoren in den Simulationen abgebildet waren, doch der Wilcoxon test und lineare Modelle (sowie deren Störfaktor-korrigierten Variationen) zeigten die beste Leistung in dieser Benchmark.

Der zweite Teil beschreibt das SIAMCAT R-Paket, eine benutzerfreundliche und validierte Software, die Workflows für das maschinelle Lernen in Mikrobiomdaten bereitstellt. Die Publikation enthält ein Fallbeispiel dafür, wie SIAMCAT Störfaktoren entdecken kann, sowie Illustrationen von häufigen Fehlerquellen bei dem Design von Workflows für maschinellen Lernen.

Der dritte Abschnitt beschreibt eine Meta-Analyse zu Darmkrebs, die auf der Grundlage von acht verfügbaren metagenomischen Datensätzen aus drei Kontinenten robuste, global prädiktive und spezifische taxonomische und funktionelle mikrobielle Biomarker für Darmkrebs ermitteln konnte. Neuere Analysen, die über die ursprünglichen Ergebnissen hinausgehen, identifizierten konsistent mit Darmkrebs assoziierte Bakterien, was den Ausgangspunkt für künftige mechanistische Studien zu der Rolle des Mikrobioms bei Darmkrebs bilden kann.

Über Darmkrebs hinausgehen untersuchte ich in einer Meta-Analyse mit verschiedenen Krankheiten, wie Modelle des maschinellen Lernens über verschiedene Studien hinweg angewendet werden können. Die naive Übertragung von Modellen auf andere Datensätze bringt erhebliche Herausforderungen mit sich, welche durch eine Strategie, die auf der Erweiterung von Datensätzen mit externen Kontrollen beruht, bewältigt werden konnten.
Das Ergebnisse meiner Doktorarbeit bestehen daher aus konkreten Empfehlungen für das Testen von differenzieller Abundanz und für die Übertragung von Modellen des maschinellen Lernens in Mikrobiomdaten, einer Software für die statistischen Analyse und maschinelles Lernen, sowie in global prädiktiven mikrobiellen Biomarkern für Darmkrebs.

List of Publications and Personal Contributions

Accepted publications

1. Wirbel*, J., Pyl*, P.T., Kartal, E. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* **25**, 679–689 (2019). https://doi.org/10.1038/s41591-019-0406-6

   ***Personal contribution**: I helped to develop the workflows for and performed the statistical analyses, which were the basis for all but three panels of the main and extended data figures. Additionally, I helped to write the original manuscript draft and to address the reviewer and editorial comments.*

2. Wirbel, J., Zych, K., Essex, M. *et al.* Microbiome meta-analysis and cross- disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol* **22**, 93 (2021). https://doi.org/10.1186/s13059-021-02306-1

   ***Personal contribution**: I helped to develop the software discussed in this article. Furthermore, I collected all data and performed most of the statistical analyses presented in the main and supplemental figures. Lastly, I helped to write the original manuscript draft and to address the reviewer and editorial comments.*

Not yet submitted manuscripts

1. Wirbel*, J., Essex*, M., *et al.* Evaluation of microbiome association models under realistic and confounded conditions. (see Appendix)

   ***Personal contribution**: I helped to develop the software and perform the statistical analyses presented in this manuscript. Additionally, I helped to write the original draft of the manuscript.*

   An updated version of this manuscript is now available on bioRxiv: https://doi.org/10.1101/2022.05.09.491139

   ** These authors contributed equally*

Introduction

The study of microbial communities through metagenomic sequencing

Microorganisms are the most ancient form of life on earth and have dominated the majority of the evolutionary history of the planet. Virtually all environmental niches on earth have been colonized by microbes, including habitats with extreme physical conditions such as deep-sea hydrothermal vents or hypersaline lakes (Merino et al., 2019). In particular, microbes can be found on the surfaces and in the cavities of all living animals and plants, where they interact with their respective host in different ways (Casadevall and Pirofski, 2000): microbial symbionts enter a mutually beneficial relationship with their host, whereas pathogens exploit and damage the host, leading to disease. Commensal microbes persist on and within the host without a clear mutualistic relationship or apparent pathogenicity.

Typically, microbes do not colonize an available environment individually. Rather, they form communities with complex interdependencies and dynamics. The entirety of all microbial species within a specific ecological niche has been referred to as ‘microbiome’ or alternatively ‘microbiota’. The human skin microbiome, for example, encompasses the community of microbes living on the human skin.

The study of microbial communities was initially restricted to individual microbes that could be grown in a laboratory setting, which is a laborious and time-consuming process and restricted to culturable microbes. With the wider availability of DNA sequencing, Carl Woese and colleagues pioneered culture-independent approaches for the study of microbes based on sequencing the 16S ribosomal RNA (rRNA) gene as a phylogenetic marker gene, leading for example to the discovery of the kingdom of Archaea (Woese and Fox, 1977). Even today, targeted sequencing of regions within the 16S rRNA gene (Caporaso et al., 2011) is a widely used method for the identification of microbial species from environmental samples (referred to as 16S amplicon sequencing from here on). Due to the historical choice of the 16S rRNA marker gene, large databases exist for the taxonomic annotation of 16S amplicon data (Bolyen et al., 2019; Matias Rodrigues et al., 2017; Quast et al., 2013).

The development of next-generation sequencing technologies made it more affordable to sequence all available DNA in a given environmental sample, allowing not only for the identification of microbes but also for the quantification of microbial functions (HMP

Consortium, 2012; Qin et al., 2010). This approach, called shotgun metagenomic sequencing, allows for more accurate and more highly resolved taxonomic characterization of microbiome samples, including species without reference genomes (Milanese et al., 2019). Additionally, shotgun metagenomic sequencing data enabled researchers *inter alia* to analyze the global distribution of microbial sub-species (Costea et al., 2017a; Karcher et al., 2020) or to create large catalogs of unexplored microbial diversity through metagenomic assembly (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019). Interestingly, even in the human gut microbiome, one of the best-studied environments, around 70% percent of putative microbial species currently lack a cultured representative (Almeida et al., 2020), highlighting the power of culture-independent methods, with recent efforts trying to close this gap (Poyet et al., 2019).

The human microbiome at the fulcrum of health and disease

The human-associated microbiota has long been the focus of researchers due to its emergence as a determinant of health and disease (Lynch and Pedersen, 2016). In a healthy state, the microbiome fulfills a crucial physiological role by training and calibrating the human immune system, which in turn shapes and controls the microbiome (Belkaid and Hand, 2014). When this intricate interplay is dysregulated, diseases, especially non-communicable diseases with an immune system component, can potentially develop (Belkaid and Hand, 2014). In this context, the human gut microbiome is of particular interest, since it contains the highest density and the largest diversity of microbial organisms in the human body (Cani, 2017).

A common approach to uncover changes in microbiome composition between healthy and diseased individuals are metagenome-wide association studies (MWAS), in which every component of the microbiome is tested for a cross-sectional association with the physiological state of interest by case-control group comparisons. This way, the human gut microbiome was found to be associated with a wide array of diseases, ranging from inflammatory bowel disease (IBD) (Gevers et al., 2014; Schirmer et al., 2018) over liver disease (Hoyles et al., 2018; Loomba et al., 2017; Qin et al., 2014) to rheumatoid arthritis (Zhang et al., 2015) or type 2 diabetes (Karlsson et al., 2013; Qin et al., 2012), to name just a few.

A bit more surprisingly, changes in the gut microbiome have also been linked to neurological disorders, possibly mediated through a `gut-brain axis` (Rhee et al., 2009). For example, patients with Parkinson's disease show differences in their gut microbiome compared to age-matched healthy controls (Bedarf et al., 2017; Scheperjans et al., 2015).

One of the earliest cases of clear disease-associated gut microbiome changes was colorectal cancer (CRC): two seminal publications identified *Fusobacterium nucleatum* to be enriched in CRC tissue using 16S amplicon and total RNA sequencing (Castellarin et al., 2012; Kostic et al., 2012). These findings were followed by mechanistic studies showing that *F. nucleatum* can adhere to and invade colon epithelial cells and activate pro-inflammatory responses (Kostic et al., 2013; Rubinstein et al., 2013). Additionally, other microbial species such as enterotoxigenic *Bacteroides fragilis* or genotoxic *Escherichia coli* were hypothesized to play a role in CRC carcinogenesis (Sears and Garrett, 2014). A metagenome-wide association study then uncovered a clear signal for CRC in the composition of the human gut microbiome, including enrichments not only for *F. nucleatum* but also for *Porphyromonas* and *Peptostreptococcus* species (Zeller et al., 2014). In the following years, more and more microbiome association studies for CRC have been published, both based on 16S amplicon (Baxter et al., 2016; Flemer et al., 2018; Zackular et al., 2014) as well as on shotgun metagenomic sequencing (Feng et al., 2015; Vogtmann et al., 2016; Yu et al., 2017).

For other cancer entities, the link between the gut microbiome and carcinogenesis is less well studied. However, recent studies have shown an association between microbial signals and cancer progression. Several studies could correlate specific microbes with response to immune checkpoint inhibition (ICI) therapy in melanoma, kidney, and lung cancer patients (Gopalakrishnan et al., 2018; Matson et al., 2018; Routy et al., 2018). The mechanism of action for this observation is not yet fully understood, but tumor antigens resembling microbial peptides (Fluckiger et al., 2020) and microbial metabolites (Mager et al., 2020) were proposed as possible explanations. Similarly, the outcome of chemotherapy treatment in pancreatic cancer has been reported to depend on microbes. However, these microbes were not assessed in the intestine , but rather directly within the tumor (Geller et al., 2017). This 'tumor microbiome' was also linked to survival in pancreatic cancer in another study (Riquelme et al., 2019). These and other initial results have fueled interest in intratumoral microbes, which were found in a wide range of cancer entities in a large-scale survey using

16S amplicon sequencing (Nejman et al., 2020). Interestingly, microbial reads could also be detected in human-focused 'omics surveys, for example in whole-exome, whole-genome, or total RNA sequencing experiments from The Cancer Genome Atlas (TCGA) (Dohlman et al., 2021; Poore et al., 2020; TCGA Research Network et al., 2013).

Potential for clinical applications of microbiome research

These and other results highlight the potential for microbiome-centered clinical applications, either as diagnostic tools or for therapeutic interventions. The most prominent example of a clinical application arising from microbiome research is arguably the treatment of recurrent *Clostridioides difficile* infection (rCDI), which can lead to severe and potentially life-threatening diarrhea and is often unresponsive to antibiotic treatment. A healthy gut microbiome is believed to be protective against rCDI (Ananthakrishnan, 2011) and therefore, restoration of the gut microbiome through fecal microbiota transplant (FMT) from healthy donors has been explored as a treatment for rCDI (van Nood et al., 2013). FMT was found to be highly effective to cure rCDI, which led to the exploration of FMT-based treatment also for other diseases, for example for inflammatory bowel disease (Colman and Rubin, 2014). Similarly, the emerging impact of a diverse gut microbiome on the response to immune checkpoint inhibition has led to clinical trials exploring the efficacy of FMT in metastatic cancer patients: stool of patients with a good response to ICI was transplanted into melanoma patients before treatment with ICI, leading to favorable clinical outcomes in a subset of patients (Baruch et al., 2021; Davar et al., 2021).

For colorectal cancer, both diagnostic and therapeutic avenues have been suggested. In Zeller et al., the authors propose a diagnostic microbial signature for CRC based on microbial marker species quantified in feces. The most commonly used non-invasive CRC diagnostic test is the fecal occult blood test, which aims to improve compliance of population screening programs for colorectal cancer. Its findings are to be confirmed by the more invasive colonoscopy, as it has limited sensitivity and specificity (Allison et al., 1990). Hence a more accurate non-invasive screening procedure could potentially transform CRC diagnosis. In their publication, the authors showed that a combination of microbial biomarkers and the fecal occult blood test could theoretically improve the sensitivity for CRC detection by over 45% compared to the fecal occult blood test alone, at least in the analyzed cohort (Zeller et al., 2014). Regarding therapeutic intervention, a study focused on

the persistence of *Fusobacterium nucleatum* in CRC-derived metastases also showed that treatment with the antibiotic metronidazole reduced *F. nucleatum* load as well as tumor growth in a mouse xenograft model (Bullman et al., 2017). Although not yet validated in humans, these results could be of relevance for the treatment of CRC patients, since *F. nucleatum* presence has been associated with a more aggressive phenotype and worse prognosis (Mima et al., 2016). Alternatively, researchers have started to develop vaccines against *F. nucleatum*, originally in the context of oral disease (Liu et al., 2009), but more recently also with a focus on CRC prevention (Brennan and Garrett, 2019; Guo et al., 2017).

Challenges for microbiome data analysis

As the findings of microbiome research are starting to move towards clinical applications, robust analysis methodology is important in order to obtain reliable results. However, several issues with statistical microbiome analyses that have the potential to lead to spurious associations are frequently encountered in the scientific literature, hence presenting serious challenges for a necessary consolidation of microbiome data analysis.

First, key data characteristics complicate the statistical analysis of microbiome data, since common assumptions about underlying distributions are often not met. For example, microbiome data are compositional by the nature of data generation through sequencing. This means that large shifts in highly abundant microbial taxa will by definition have influences on the measured abundances of all other taxa. This issue cannot be addressed without cumbersome protocol modifications (Vandeputte et al., 2017), but its importance for subsequent data analysis is still discussed in the literature (Tsilimigras and Fodor, 2016). Furthermore, large inter-individual variation (Voigt et al., 2015) reduces the power of statistical tests and leads to zero inflation in taxonomic profiles, meaning that many microbial species are not present in most of the samples, which in turn necessitates large sample sizes to detect differences in rare taxa. Additionally, the experimental process, for example the DNA extraction or 16S amplification step, is biased towards specific taxa over others, which can distort microbiome composition estimates and is often unaccounted for (McLaren et al., 2019).

Second, study effects - differences between studies in large parts attributable to technical factors - have been shown to strongly affect microbiome cohorts. For example, in the first meta-analysis of 16S-based microbiome studies published, the variation associated with

the study covariate outweighed the association with biologically meaningful factors, such as antibiotics treatment (Lozupone et al., 2013). Two large-scale efforts to assess data reproducibility across different laboratories and analysis pipelines reported that the amount of variance attributable to DNA extraction method or bioinformatic processing is comparable to or larger than biologically meaningful inter-individual differences, both for 16S amplicon (Sinha et al., 2017) as well as for shotgun metagenomics data (Costea et al., 2017b). Especially for 16S data, the largest fraction of variation can often be attributed to the choice of primers or the region in the 16S gene targeted for sequencing, as apparent in a meta-analysis of fecal 16S amplicon sequencing studies for CRC (Shah et al., 2018).

Lastly, some studies have shown that the gut microbiome composition can be subject to considerable confounding, that is when meta-variables (also called covariates) which are not the main variable of interest are strongly associated with differences in the gut microbiome. For the case of type 2 diabetes, two studies reported a strong diabetes signal in the gut microbiome composition but showed little overlap in significantly associated microbial taxa (Karlsson et al., 2013; Qin et al., 2012). A third study incorporating the two previous datasets later concluded that the majority of the reported type 2 diabetes signal could be explained by treatment with metformin prescribed to some of the type 2 diabetes patients in these studies (Forslund et al., 2015). Indeed, follow-up studies refined our understanding of how metformin leads to alterations in the gut microbiome, which might actually contribute to its efficacy against type 2 diabetes (Wu et al., 2017). Another example is the impact of proton-pump inhibitor medication on several microbial taxa (Imhann et al., 2016; Jackson et al., 2016), which could similarly confound microbiome association studies. In general, human-targeted drugs seem to be widely metabolized by commensal gut microbes (Zimmermann et al., 2019) and have likewise broad impacts on their viability (Maier et al., 2018), so that medication overall is of prime concern as a potential confounder in microbiome association studies. Other sources of confounding are linked to diet or lifestyle (Schmidt et al., 2018). For example, a recent study using the large American Gut Project dataset (McDonald et al., 2018) found differences in alcohol consumption, diet, and host physiological measurements to have a large impact on association statistics in case-control settings (Vujkovic-Cvijin et al., 2020).

The current state of microbiome analysis methodology

Faced with these challenges, the analysis methodology in microbiome research has to be rigorous and robust in order to lead to reliable biological insights and maximize the potential for clinical applications of microbiome-based findings. Although method development for microbiome data is an area of active interdisciplinary research, the design of evaluation frameworks and benchmarks is complex and existing attempts often suffer from a lack of realism leading to over-optimistic conclusions (Buchka et al., 2021).

One fundamental question currently discussed in the microbiome literature is which statistical test would be the most appropriate for the detection of differentially abundant taxa in metagenomic case-control studies. Some researchers emphasize the compositional nature of microbiome data and proposed methods that take this into account, such as ANCOM (Lin and Peddada, 2020; Mandal et al., 2015) or testing of ratios against reference frames (Morton et al., 2019). Another method, metagenomeSeq, tries to model a zero-inflated distribution for each microbial taxon to account for the observed sparsity of microbial sequencing count data (Paulson et al., 2013). Several benchmarking efforts applied those newly developed and other, more traditional methods to simulated data, but no consensus concerning the best statistical tool has emerged yet (Hawinkel et al., 2019; McMurdie and Holmes, 2014; Weiss et al., 2017), partly due to the employed simulation frameworks making different assumptions about the underlying distributions.

Another open question is how confounding variables should be taken into account when performing differential abundance tests. In a recent publication (Vujkovic-Cvijin et al., 2020), the authors argue to extensively match for possible confounders when performing MWAS, based on an exploration of the large dataset from the American Gut Project (McDonald et al., 2018). However, this strategy might not be feasible in each case, especially since datasets large enough for comprehensive matching are not commonly available. A special case of this consideration is the methodology for metagenomic meta-analyses, in which technical (various DNA extraction or sample handling protocol) as well as biological (diet or lifestyle) differences between studies need to be accounted for and are typically modeled using a study identifier as a covariate. Also here, a clear consensus about best practices has not been reached in the research community.

Individual microbial biomarkers for diseases can be at a high risk of confounding and usually carry limited signal. Therefore, combining signals across multiple associated

microbial taxa will often result in more robust disease signatures. In this context, multivariable statistical modeling through machine learning is a crucial tool for the identification and validation of potential biomarkers. Machine learning models can make predictions on external samples, thereby assessing the robustness of the biological signal and generalization of the signature. In most cases, a truly external dataset is not available, but the estimation of a model's accuracy is possible through cross-validation. Here, a part of the available data is set aside prior to model training, and the trained model is then evaluated on the left-out portion of the data.

In a setting with a binary outcome, such as case-control studies, the evaluation of a machine learning model is usually done via the receiver operating characteristic (ROC) analysis. The true-positive rate (TPR or sensitivity, defined as the fraction of correctly recognized instances among all positive ones) is plotted against the false positive rate (FPR, defined as the fraction of instances incorrectly labeled as positive among all truly negative ones, sometimes also denoted as 1 - specificity) at different prediction thresholds. The area under the ROC curve (AUC) subsequently serves as an aggregate estimate of the performance of the model, with an AUC of 1 corresponding to perfect classification accuracy and an AUC of 0.5 to a class assignment no better than random.

Although machine learning is widely used in the microbiome literature (Bang et al., 2019; Duvallet et al., 2017; Knights et al., 2011a, 2011b; Le Goallec et al., 2020; Pasolli et al., 2016; Wang et al., 2018), a user-friendly machine learning toolkit tailored towards the specifics of metagenomic data has not been published. This is particularly pressing since setting up machine learning workflows is a complex task for many researchers and some commonly made mistakes can lead to over-optimistic performance evaluations (He et al., 2018; Quinn, 2021).

Aims of this Work

Despite promises for biological insights and clinical applications, analysis of microbiome association studies using statistical tests and ML models is complicated by several fundamental challenges, ranging from complex data distributions to study effects and confounding. Additionally, over-optimistic performance evaluation and biased benchmarking studies are prevalent in the literature, which makes it difficult to choose the appropriate methodology for microbiome data analysis.
The goal of my doctoral research was to develop methods for the comparative analysis of data from clinical microbiome studies, with a special focus on colorectal cancer. My work can be separated into three parts: an unbiased evaluation benchmark for association tests under realistic and confounded conditions, the development of a machine learning toolbox for microbiome data, and a meta-analysis of metagenomic CRC studies.

For the **benchmarking of microbiome association tests**, I aimed to develop a new simulation framework that can produce data which recapitulates key characteristics observed in real metagenomics data. Additionally, the aim was to include confounding factors into this simulation framework. In a second step, the performance of different association tests was to be benchmarked in this new framework.

For the **machine learning toolbox**, my aim was to develop a user-friendly R package for the statistical and machine learning analysis of microbiome datasets. Additionally, I wanted to validate the package in a large machine learning meta-analysis, which presented the perfect setting to systematically explore the **cross-study application and generalization** of machine learning models.

For the **CRC meta-analysis**, my aim was to establish robust taxonomic and functional microbial CRC biomarkers using both univariate association tests as well as machine learning classifiers. These models were based on eight available shotgun metagenomic datasets from seven different countries to assess their global cross-study generalization and disease specificity for noninvasive detection of CRC.

Results and Discussion

Association testing for microbiome data

The benchmarking study (see Appendix) aimed to answer the question which statistical test would be most appropriate for identifying differentially abundant taxa in case-control microbiome studies. Although a fundamental question, a consensus has not emerged yet, as different benchmarking efforts reached widely different conclusions (Calgaro et al., 2020; Hawinkel et al., 2019; McMurdie and Holmes, 2014; Weiss et al., 2017). Additionally, the question of how confounding in microbiome studies can best be addressed with statistical models remains unanswered, which is ever more pressing as confounding is more and more recognized as a potential problem for inference in microbiome studies (Forslund et al., 2015; Schmidt et al., 2018).

Empirical assessment of realism for simulated metagenomic data

A common approach to evaluate methods in any benchmarking effort is to simulate data that include a ground truth of differentially abundant features and then to apply the methods to the simulated dataset, assessing how well the ground truth features are identified. Previous benchmarking efforts in the microbiome field used different models for the simulation of metagenomic profiles, based amongst others on multinomial, beta-binomial, or Dirichlet distributions (Hawinkel et al., 2019; McMurdie and Holmes, 2014). Data generated using those models, however, fails to capture key characteristics observed in real microbiome datasets (see **Fig. 1a**), particularly when considering feature variance -- a defining characteristic of microbiome data is the overdispersion not observed to the same extent in other sequencing data types, such as RNA sequencing. Machine learning models, commonly used in metagenomic association studies to detect biomarkers, are able to reveal even subtle differences between groups and exhibit larger sensitivity than ordination-based analyses (see also next section). Accordingly, machine learning models trained to distinguish between real and simulated data points could do so with almost perfect accuracy in most cases.

Since all evaluated parametric methods for simulation failed to generate realistic datasets for benchmarking, we developed an alternative method making no distributional assumptions, based on implanting differentially abundant features into a real dataset. This

way, essentially replacing parametric modeling by sampling, the original dataset is modified as little as possible and therefore, key data characteristics such as sparsity or feature variance are preserved (see **Fig. 1a**). The advantage of this method is that it combines realism with a ground truth: It can faithfully reproduce the challenges inherent to real data into which differentially abundant features with known effect sizes are then implanted in a controlled manner. As baseline dataset, we used the taxonomic profiles of a study investigating healthy adults (Zeevi et al., 2015), into which features with differential abundance were randomly implanted using repeated random splits of the dataset. Machine learning models, although extremely sensitive, can not distinguish between real and "simulated" data points in this setting, affirming the realism of the framework.

Other benchmarking efforts had already relied on real data, but analyzed case-control datasets without clearly defined differentially abundant features or varying effect sizes. In one case, a consensus vote across methods was used to identify differentially abundant features (Hawinkel et al., 2019), which can easily be biased by methods with extreme results or the selection of methods to be included, since methods with similar distributional assumptions can be expected to make similar errors. Two other studies evaluated concordance between methods across different datasets (Calgaro et al., 2020; Nearing et al., 2021), again without a ground truth of differentially abundant features.

Evaluation of differential abundance testing methods

To test the performance of statistical methods for the identification of differentially abundant features, several methods were applied to the data with implanted effects. As the main evaluation metrics, false discovery rate (FDR) as well as AUC (as a measure for the enrichment of true effects among those with a low *P* value) were recorded across the different repeats. Most methods, especially those developed for RNAseq data analysis (DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010)), failed to control the FDR at the nominal 5% level, with the extreme case of metagenomeSeq (*mgs*) (Paulson et al., 2013) recording an FDR of ~80% (see **Fig. 1b**). This issue was more pronounced at smaller sample sizes (N<100), with 8 of 11 methods showing FDRs above twice the nominal level.

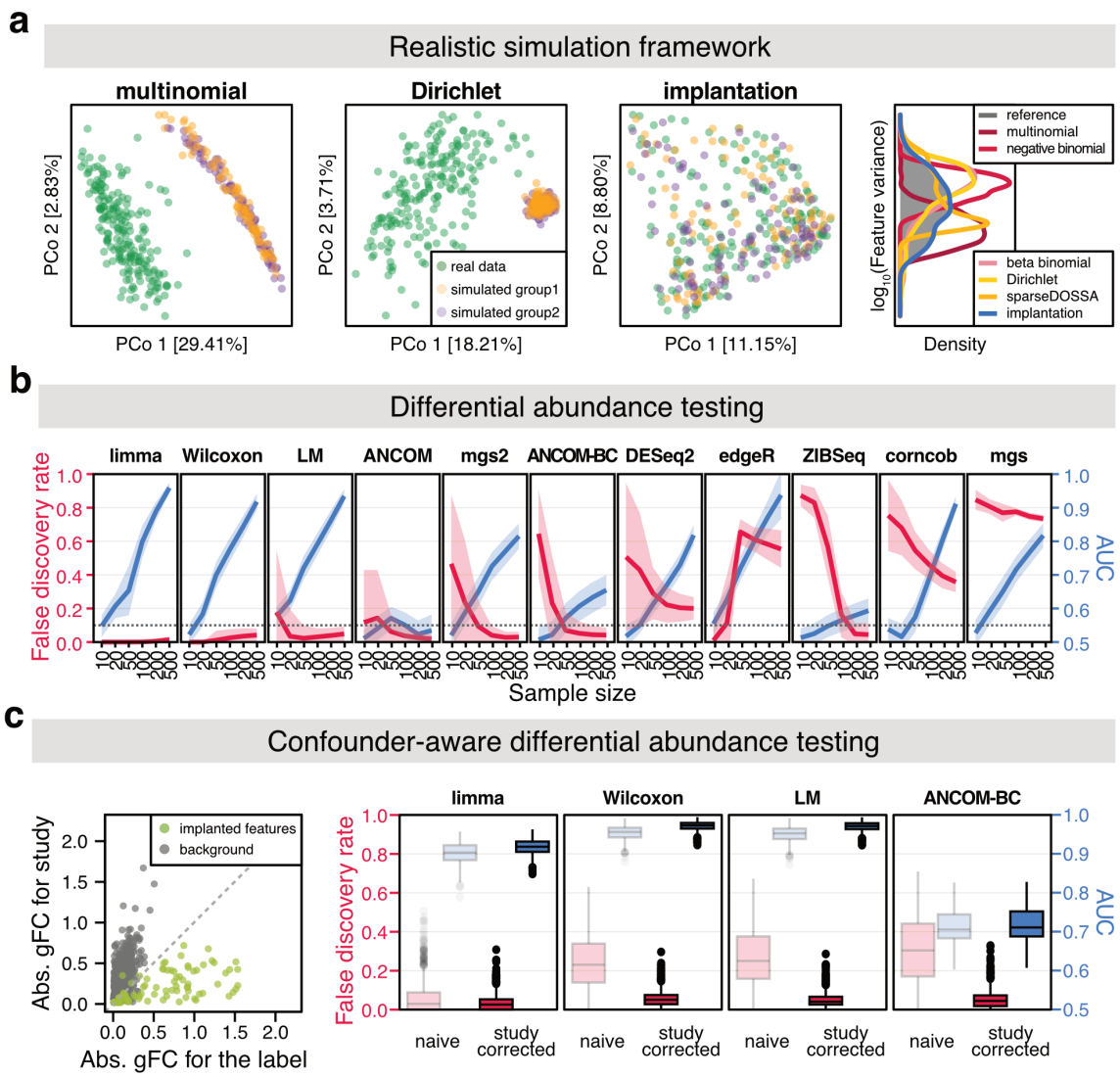


Figure 1: Summary of the benchmarking study

a) Principal coordinate analysis shows that the data simulated using parametric methods fail to reproduce key characteristics of microbiome data. The alternative approach, feature implantation, preserves general data distributions while at the same time allowing for control of differentially abundant features. Importantly, implantation -- in contrast to the parametric simulation methods -- reproduces realistic feature variances. **b)** False discovery rate (FDR) and AUC for the detection of implanted features (using taxon-wise *P* values as predictor) over different sample sizes are shown across all repetitions of the simulations shown in a). **c)** Using the implantation framework, differentially abundant features were implanted into two datasets using study as artificial confounding factor. On the left, the effect size associated with the implanted features is contrasted with the study effect size. For the methods with reasonable performance in b), the naive results are compared to the results after confounder adjustment at a fixed effect and sample size, given a strong study confounding.

A theoretical explanation for these results could be that the *P* values of these methods are poorly calibrated, that is, that a more appropriate cutoff for detection would reduce false detections while still identifying truly differentially abundant features. However, most methods exhibiting an elevated FDR also recorded relatively low AUC values for detection of

differentially abundant features, implying more general issues with the distinction between implanted and background features. Interestingly, even the methods that are specifically tailored towards microbiome data (ANCOM, ANCOM-BC, corncob, and *mgs*) exhibited inadequate performance. Lastly, the evaluation benchmark indicates that the Wilcoxon test, linear models, or limma (Ritchie et al., 2015) show the best performance for the detection of differentially abundant features in microbiome data, controlling both the FDR and exhibiting better statistical power than the other methods evaluated here.

The finding that most methods fail to control the FDR is in line with previous reports (Hawinkel et al., 2019). However, the same publication did also report elevated FDRs for the methods with reasonable performance in this benchmark, indicating that the choice of simulation framework is crucial for the resulting evaluation. Similarly, other benchmarks based on multinomial data (McMurdie and Holmes, 2014; Weiss et al., 2017) suggested using ANCOM (Mandal et al., 2015) or DESeq2 (Love et al., 2014), which showed poor performance in this more realistic benchmark.

Confounding poses additional challenges for differential abundance testing methods

To test how confounding factors can be addressed in differential abundance testing, we specifically simulated confounding through differences between studies (Costea et al., 2017b; Sinha et al., 2017) by combining data from two different publications in the implantation framework (Xie et al., 2016; Zeevi et al., 2015). Here, the propensity of a sample to be selected for one of the groups in the repeated random splits of the datasets was contingent on the study affiliation. In this setup, the degree of study confounding can be modulated by sampling one group from one dataset and the other group from the other dataset with different proportions. When these proportions are equal, data sets with minimal confounding are created.

Only methods with satisfactory performance (mean FDR across effect sizes not exceeding 10% for sample sizes 50 to 200) in the unconfounded benchmark were included, additionally requiring that the method could be adjusted for putative confounders through the inclusion of the corresponding covariate (‘study’ in our case) in the test formula. These conditions limited the exploration to the blocked Wilcoxon test, linear mixed effect (LME) models (instead of simple linear models), limma, and ANCOM-BC (Lin and Peddada, 2020).

The results from this evaluation show that applying naive testing strategies in the presence of confounding will lead to an excess of spurious associations (mean FDR between 24 and 31% for all methods except limma), even for confounders of moderate effect size (see **Fig. 1c**), since many features artificially show a strong association with the label. Explicitly modeling the covariate in the testing methods can rescue the performance of both the Wilcoxon test and linear models, resulting in mean FDRs close to the nominal level (5% vs 27% for the Wilcoxon and 5% versus 24% for the linear models at moderate study confounding) and comparable AUC values as observed for non-confounded simulations. While ANCOM-BC can control the FDR after confounder-adjustment, the resulting AUC for the detection of implanted features is still lower compared to the other three methods. For limma, the inclusion of the covariate in the model does reduce the FDR, but to a level that is still markedly above the nominal level (mean FDR of 28% with strong study confounding).

In summary, accounting for confounding factors with LME models or the blocked Wilcoxon test seems to be a promising approach to minimize spurious associations in confounded microbiome studies, although further exploration of other confounding factors are warranted. For example, a recent study highlighted alcohol consumption as an impactful dietary covariate that could confound disease associations (Vujkovic-Cvijin et al., 2020). This meta-variable was measured as different frequencies, representing therefore an ordinal variable instead of a binary one as evaluated in the presented benchmark.

As the findings of microbiome research are increasingly moving towards clinical application, consolidation of the statistical methodology is urgently needed. In other fields, this process has been aided by community-driven benchmarking projects, such as the critical assessment of protein structure prediction (CASP) (Kryshtafovych et al., 2019) or the critical assessment of metagenome interpretation (CAMI) for taxonomic profiling and assembly (Sczyrba et al., 2017). Similarly, several DREAM challenges have been organized to crowdsource challenges such as the inference of signaling networks (Prill et al., 2011). Although the presented benchmark currently represents the effort of only two research groups, we designed our benchmarking software to facilitate the addition of methods and the exploration of different baseline datasets to name just two possible extensions. In the future, a critical assessment project for the identification of DA features in metagenomic data might further accelerate our efforts to promote consolidation of statistical methodology in microbiome research.

Machine learning workflows for microbiome data

My second publication (Wirbel et al., 2021) describes the SIAMCAT R package as a toolbox for statistical and machine learning analysis of microbiome data. Machine learning is a crucial tool for the identification and validation of microbial biomarkers, since machine learning models can be applied to truly independent datasets for an unbiased evaluation of their prediction accuracy, thereby extending beyond simple differential abundance testing. Additionally, the combination of different biomarkers can lead to more robust classifiers. However, setting up machine learning workflows can be challenging for non-experts, especially when more advanced cross-validation procedures are needed. Machine learning has been used in more and more microbiome publications (Bang et al., 2019; Duvallet et al., 2017; Knights et al., 2011a, 2011b; Le Goallec et al., 2020; Pasolli et al., 2016; Wang et al., 2018), but the employed software is usually not made public as an easy-to-use package that can readily be applied to other datasets. Therefore, a flexible and user-friendly toolbox for the comparative analysis of clinical microbiome data with validated and rigorous machine learning workflows was still missing before I started my doctoral studies.

The SIAMCAT R package

The SIAMCAT package is implemented in the R programming language and available through the Bioconductor framework (Huber et al., 2015). The standard pipeline (see **Fig. 2a**) is based on the analyses performed in a case-control study exploring associations between the fecal microbiome composition and colorectal cancer (Zeller et al., 2014). It includes steps to filter and normalize the data, perform differential abundance testing using a Wilcoxon test, split the samples into cross-validation folds and then train and evaluate a machine learning model. The functions are tailored towards microbiome data and its specific characteristics, e.g. filtering can be performed to specify a minimum prevalence or abundance of the taxa to be retained in the analysis. Individual steps of the pipeline can be flexibly omitted or combined, allowing more advanced users to build more complicated workflows. The input for SIAMCAT consists of relative abundances of microbial species or functional groups and a binary label describing the group membership for each sample (e.g. cancer versus tumor-free in the mentioned example). Optionally, meta-variables can be supplied as well, which can then be tested as potential confounding factors.

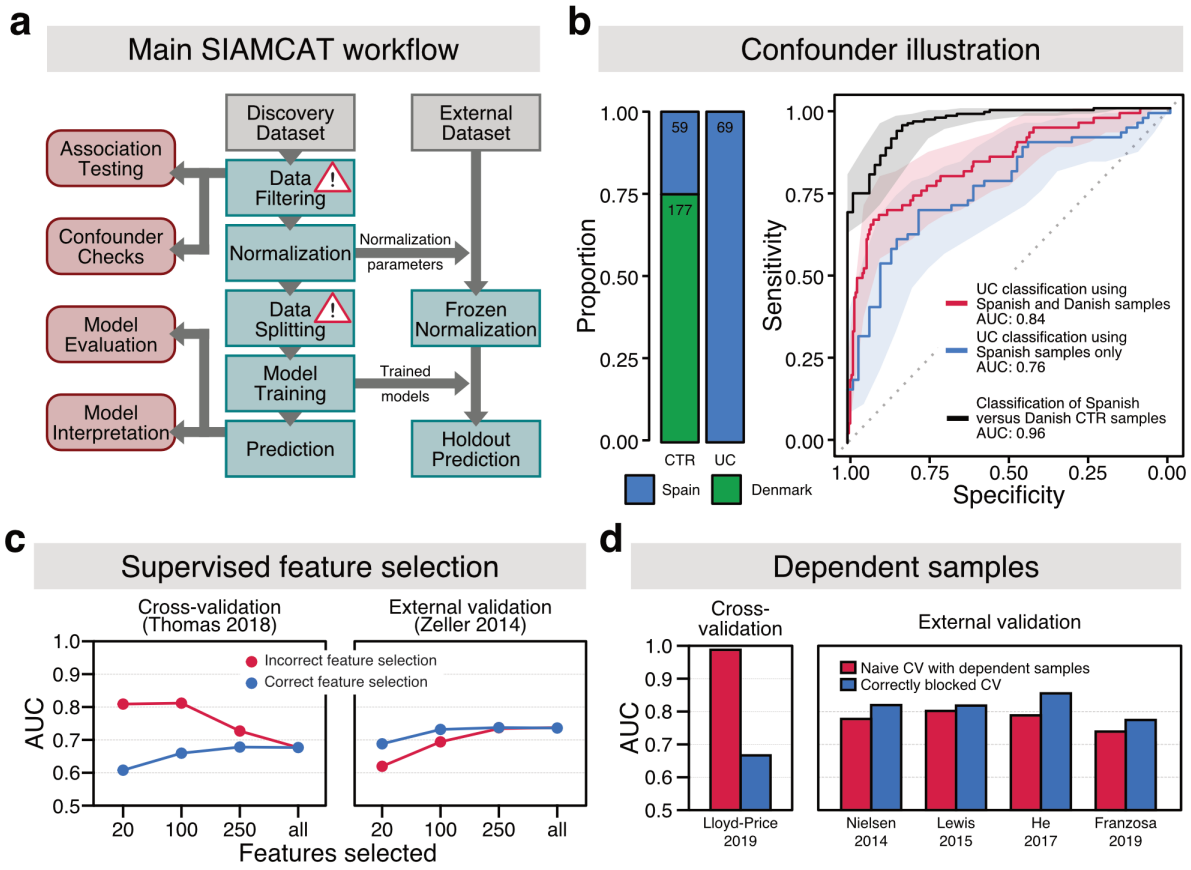


Figure 2: Summary of the SIAMCAT publication

a) Each step (green boxes) in the standard pipeline of a SIAMCAT analysis is implemented as a modular function. Workflow steps that produce visual outputs are shown on the left in the red boxes. Additionally, the pipeline to transfer machine learning models to external data is shown on the right. Lastly, the steps where SIAMCAT safeguards against common machine learning pitfalls are indicated by caution signs. **b)** To illustrate confounding in real datasets, the data from (Nielsen et al., 2014) was analyzed with SIAMCAT. Study was flagged as a potential confounder, since there is a strong association with the label (Fisher's test *P* value: 4.47E-33). As a result, machine learning model performance is affected: A model trained with all data (red line) performs better than a model based only on the Spanish samples (blue line), since there is an apparent difference between Spanish and Danish control samples (black line). Panels **c)** and **d)** illustrate two common pitfalls for machine learning workflows: in **c)**, the cross-validation performance is higher for the incorrect feature selection than for the correct strategy. However, the models resulting from the incorrect procedure fail to generalize to external data compared to the correct one. Similar results are shown in panel **d)** for the naive and blocked cross-validation procedure given repeated samples for the same individual. Figure adapted from (Wirbel et al., 2021).

The main outputs for a SIAMCAT pipeline are visualizations for the results of the differential abundance testing, the confounder analysis, and the machine learning model. For the model, the package produces both an evaluation plot (showing ROC and precision-recall curves) and an interpretation plot that allows for model introspection by displaying the weights assigned to the different features, i.e. microbial taxa or functions. As of now, the SIAMCAT R package is limited to a case-control setting with a binary classification task.

Other types of machine learning tasks such as regression analysis are planned to be implemented in the future. For multi-class classification, SIAMCAT can be employed after converting the task into binary classification problems and ongoing developments will include convenience functions to make this easier for users.

Illustration of confounding via SIAMCAT

The SIAMCAT R package includes a step to check the supplied meta-variables for potential sources of confounding, since it has been demonstrated that medication, study effects, or other biological or technical variables can dramatically impact microbiome composition (Schmidt et al., 2018), potentially leading to spurious associations (Forslund et al., 2015; Imhann et al., 2016; Jackson et al., 2016). In the *check.confounder* function, meta-variables are checked for an association with the label using either a Wilcoxon or a Fisher's test for continuous or categorical variables, respectively. Additionally, the amount of variance explained by the meta-variable is contrasted with the variance explained by the label, for each feature independently.

As a demonstration of the dramatic effects confounding can have, the dataset from (Nielsen et al., 2014), containing samples from ulcerative colitis (UC) patients and controls, was analyzed using a naive SIAMCAT workflow (see **Fig. 2b**). In this dataset, control samples were taken from both Spanish and Danish subjects, while UC samples were obtained only from Spanish individuals. Here, the country variable can be seen as a proxy for other difficult-to-measure factors, such as diet or lifestyle, while it captures at the same time technical differences within the same study. SIAMCAT can help to identify this type of confounding by indicating which measured meta-variables have an association with the label. In this dataset, the study confounder does lead to spurious associations when features are naively tested for differentially abundance. Additionally, the results from machine learning models can also be confounded: A model to distinguish between UC cases and controls performs seemingly better when the Danish samples are included, since the strong differences between Spanish and Danish control samples are exploited by the model.

SIAMCAT can indicate when meta-variables are potential confounders that could lead to spurious associations or inflated machine learning model performance estimates. The problem remains that sources of confounding are often unknown or unmeasured and can

therefore not be tested. Additionally, SIAMCAT in its current implementation will only identify a potential confounder. Planned future developments will then allow it to incorporate the confounder information into differential abundance testing, using the blocked Wilcoxon test (Hothorn et al., 2006) or LME models as discussed in the previous section. In machine learning, however, confounding can lead to over-optimistic performance estimates, which can not be corrected in a similar way. Here, removal of data according to the confounder information (as done in the example discussed above) seems to be the easiest way to account for confounding. Alternatively, a strategy of adjusting the data for potential confounders before machine learning, for example through modeling the confounder effect in principal coordinate space (Price et al., 2006), has been employed previously (Qin et al., 2012). As this issue is prevalent outside the microbiome field as well, the hope is that future developments in confounder-aware machine learning (Zhao et al., 2020) might be applicable to microbiome studies.

Illustration of machine learning pitfalls via SIAMCAT

Machine learning models are often used in microbiome publications for the identification of biomarkers and to assess the generalizability of a microbial signature. For this task, it is important that the model is evaluated on an independent test set so that its performance on new data can be estimated. If no truly external dataset is available, researchers usually resort to cross-validation, in which a part of the data is excluded from training and reserved for testing. However, setting up the cross-validation workflow incorrectly can lead to over-optimistic estimates for the performance of the model, which is unfortunately commonly seen in microbiome literature. In the SIAMCAT publication (Wirbel et al., 2021), two well-described (Roberts et al., 2017; Smialowski et al., 2010) machine learning pitfalls are illustrated using real datasets.

The first pitfall occurs when label-aware (or supervised) feature selection is performed on the complete dataset before splitting the data into cross-validation folds (Smialowski et al., 2010). In this case, the label information of all samples is taken into account for feature selection, for example through differential abundance testing. When this is combined with a naive cross-validation to train and test a model on the retained features, the label information from the test set has leaked into the model, resulting in performance estimates that can be overly optimistic (also called overfitting), especially for sample sizes that are

small relative to the number of features. On the other hand, supervised feature selection is sometimes needed, for example when functional data with tens of thousand functional groups are used as input. To correctly incorporate supervised feature selection into a machine learning workflow, the selection step needs to be folded into the cross-validation, that is, only the information of the training data can be used for feature selection, separately in each cross-validation fold.

In the worst case, this pitfall can lead to the incorrect biological interpretation of results (see **Additional Figure 1**): In a recent publication, the authors claim that microbiome-based machine learning models for the detection of common diseases might be accurate within a region but fail to extrapolate across different regions in China (He et al., 2018). They come to this conclusion because they observe good model performance in cross-validation within a region and a stark performance drop when the models are applied to different regions. However, in their machine learning workflow, supervised feature selection was performed incorrectly before cross-validation within a region, thereby leading to inflated within-region performance estimates. With the correct workflow (**Additional Figure 1**), the original conclusion is not supported anymore, since disease models do not show a strong and consistent drop in performance when applied across regions, partly because the initial within-region models do not show such a high (that is, inflated) performance. To raise awareness of this commonly encountered issue, the SIAMCAT publication included an illustration of this behaviour using two CRC datasets (see **Fig. 2c**).

The second issue arises when samples are not independent from each other. This can occur when multiple samples from the same individual are analyzed, which tend to be more similar to each other than across individuals (Voigt et al., 2015). Likewise, geographical structure from environmental samples need to be taken into account for cross-validation in some instances (Roberts et al., 2017). If repeated samples from the same individual (or geographical region) are naively split in cross-validation, so that some end up in training folds, others in the test fold, the cross-validation accuracy effectively measures how well the model can generalize across time points rather than across individuals (or across short spatial distances rather than globally). To counter this pitfall, the cross-validation needs to be blocked so that the samples from the same individual or from the same region are always kept in the same fold and are not split across training and test fold (see **Fig. 2d**). As an example with the correct cross-validation setup, a study investigating the relationship

between surface temperature and ocean microbiome composition kept samples from the same ocean basin together in cross-validation (Sunagawa et al., 2015).

Avoiding common pitfalls in machine learning can be challenging for non-experts and therefore, problems are found in many publications (Quinn, 2021). SIAMCAT tries to safeguard against the most common pitfalls in the design of machine learning workflows by allowing for nested supervised feature selection and blocked cross-validation.

The human gut microbiome in colorectal cancer

The CRC meta-analysis publication (Wirbel et al., 2019) focuses on the in-depth characterisation of the fecal microbiome in colorectal cancer patients and healthy controls. Several metagenome-wide association studies for CRC had been published before (Feng et al., 2015; Vogtmann et al., 2016; Yu et al., 2017; Zeller et al., 2014), but it was unclear to what extent reported associations were consistent across studies, given large differences in biological as well as technical characteristics between individual studies. Additionally, we had gained access to a so-far unpublished cohort recruited at the DKFZ in Heidelberg, presenting an ideal setting for a meta-analysis.

Microbiome-centered meta-analyses for CRC had been published before but suffered from certain limitations: one meta-analysis that investigated 16S studies was restricted by the taxonomic resolution of targeted 16S amplicon sequencing and additionally reported extreme technical differences between studies, leading to low effect sizes for disease associations (Shah et al., 2018). Another study, investigating the four publicly available shotgun metagenomic datasets mentioned above, employed Kraken for taxonomic profiling, which is known to produce many false positive taxonomic assignments (Milanese et al., 2019), and thus reported results that were in part biologically implausible (Dai et al., 2018). Additionally, training and test sets were not strictly separated for feature selection, causing over-optimistic performance estimates as discussed above.

The aim of the shotgun metagenomic meta-analysis for CRC was to test the robustness of microbial biomarkers in the face of biological and technical cross-study heterogeneity and to assess how well a machine learning-based CRC signature trained with rigorous methodology would be able to generalize across studies. Additionally, the functional potential of CRC metagenomes was to be explored to identify bacterial functions enriched in cancer, possibly uncovering bacterial contributions to carcinogenesis.

Study effects and confounders

To assess the influence of study effects across CRC studies and other confounders, two analyses were performed: To gain a global view, principal coordinate analysis was performed on taxonomic profiles. Study effects generally outweighed the disease signal in principle coordinate space and the dataset from Feng et al. presented itself as an outlier. However, the severity of study effects was not comparable to what had been observed for

16S amplicon sequencing data (Shah et al., 2018). To refine this analysis towards individual taxa, the amount of variance attributable to available meta-variables, including study affiliation, was calculated for each microbial taxon and contrasted to the amount of variance that could be explained by cancer status. As expected, the strongest effects could be observed for the study variable, which is a proxy for both biological (for example different genetic predisposition or environmental exposures) and technical (such as differences in sample handling or DNA extraction protocols) heterogeneity. Other meta-variables generally had a smaller influence on most taxa, with the exception of the variable indicating if the sample was collected before or after colonoscopy. Therefore, all subsequent analyses were corrected for both the study as well as sampling relative to colonoscopy, using the blocked Wilcoxon test. Generally, the taxa most strongly affected by confounders were not strongly associated with the disease label and vice versa (see **Fig. 3a**), indicating that there is a consistent signal for CRC in a subset of microbial taxa.

Although several covariates were investigated as potential confounders, this analysis was limited by the availability of metadata consistently recorded across the studies. For example, diet information (vegetarian or non-vegetarian) or smoking status was available in only two of the included studies. Other potential confounding factors, for example exposure to drugs or antibiotics, were either not recorded in the original studies or had not been made publicly available. Therefore, the potential for confounding from unmeasured sources exists and more extensive characterization of cancer samples will be needed in future studies (Ogino et al., 2018), especially since older populations (as investigated in these studies) often present with several co-morbidities and regular drug intake. However, in contrast to the prominent example of confounding through metformin treatment in type 2 diabetes (Forslund et al., 2015), all samples in this meta-analysis had been collected prior to cancer diagnosis and treatment, rendering systematic confounding through anticancer medication unlikely.

Univariate association testing

To identify differentially abundant microbial taxa while at the same time accounting for the confounders described above, a blocked Wilcoxon test was used, which found more than 90 microbial species and 32 genera to be differentially abundant at a conservative false discovery rate of 0.005, with a core set of 29 microbial species showing an FDR of less than

1E-05 (see **Fig. 3b**). In general, our meta-analysis results were concordant with other publications, especially concerning the enrichment of *Fusobacterium*, *Peptostreptococcus*, or *Parvimonas* species (Feng et al., 2015; Yu et al., 2017; Zeller et al., 2014). Additional CRC-associated microbial species detected by this uniquely powered meta-analysis, especially from the Clostridiales order or species without a genomic reference (Milanese et al., 2019) were not previously associated with CRC. Notably, all 29 differentially abundant species were enriched in CRC samples and often undetected in control samples, which would support a model for the role of the gut microbiome in CRC in which specific, predominantly oral microbes either benefit from or contribute to the process of carcinogenesis (Flynn et al., 2016).

A companion publication (Thomas et al., 2019) analyzed a partly overlapping set of data and two additional cohorts from Italy in a complementary CRC meta-analysis. Although the method of taxonomic profiling, MetaPhlAn2 (Truong et al., 2015) versus mOTUs2 (Milanese et al., 2019), and the test for differential abundance (LME models versus the blocked Wilcoxon test) were different between the two publications, the results of the differential abundance testing were similar, with 13 of 17 CRC-enriched microbial taxa replicated in our studies, further highlighting the robustness of CRC associations for these microbial species.

Machine learning model transfer

To test the capacity of the fecal microbiome for the prediction of colorectal cancer, machine learning models were trained for each study individually using a ten times repeated ten-fold cross-validation scheme implemented in the SIAMCAT toolbox (Wirbel et al., 2021). To avoid potential leakage of label information from the training to the test set, machine learning model training was independent of the differential abundance analysis described above. Instead, feature selection was performed through training least absolute shrinkage and selection operator (LASSO) logistic regression models, which internally select informative features (Tibshirani, 1996).

In study transfer, single-study machine learning models generally retained similar accuracy, indicating again that the CRC signal is largely consistent across different geographies and experimental pipelines. A notable exception is the model trained on the dataset from Feng et al., which showed lower generalization to other datasets, in agreement with the confounder analysis performed before. The study from Vogtmann et al. generally exhibited

a lower signal than the other studies, most probably since fecal samples had been stored at -80°C for over 20 years before processing (Vogtmann et al., 2016).

To further improve the accuracy of the machine learning models, data were pooled across four of the five studies for training and evaluated on the left-out study (leave-one-study-out or LOSO validation). The availability of more training data markedly improved the accuracy of classifiers on the test set, with AUC values of >0.8 in most cases (see **Fig. 3c**). These results could also be replicated on three completely external datasets that became available during the revision process (Thomas et al., 2019; Yachida et al., 2019). Overall, these results suggest that more diverse training data can lead to improved performance of machine learning models (see also the next section) and that the fecal microbiome might be the basis for the development of a diagnostic test for the detection of colorectal cancer, again supported by similar results derived with a different methodology (LASSO versus random Forest models) in the companion publication (Thomas et al., 2019).

Functional profiles were also explored as input for the machine learning pipeline. Classifiers based on the KEGG (Kanehisa et al., 2014) or eggNOG (Huerta-Cepas et al., 2016) database showed very similar or slightly improved accuracies compared to those based on taxonomic profiles. Since the number of input features was much higher for functional profiles, feature selection had to be performed before model training, which was folded into the cross-validation procedure (as described above). However, many of the predictive gene groups are of unknown function, hampering the interpretation of these models.

Two previous machine learning meta-analyses had highlighted that several different diseases shared a general signal of dysbiosis (Duvallet et al., 2017; Pasolli et al., 2016). To test this expectation in our CRC machine learning meta-analysis, disease specificity of the CRC models was assessed using datasets from other diseases, including type 2 diabetes (Karlsson et al., 2013; Qin et al., 2012), inflammatory bowel disease (Qin et al., 2010; Schirmer et al., 2018), and Parkinson's disease (Bedarf et al., 2017). Single-study models did indeed show elevated predictions for other diseases, but LOSO models made much more disease-specific predictions in the tested datasets (see Fig. 3c in the CRC meta-analysis publication, Appendix). Overall, pooling of data improved both transfer accuracy and disease specificity of the machine learning models.

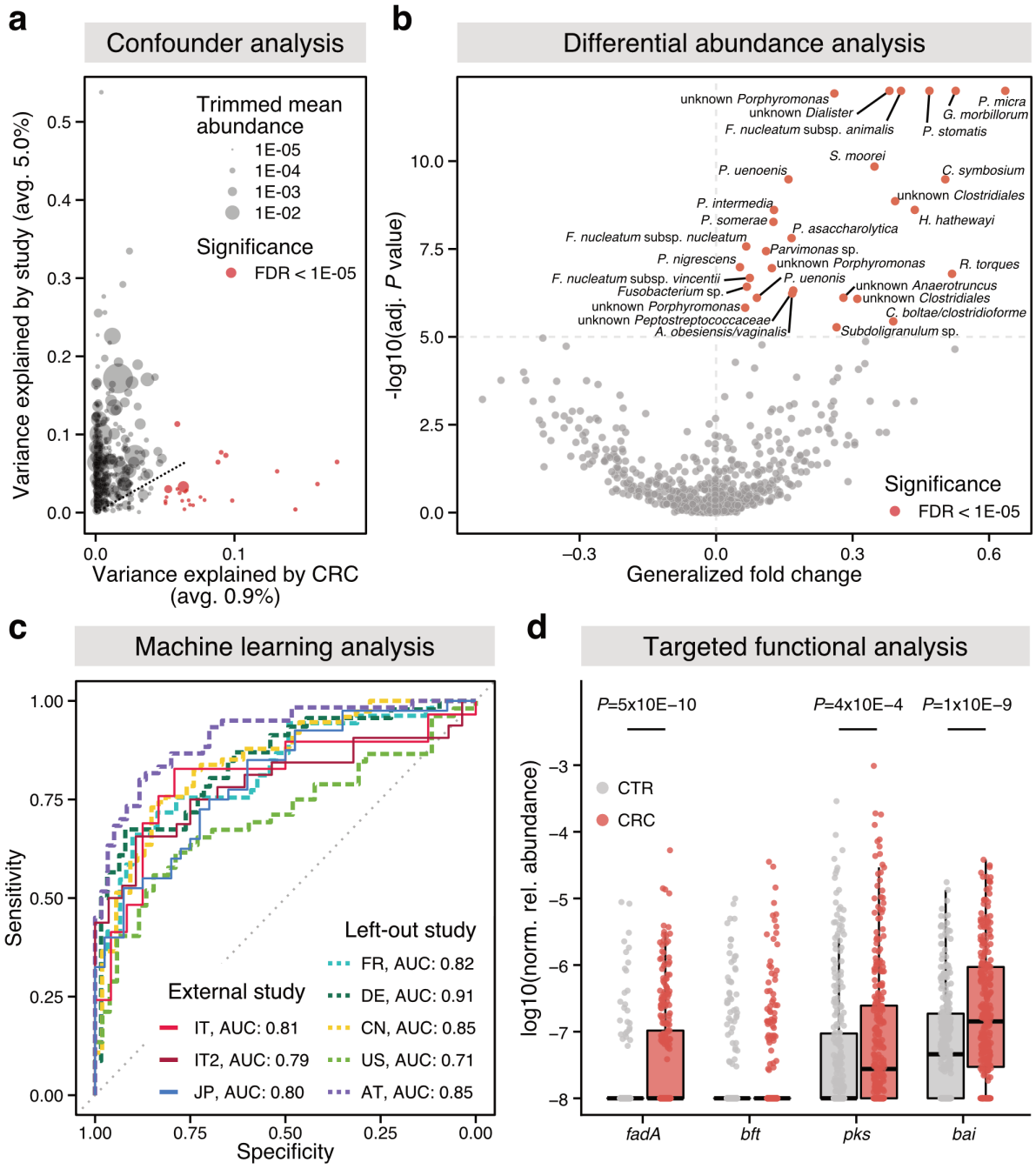


Figure 3: Summary of the CRC meta-analysis

a) An initial confounder analysis revealed that a subset of features vary more by disease status than by study. **b)** The results of the blocked Wilcoxon test for differential abundance analysis are shown as a volcano plot. The core set of microbial species with a strong association with CRC (FDR<1E-05) are labeled, highlighting previously reported species (e.g. *Fusobacterium* or *Peptostreptococcus* species) as well as others, some of which are without genomic reference. **c)** ROC-curves from the LOSO machine learning models show that in most cases, an AUC of >0.8 is possible. Similar results could be obtained for the truly external datasets from (Thomas et al., 2019; Yachida et al., 2019). Studies are labeled by the country where the main study population was recruited (see the manuscript in the Appendix for a legend). **d)** Targeted functional analysis of metagenomes reveals strong enrichment of virulence factors in CRC metagenomes. Figure adapted from (Wirbel et al., 2019).

Targeted exploration of functional CRC signals

Shotgun metagenomic sequencing allows not only for the taxonomic but also for the functional characterization of proteins encoded in a microbial community. However, databases such as KEGG or eggNOG are usually focussed on human or eukaryotic organisms and thus contain many (prokaryotic) gene groups without clearly defined functional annotation, impeding the interpretation of enriched gene groups. To overcome these limitations, functional data was mined in a targeted way for potential enrichments, resorting to metabolic modules and previously described virulence factors.

First, the metabolic potential of the gut microbiome was assessed using the gut metabolic module (GMM) framework developed in (Vieira-Silva et al., 2016). In CRC metagenomes, modules encoding for the degradation of amino acids, organic acids, or glycoproteins such as mucin were enriched, whereas carbohydrate utilization modules showed a higher abundance in controls. These associations have to be interpreted with caution, as metagenomic enrichments are only proxies for actual functional differences, such as differences in protein level or activity. However, the observed enrichments largely agree with studies finding increased amino acid concentrations in CRC feces (Goedert et al., 2014; Weir et al., 2013; Yachida et al., 2019) or tissue (Denkert et al., 2008; Hirayama et al., 2009; Mal et al., 2012) and with well-established dietary risk factors for CRC such as low fiber intake or increased red meat consumption (WCRFI, 2018).

Secondly, since a diverse set of bacterial virulence factors have been hypothesized to play a role in CRC (Sears and Garrett, 2014), we identified and quantified them using Hidden Markov models (see **Fig. 3d**). One virulence factor enriched in CRC metagenomes was the *fadA* adhesin of *Fusobacterium nucleatum*, which has been shown to be crucial for binding and subsequent invasion of colonic epithelial cells (Rubinstein et al., 2013). This functional enrichment is in agreement with taxonomic enrichment of *Fusobacterium* species in CRC microbiomes. The *bft* gene from specific *Bacteroides fragilis* strains had previously been implicated in CRC carcinogenesis as a potential genotoxin in mice (Wu et al., 2009). Furthermore, co-associations of the *bft* gene with the *pks* genomic island from *Escherichia coli* (which encodes for colibactin, a genotoxin inducing DNA double strand breaks (Nougayrède et al., 2006)) had been observed in patients with genetic risk factors for CRC (Dejea et al., 2018). In the meta-analysis, the *bft* gene showed no differential abundance and lower prevalence overall in CRC metagenomes, whereas *B. fragilis* was strongly

CRC-enrichmed in the taxonomic analyses (*P*=2E-5), just barely not being included in the core set of 29 marker species. The taxonomic profiles are currently unable to distinguish between strains that either carry the toxin or not, so that it is unclear whether the gene was truly not present or not well detectable with our Hidden Markov model-based approach. The *pks* operon, on the other hand, was found to be strongly enriched in CRC metagenomes. Interestingly, *pks*-positive *E. coli* strains were later shown to induce a specific mutational signature in human-derived colonic organoids (Pleguezuelos-Manzano et al., 2020) which is consistent with a proposed mechanism for how colibactin damages DNA through alkylation (Wilson et al., 2019). The last factor investigated was the *bai* operon present in some *Clostridium* species, which encodes for enzymes that metabolize bile acids via 7α-dehydroxylation (Ridlon et al., 2016). The resulting microbiome-derived secondary bile acids, deoxycholate and lithocholate, had been associated with liver cancer before (Yoshimoto et al., 2013), potentially through the induction of oxidative stress (Payne et al., 2007). Interestingly, the *bai* operon was consistently enriched in CRC metagenomes across all five studies, which was validated for one gene in the operon through quantitative PCR in a subset of the in-house processed samples, additionally revealing elevated transcript levels in CRC. In the companion publication from Thomas et al., the authors investigated the microbial *cutC* gene, which is required for the production of trimethylamine, a potential cancer-inducing metabolite (Oellgaard et al., 2017), and likewise found a consistent enrichment in CRC metagenomes.

Overall, the targeted analysis of the functional potential of the CRC metagenome confirmed previously reported pathways and revealed novel ones by which microbes might functionally contribute to CRC development. However, this targeted analysis was largely based on the limited mechanistic knowledge of host-microbe interactions revealed by hypothesis-driven experimental work. Therefore, a more comprehensive and broader exploration of potential virulence factors, including also bacterial functions not previously associated with CRC, could be expected to uncover even more relevant associations.

A broader view on the microbiome and CRC

The presented CRC meta-analysis focuses on the fecal microbiome, analyzed using shotgun metagenomic sequencing. The results indicate that a stool-based diagnostic application for CRC might be feasible, since the microbial biomarkers were found to be globally

generalizable, predictive, and generally disease-specific. However, the analyzed fecal metagenomic data represent a single view on the relationship between the microbiome and CRC, which has been analyzed under different aspects in other studies.

First, the fecal microbiome in CRC has also been investigated via 16S amplicon sequencing. The taxonomic enrichments reported in those studies largely agree with the results from our meta-analysis, but a thorough comparison has not been performed yet. Integrating these two types of data is challenging, however, since 16S amplicon sequencing provides limited taxonomic resolution, usually up to the level of genera, and since study effects appear to be more prominent in 16S amplicon data (Shah et al., 2018).

Secondly, several studies have subjected CRC tissue biopsies to 16S amplicon sequencing in order to get a better understanding of which microbes are adhering to or invading CRC tissue. In this setting, shotgun metagenomic experiments for the detection of microbes are made impractical by the large amount of human DNA in the sample, as typically more than 99% of reads are of human origin. In contrast to fecal samples, these tissue-derived datasets are not collected to explore potential microbiome-based diagnostic applications for CRC. Instead, they might more directly reflect biological processes linking the gut microbiome and colonic tumors. Additionally, a recent study searched for prognostic microbial biomarkers in CRC tissue and found *Fusobacterium* presence linked to shorter survival (Mima et al., 2016).

Lastly, human-targeted sequencing projects were recently found to contain microbial reads as well (Dohlman et al., 2021; Poore et al., 2020). In both publications, the authors analyzed data from the Cancer Genome Atlas (TCGA) project, which contains tumor samples from several hundred CRC cases (TCGA Research Network et al., 2013). Since the microbial reads represent a tiny fraction of all reads in a human-targeted sequencing sample, contamination and sequencing artifacts present substantial challenges for the identification of microbial taxa. Nonetheless, those analyses offer an alternative avenue to investigate the CRC microbiome and the newly developed microbial profiling methodology could therefore dramatically expand the scope of potential samples, especially when considering the scale of human-focussed sequencing projects such as TCGA (TCGA Research Network et al., 2013).

To gain a broader view of the role of the microbiome in CRC, I calculated statistical significance and effect sizes for enrichments of microbial genera in CRC using all of the data

types described above (see **Additional Methods**), effectively representing four meta-analyses encompassing 2.864 samples and 21 studies. All associations were calculated using LME models as a more versatile alternative for the blocked Wilcoxon test, since some data had to be adjusted not only for the study confounder but also for participant ID in studies with matched cancer and control tissue. The taxonomic profiles from (Poore et al., 2020) were excluded, since the calculated enrichments included many biologically implausible results, likely from contaminations or bioinformatic mis-annotations, highlighting again the challenges to identify microbial signals in human-centered sequencing projects (see **Additional Figure 2**).

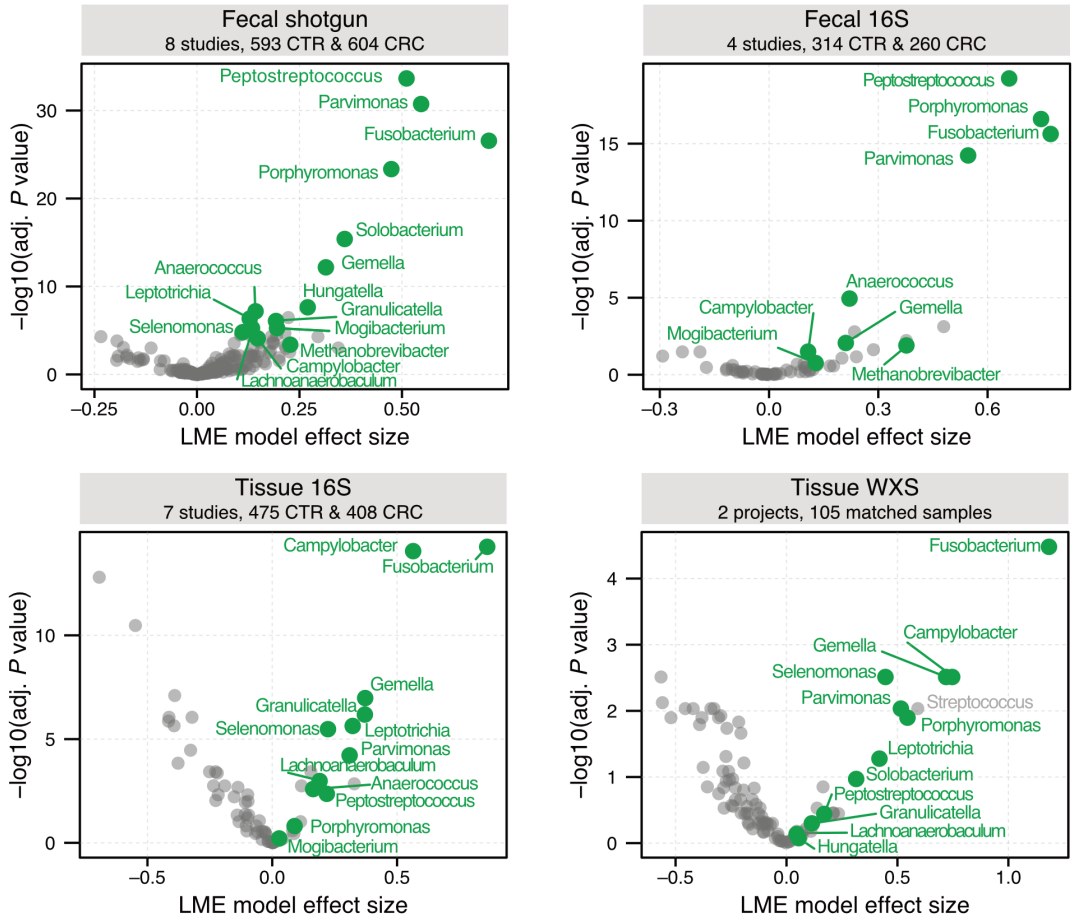


Figure 4: Different data types show similar bacterial enrichments for CRC
Volcano plots show the meta-analysis significance (FDR-adjusted *P* value) and effect sizes from linear mixed effect (LME) models on the basis of different data types (see **Additional Methods** for details). The top 15 genera (mean ranking across data types and CRC-enriched in at least two data types) are highlighted in green.

Overall, I found a striking similarity between the enrichment results in the remaining data types, with the most strongly CRC-associated genera being consistently enriched in all of the data types (see **Fig. 4**). Since both biological and technical variation between the

different data types is considerable, it could have been expected to obscure true biological signals as seen in the previously mentioned 16S meta-analysis (Shah et al., 2018). Nevertheless, the enrichments calculated here on different data types result in very similar rankings for the CRC-enriched genera, with only a few genera appearing as strongly CRC-enriched in a single data type. Differences between the enrichments seem to most strongly relate to the analyzed material: For example, stronger enrichments for the *Campylobacter* or the *Selenomonas* genus are found in both tissue-based analyses, whereas *Peptostreptococcus* is more strongly enriched in feces, independent of the sequencing approach. On the other hand, *Fusobacterium*, *Parvimonas*, and *Gemella* are strongly enriched across all sequenced materials and analytical methods.

The observed effect sizes and significance of these associations warrants further investigations of the biological role of bacteria in CRC: the most strongly CRC-enriched bacteria might serve as the basis for diagnostic tests and additionally, their role at the interface between tumor and gut microbiome needs to be explored in more detail to uncover causal relationships and additional biological processes involved in CRC carcinogenesis. For a subset of CRC-associated bacteria, for example *Fusobacterium* or *pks*-positive *Escherichia coli*, mechanistic studies have shed light on their specific biological role in CRC. For the majority of genera found in this meta-analysis, however, the mechanisms by which they might contribute to CRC is not well understood at all. Therefore, mechanistic studies, for example investigating *Gemella* or *Parvimonas* species, have the potential to uncover new biological processes and significantly expand our understanding of how bacteria interact with the host to promote carcinogenesis.

Recent technological developments have the potential to transform how we can study the interface between microbes and cancer. New multiplexed imaging technologies (Gyllborg et al., 2020; Sheth et al., 2019; Shi et al., 2020) will enable comprehensive mapping of the spatial organization of the microbe-host interface, providing important information about which bacteria adhere to or invade colonic epithelial cells. On the other hand, the study of both host gene expression or genomic alterations in the tumor and at the same time bacterial presence will be made possible either through cross-species RNA sequencing approaches (Westermann and Vogel, 2021) or through in-depth analysis of bacterial components in human-targeted sequencing projects (Dohlman et al., 2021; Poore et al., 2020).

Cross-study application of machine learning models in microbiome research

As part of the SIAMCAT publication (Wirbel et al., 2021), a machine learning meta-analysis was performed to gain insights into the performance of different machine learning algorithms and the influence of preprocessing steps on the results of the workflow. In a second step, the application of disease models across datasets was to be explored, since cross-study application of models, although crucial to establish validity, had not been evaluated systematically in a larger set of studies yet.

Machine learning meta-analysis

Previous studies had made recommendations regarding machine learning algorithms and data preprocessing, but were limited by the small number of analyzed datasets (Pasolli et al., 2016) or by focussing only on data generated using a single pipeline (Duvallet et al., 2017). In the SIAMCAT publication, over 130 classification tasks derived from 50 metagenomic studies were included in a machine learning meta-analysis (see **Fig. 5a**). The raw data had been profiled with technically heterogeneous pipelines, including the RDP classifier (Wang et al., 2007) for 16S rRNA data, mOTUs2 (Milanese et al., 2019) and MetaPhlAn2 (Truong et al., 2015) for the taxonomic and eggNOG4.5 (Huerta-Cepas et al., 2016) and HUMAnN2 (Franzosa et al., 2018) for the functional profiling of shotgun metagenomic studies. For each classification task, over 7000 distinct parameter combinations were explored for the machine learning workflow, including variations in the learning algorithm, the normalization method, and filtering parameters.

Other than biological differences between diseases, the learning algorithm and the choice of normalization method had the biggest influence on the resulting model performance. However, when comparing the best-performing parameter settings, the three included machine learning algorithms, LASSO (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005), and random forest (Tin Kam Ho, 1995), showed overall similar performance, with the LASSO and the Elastic Net having a slight edge over the random forest. This edge was only realized, however, when data were appropriately normalized.

In summary, the results suggest using the Elastic Net algorithm in combination with a suitable normalization method (total sum scaling followed by z-scoring) for optimal performance. Since a large number of technically heterogeneous datasets were included in this meta-analysis, the recommendations derived from it are expected to be more robust

than previous efforts that had advised using the random forest algorithm (Pasolli et al., 2016).

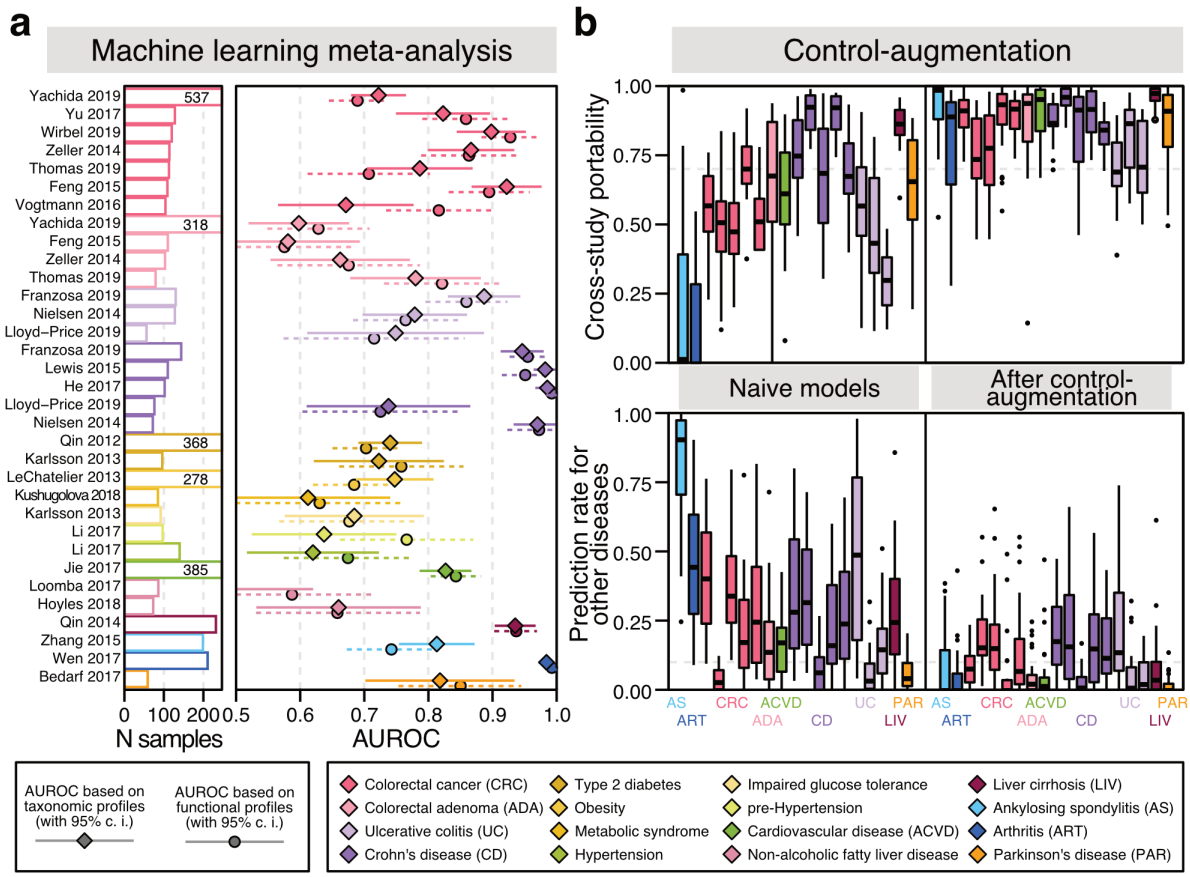


Figure 5: Machine learning meta-analysis and cross-study application of models
a) As part of the SIAMCAT publication, a large set of case-control metagenomic studies were collected for a machine learning meta-analysis. The figure shows the AUC values derived with the best-performing parameter set for datasets that were profiled taxonomically with mOTUs2 and functionally with eggNOG4.5. **b)** Using the datasets from the machine learning meta-analysis, cross-study application of models was explored systematically, revealing low accuracy and disease-specificity for naively transferred models. In contrast to that, control-augmented models (see text for explanation) showed uniform improvements in both measures. Figure adapted from (Wirbel et al., 2021).

Extensive loss of accuracy and disease-specificity in cross-study application

A critical advantage of machine learning models is that they can be applied to completely external data to make predictions, including datasets investigating a different disease. For example, the single-study models in the CRC meta-analysis could be applied to datasets from IBD and type 2 diabetes, showing elevated false positive detections for those other diseases, which was alleviated in the LOSO setting. Although these types of analyses are important to evaluate the validity of biomarkers across different populations and technical

heterogeneity, cross-study application of machine learning models is not routinely investigated in the microbiome literature.

The machine learning meta-analysis included in the SIAMCAT publication presented the ideal setting to systematically study how models behave in cross-study application. To do so, models trained with the best-performing parameter sets were applied to all other datasets in the meta-analysis and the predictions were recorded. In contrast to same-disease analyses (as in the CRC meta-analysis, for example), the evaluation of cross-study application is not straightforward across different diseases, since true positives are lacking for a ROC analysis, because the classifiers were trained to recognize another disease (positive class) than is present in the other studies. Therefore, two measures were calculated to answer different questions. First, the cross-study portability asks if a model can be applied to another dataset and retain its accuracy by evaluating the separation between cases from the data set the classifier was cross-validated on and controls from the external data set. The second measure is the prediction rate for other diseases and effectively measures the disease-specificity of the classifier.

The evaluations show an extensive loss of accuracy and low disease specificity in almost all analyzed models (see **Fig. 5b**), manifesting in low cross-study portability scores and high prediction rates for other diseases. In the most extreme case of the model for ankylosing spondylitis, a median of 90% false positive predictions for the diseased cases from other datasets was recorded. These results suggest that machine learning model transfer across different datasets is extremely challenging, which motivated modifications of the machine learning workflows.

One reason for these observations could be that technical differences between studies dominate most, and therefore also predictive, features, precluding the naive transfer of machine learning models. Correction techniques aiming to normalize out some of the observed study differences, for example based on quantile normalization, might be applicable here, but they are not well explored for microbiome data. Moreover, these procedures often use the label information for correction to avoid normalizing out the disease signal as well (Gibbons et al., 2018; Leek and Storey, 2007), which again opens up the possibility for overfitting across datasets and thus complicates their evaluations. Additionally, studies might differ in terms of the study population, for example, age structure of the participants, and other life-style, co-morbidity or medication-related

covariates, which would represent real biological differences. Lastly, other sources of heterogeneity might be how studies define diseased and healthy participants (Thompson, 1994), further impeding model transfer due to label noise.

Control-augmentation as strategy to improve cross-study application

These results indicate that the naive transfer of machine learning models across studies is extremely challenging, if not impossible. However, the machine learning exploration in the CRC meta-analysis had shown that pooling of data improved both disease specificity as well as model accuracy. These results therefore motivated the hypothesis that inclusion of data from multiple, technically heterogeneous studies would enable the classifiers to learn more robust and broadly applicable models, which in turn minimizes overfitting and improves disease specificity; similar observations have been made in other research fields as well (Zhang et al., 2021). In the general setting of cross-study application, data pooling across studies is not always possible, since sometimes only a single dataset is available for a given disease, which makes it impossible to pool the cases from multiple studies. Therefore, augmentation with external controls was explored as a strategy to improve cross-study transfer of machine learning models. Again the idea here being that such a classifier could learn a more precise disease representation to be discriminated from a larger, more heterogeneous pool of control samples in the training set.

For the control-augmentation, external controls from several cohort studies without disease signal (such as (Schirmer et al., 2016; Xie et al., 2016; Zeevi et al., 2015)) were randomly added during model training within the cross-validation. Indeed, control-augmented models were found to have greatly improved cross-study portability as well as disease-specificity, uniformly improving both measures across all investigated models (see **Fig. 5b**).

The control-augmentation strategy is a first attempt to alleviate the problems with cross-study applications that were uncovered in this meta-analysis. The best parameters for control-augmentation need to be explored in more detail, e.g. how many external studies are needed for robust improvements. Additionally, since the definition of “control” samples can vary dramatically across studies (Thompson, 1994), control-augmentation has the potential to bias the resulting models, for example by introducing label noise. However,

in the presented study, measures for cross-study applications improved consistently -- independent of the choice of datasets used for the control-augmentation.

Other microbiome meta-analyses highlighted a general dysbiosis signal that was common to multiple diseases (Duvallet et al., 2017) and even devised a classifier for a “healthy” and “diseased” microbiome (Gupta et al., 2020). While naive models also showed a high level of cross-disease classification in the presented study, control-augmented models were generally rather disease-specific, indicating that different diseases can be associated with distinct microbial biomarkers.

In summary, the control-augmentation strategy greatly improves both accuracy and disease specificity of machine learning models in cross-study application, allowing for the investigation of microbial biomarkers specific to a single disease as opposed to ones that are more unspecific and shared across different diseases.

Concluding Remarks

The goal of my doctoral studies was to advance the methodology for comparative statistical analysis in clinical microbiome studies and apply these tools to investigate the fecal microbiome in colorectal cancer in depth. The key outcomes can be summarized as follows.

I first developed the necessary statistical methods for comparative metagenomic studies, and made them available to the community in the form of the SIAMCAT R package as a user-friendly toolbox for statistical and machine learning analysis of microbiome data. To guide the choice of microbiome analytics methods, I broadly validated differential abundance testing methods through an unbiased and realistic benchmark that could mirror study confounding; I furthermore extensively validated the machine learning workflows and their parameter settings and made them robust against commonly encountered statistical flaws in their design and evaluation.

I then applied those methods to fecal shotgun metagenomic studies of CRC for an in-depth characterization of microbiome alterations in this common cancer type, revealing taxonomic and functional biomarkers as well as globally predictive and disease-specific CRC signatures. I additionally explored virulence factors that might contribute to carcinogenesis, such as bile acid conversion enzymes. Following up on these findings through meta-analyses with other data types, I established a solid foundation for ongoing and future explorations of the microbiome-host interface in CRC.

Lastly, I broadened the focus of machine learning applications in a meta-analysis encompassing not a single but multiple diseases, uncovering substantial challenges for naive cross-study application of microbiome-based machine learning models. Motivated by the observations from the CRC meta-analysis, I empirically validated a novel strategy of dataset augmentation with external control samples as an effective means to improve model transfer across studies effectively reducing their propensity to make false-positive predictions due to between-study differences (of both technical and biological nature) and substantially improving their disease specificity.

Acknowledgments

This effort would have been impossible without all the people around me. Therefore, I want to thank everyone who made my scientific journey possible by helping and supporting me throughout the last years.

First and foremost, I want to thank **Georg** for -- everything. Your guidance and input were invaluable to me at every step and I thoroughly enjoyed my experience of working with you. I cannot even put into words all the things that I learned from you.

I also want to thank everyone who is and was in the **Zeller team** at various points in time. The atmosphere in the group was always fantastic and every day I was happy to come to work with you. Special thanks of course to **Nic** for his good friendship and company in all the bouldering sessions.

I would like to thank everyone at **EMBL**, especially in the Bork, Typas, and Zimmermann group, for stimulating conversations about science and fun beer sessions in the staff lounge.

None of the publications included in this document would have been possible without **collaborators** that shared their data, their insights, and their time. Thank you all, in particular **Morgan**, **Lisa**, and **Peer**, for a wonderful experience in the scientific environment!

I especially want to thank **Daniel Huson** and **Stephan Ossowski** for taking the time and effort to assess my work.

I also want to thank all the friends that I made at EMBL, especially **Tim**, **Pamela**, **Holly**, and **Martine**. Thank you for being amazing human beings and for your friendship.

Vielen Dank auch an meine **Familie**, meine Eltern, meine Geschwister und Neffen, die ich nicht so oft gesehen habe, wie ich oder sie es gerne gewollte hätten. Danke, dass ihr immer für mich da seid!

Vielen Dank auch an **Matze, Zoey,** und **Vlad** für die vielen Abende mit gutem asiatischen Essen und exzellenten Wein bei euch in der Küche oder auf eurem Balkon.

Zu guter Letzt geht mein Dank natürlich auch an **Jens, Jasper**, und **Simon**. Für alle ZEIT Kreuzworträtsel-Abende, die mich durch die dunklen Tage der Pandemie gebracht haben, für das Stranden mit Bier in Holland, für alle Gespräche über die Wissenschaft, die mich immer wieder neu motiviert haben, und für alle Gespräche über jedwede andere Themen.

References

Allison, J.E., Feldman, R., and Tekawa, I.S. (1990). Hemoccult screening in detecting colorectal neoplasm: sensitivity, specificity, and predictive value. Long-term follow-up in a large group practice setting. Ann. Intern. Med. *112*, 328–333.

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol.

Ananthakrishnan, A.N. (2011). Clostridium difficile infection: epidemiology, risk factors and management. Nat. Rev. Gastroenterol. Hepatol. *8*, 17–26.

Bang, S., Yoo, D., Kim, S.-J., Jhang, S., Cho, S., and Kim, H. (2019). Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. Sci. Rep. *9*, 10189.

Baruch, E.N., Youngster, I., Ben-Betzalel, G., Ortenberg, R., Lahat, A., Katz, L., Adler, K., Dick-Necula, D., Raskin, S., Bloch, N., et al. (2021). Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. Science *371*, 602–609.

Baxter, N.T., Ruffin, M.T., 4th, Rogers, M.A.M., and Schloss, P.D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. *8*, 37.

Bedarf, J.R., Hildebrand, F., Coelho, L.P., Sunagawa, S., Bahram, M., Goeser, F., Bork, P., and Wüllner, U. (2017). Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. Genome Med. *9*, 39.

Belkaid, Y., and Hand, T.W. (2014). Role of the microbiota in immunity and inflammation. Cell *157*, 121–141.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857.

Brennan, C.A., and Garrett, W.S. (2019). Fusobacterium nucleatum - symbiont, opportunist and oncobacterium. Nat. Rev. Microbiol. *17*, 156–166.

Buchka, S., Hapfelmeier, A., Gardner, P.P., Wilson, R., and Boulesteix, A.-L. (2021). On the optimistic performance evaluation of newly introduced bioinformatic methods. Genome Biol. *22*, 152.

Bullman, S., Pedamallu, C.S., Sicinska, E., Clancy, T.E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science *358*, 1443–1448.

Burns, M.B., Lynch, J., Starr, T.K., Knights, D., and Blekhman, R. (2015). Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. Genome Med. *7*, 55.

Calgaro, M., Romualdi, C., Waldron, L., Risso, D., and Vitulo, N. (2020). Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. Genome Biol. *21*, 191.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods *13*, 581–583.

Cani, P.D. (2017). Gut microbiota - at the intersection of everything? Nat. Rev. Gastroenterol. Hepatol. *14*, 321–322.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., and Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc.

Natl. Acad. Sci. U. S. A. *108 Suppl 1*, 4516–4522.

Casadevall, A., and Pirofski, L.-A. (2000). Host-Pathogen Interactions: Basic Concepts of Microbial Commensalism, Colonization, Infection, and Disease. Infect. Immun. *68*, 6511–6518.

Castellarin, M., Warren, R.L., Freeman, J.D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R.A., et al. (2012). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. *22*, 299–306.

Colman, R.J., and Rubin, D.T. (2014). Fecal microbiota transplantation as therapy for inflammatory bowel disease: a systematic review and meta-analysis. J. Crohns. Colitis *8*, 1569–1581.

Costea, P.I., Coelho, L.P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017a). Subspecies in the global human gut microbiome. Mol. Syst. Biol. *13*, 960.

Costea, P.I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., Tramontano, M., Driessen, M., Hercog, R., Jung, F.-E., et al. (2017b). Towards standards for human fecal sample processing in metagenomic studies. Nat. Biotechnol. *35*, 1069–1076.

Dai, Z., Coker, O.O., Nakatsu, G., Wu, W.K.K., Zhao, L., Chen, Z., Chan, F.K.L., Kristiansen, K., Sung, J.J.Y., Wong, S.H., et al. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome *6*, 70.

Davar, D., Dzutsev, A.K., McCulloch, J.A., Rodrigues, R.R., Chauvin, J.-M., Morrison, R.M., Deblasio, R.N., Menna, C., Ding, Q., Pagliano, O., et al. (2021). Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. Science *371*, 595–602.

Dejea, C.M., Fathi, P., Craig, J.M., Boleij, A., Taddese, R., Geis, A.L., Wu, X., DeStefano Shields, C.E., Hechenbleikner, E.M., Huso, D.L., et al. (2018). Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science *359*, 592–597.

Denkert, C., Budczies, J., Weichert, W., Wohlgemuth, G., Scholz, M., Kind, T., Niesporek, S., Noske, A., Buckendahl, A., Dietel, M., et al. (2008). Metabolite profiling of human colon carcinoma--deregulation of TCA cycle and amino acid turnover. Mol. Cancer *7*, 72.

Dohlman, A.B., Arguijo Mendoza, D., Ding, S., Gao, M., Dressman, H., Iliev, I.D., Lipkin, S.M., and Shen, X. (2021). The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. Cell Host Microbe *29*, 281–298.e5.

Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., and Alm, E.J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat. Commun. *8*, 1784.

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat. Commun. *6*, 6528.

Flemer, B., Warren, R.D., Barrett, M.P., Cisek, K., Das, A., Jeffery, I.B., Hurley, E., O'Riordain, M., Shanahan, F., and O'Toole, P.W. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. Gut *67*, 1454–1463.

Fluckiger, A., Daillère, R., Sassi, M., Sixt, B.S., Liu, P., Loos, F., Richard, C., Rabu, C., Alou, M.T., Goubet, A.-G., et al. (2020). Cross-reactivity between tumor MHC class I–restricted antigens and an enterococcal bacteriophage. Science.

Flynn, K.J., Baxter, N.T., and Schloss, P.D. (2016). Metabolic and Community Synergy of Oral Bacteria in Colorectal Cancer. mSphere *1*.

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, S., Gudmundsdottir, V., Pedersen, H.K., et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature *528*, 262–266.

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods *15*, 962–968.

Geller, L.T., Barzily-Rokni, M., Danino, T., Jonas, O.H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z.A., Shee, K., et al. (2017). Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. Science *357*, 1156–1160.

Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe *15*, 382–392.

Gibbons, S.M., Duvallet, C., and Alm, E.J. (2018). Correcting for batch effects in case-control microbiome studies. PLoS Comput. Biol. *14*, e1006102.

Goedert, J.J., Sampson, J.N., Moore, S.C., Xiao, Q., Xiong, X., Hayes, R.B., Ahn, J., Shi, J., and Sinha, R. (2014). Fecal metabolomics: assay performance and association with colorectal cancer. Carcinogenesis *35*, 2089–2096.

Gopalakrishnan, V., Spencer, C.N., Nezi, L., Reuben, A., Andrews, M.C., Karpinets, T.V., Prieto, P.A., Vicente, D., Hoffman, K., Wei, S.C., et al. (2018). Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science *359*, 97–103.

Guo, S.-H., Wang, H.-F., Nian, Z.-G., Wang, Y.-D., Zeng, Q.-Y., and Zhang, G. (2017). Immunization with alkyl hydroperoxide reductase subunit C reduces Fusobacterium nucleatum load in the intestinal tract. Sci. Rep. *7*, 10566.

Gupta, V.K., Kim, M., Bakshi, U., Cunningham, K.Y., Davis, J.M., 3rd, Lazaridis, K.N., Nelson, H., Chia, N., and Sung, J. (2020). A predictive index for health status using species-level gut microbiome profiling. Nat. Commun. *11*, 4635.

Gyllborg, D., Langseth, C.M., Qian, X., Choi, E., Salas, S.M., Hilscher, M.M., Lein, E.S., and Nilsson, M. (2020). Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. Nucleic Acids Res. *48*, e112.

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2019). A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief. Bioinform. *20*, 210–221.

He, Y., Wu, W., Zheng, H.-M., Li, P., McDonald, D., Sheng, H.-F., Chen, M.-X., Chen, Z.-H., Ji, G.-Y., Zheng, Z.-D.-X., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat. Med. *24*, 1532–1535.

Hirayama, A., Kami, K., Sugimoto, M., Sugawara, M., Toki, N., Onozuka, H., Kinoshita, T., Saito, N., Ochiai, A., Tomita, M., et al. (2009). Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. Cancer Res. *69*, 4918–4925.

HMP Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Hothorn, T., Hornik, K., van de Wiel, M.A., and Zeileis, A. (2006). A Lego System for Conditional Inference. Am. Stat. *60*, 257–263.

Hoyles, L., Fernández-Real, J.-M., Federici, M., Serino, M., Abbott, J., Charpentier, J., Heymes, C., Luque, J.L., Anthony, E., Barton, R.H., et al. (2018). Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. Nat. Med. *24*, 1070–1080.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods *12*, 115–121.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. *44*, D286–D293.

Imhann, F., Bonder, M.J., Vich Vila, A., Fu, J., Mujagic, Z., Vork, L., Tigchelaar, E.F., Jankipersadsing, S.A., Cenit, M.C., Harmsen, H.J.M., et al. (2016). Proton pump inhibitors affect the gut microbiome. Gut *65*, 740–748.

Jackson, M.A., Goodrich, J.K., Maxan, M.-E., Freedberg, D.E., Abrams, J.A., Poole, A.C., Sutter, J.L., Welter, D., Ley, R.E., Bell, J.T., et al. (2016). Proton pump inhibitors alter the composition of the gut microbiota. Gut *65*, 749–756.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. *42*, D199–D205.

Karcher, N., Pasolli, E., Asnicar, F., Huang, K.D., Tett, A., Manara, S., Armanini, F., Bain, D., Duncan, S.H., Louis, P., et al. (2020). Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol. *21*, 138.

Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99–103.

Knights, D., Costello, E.K., and Knight, R. (2011a). Supervised classification of human microbiota. FEMS Microbiol. Rev. *35*, 343–359.

Knights, D., Parfrey, L.W., Zaneveld, J., Lozupone, C., and Knight, R. (2011b). Human-associated microbial signatures: examining their predictive value. Cell Host Microbe *10*, 292–296.

Kostic, A.D., Gevers, D., Pedamallu, C.S., Michaud, M., Duke, F., Earl, A.M., Ojesina, A.I., Jung, J., Bass, A.J., Tabernero, J., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. *22*, 292–298.

Kostic, A.D., Chun, E., Robertson, L., Glickman, J.N., Gallini, C.A., Michaud, M., Clancy, T.E., Chung, D.C., Lochhead, P., Hold, G.L., et al. (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe *14*, 207–215.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins *87*, 1011–1020.

Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, Articles *82*, 1–26.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.

Le Goallec, A., Tierney, B.T., Luber, J.M., Cofer, E.M., Kostic, A.D., and Patel, C.J. (2020). A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. PLoS Comput. Biol. *16*, e1007895.

Lin, H., and Peddada, S.D. (2020). Analysis of compositions of microbiomes with bias correction. Nat. Commun. *11*, 3514.

Liu, P.-F., Haake, S.K., Gallo, R.L., and Huang, C.-M. (2009). A novel vaccine targeting Fusobacterium nucleatum against abscesses and halitosis. Vaccine *27*, 1589–1595.

Loomba, R., Seguritan, V., Li, W., Long, T., Klitgord, N., Bhatt, A., Dulai, P.S., Caussy, C., Bettencourt, R., Highlander, S.K., et al. (2017). Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab. *25*, 1054–1062.e5.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lozupone, C.A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J.K., Gordon, J.I., and Knight, R. (2013). Meta-analyses of studies of the human microbiota. Genome Res. *23*, 1704–1714.

Lynch, S.V., and Pedersen, O. (2016). The Human Intestinal Microbiome in Health and Disease. N. Engl. J. Med. *375*, 2369–2379.

Mager, L.F., Burkhard, R., Pett, N., Cooke, N.C.A., Brown, K., Ramay, H., Paik, S., Stagg, J., Groves, R.A., Gallo, M.,

et al. (2020). Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy. Science *369*, 1481–1489.

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E.E., Brochado, A.R., Fernandez, K.C., Dose, H., Mori, H., et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. Nature *555*, 623–628.

Mal, M., Koh, P.K., Cheah, P.Y., and Chan, E.C.Y. (2012). Metabotyping of human colorectal cancer using two-dimensional gas chromatography mass spectrometry. Anal. Bioanal. Chem. *403*, 483–493.

Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., and Peddada, S.D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Health Dis. *26*, 27663.

Matias Rodrigues, J.F., Schmidt, T.S.B., Tackmann, J., and von Mering, C. (2017). MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics *33*, 3808–3810.

Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J.J., and Gajewski, T.F. (2018). The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. Science *359*, 104–108.

McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems *3*.

McLaren, M.R., Willis, A.D., and Callahan, B.J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. Elife *8*, e46923.

McMurdie, P.J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput. Biol. *10*, e1003531.

Merino, N., Aronson, H.S., Bojanova, D.P., Feyhl-Buska, J., Wong, M.L., Zhang, S., and Giovannelli, D. (2019). Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. Front. Microbiol. *10*, 780.

Milanese, A., Mende, D.R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., Hingamp, P., Alves, R., Costea, P.I., Coelho, L.P., et al. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. Nat. Commun. *10*, 1014.

Mima, K., Nishihara, R., Qian, Z.R., Cao, Y., Sukawa, Y., Nowak, J.A., Yang, J., Dou, R., Masugi, Y., Song, M., et al. (2016). Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. Gut *65*, 1973–1980.

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K., and Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. Nat. Commun. *10*, 2719.

Nakatsu, G., Li, X., Zhou, H., Sheng, J., Wong, S.H., Wu, W.K.K., Ng, S.C., Tsoi, H., Dong, Y., Zhang, N., et al. (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. Nat. Commun. *6*, 8727.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510.

Nearing, J.T., Douglas, G.M., Hayes, M.G., and MacDonald, J. (2021). Microbiome differential abundance methods produce disturbingly different results across 38 datasets. bioRxiv.

Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L.T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. Science *368*, 973–980.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. *32*, 822–828.

van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E.G., de Vos, W.M., Visser, C.E., Kuijper, E.J.,

Bartelsman, J.F.W.M., Tijssen, J.G.P., et al. (2013). Duodenal infusion of donor feces for recurrent Clostridium difficile. N. Engl. J. Med. *368*, 407–415.

Nougayrède, J.-P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., Buchrieser, C., Hacker, J., Dobrindt, U., and Oswald, E. (2006). Escherichia coli induces DNA double-strand breaks in eukaryotic cells. Science *313*, 848–851.

Oellgaard, J., Winther, S.A., Hansen, T.S., Rossing, P., and von Scholten, B.J. (2017). Trimethylamine N-oxide (TMAO) as a New Potential Therapeutic Target for Insulin Resistance and Cancer. Curr. Pharm. Des. *23*, 3699–3712.

Ogino, S., Nowak, J.A., Hamada, T., Phipps, A.I., and Peters, U. (2018). Integrative analysis of exogenous, endogenous, tumour and immune factors for precision medicine. Gut.

Pasolli, E., Truong, D.T., Malik, F., Waldron, L., and Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput. Biol. *12*, e1004977.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20.

Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. Nat. Methods *10*, 1200–1202.

Payne, C.M., Weber, C., Crowley-Skillicorn, C., Dvorak, K., Bernstein, H., Bernstein, C., Holubec, H., Dvorakova, B., and Garewal, H. (2007). Deoxycholate induces mitochondrial oxidative stress and activates NF-kappaB through multiple mechanisms in HCT-116 colon epithelial cells. Carcinogenesis *28*, 215–222.

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H.M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P.B., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature *580*, 269–273.

Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S.J., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature *579*, 567–574.

Poyet, M., Groussin, M., Gibbons, S.M., Avila-Pacheco, J., Jiang, X., Kearney, S.M., Perrotta, A.R., Berdy, B., Zhao, S., Lieberman, T.D., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat. Med. *25*, 1442–1452.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Prill, R.J., Saez-Rodriguez, J., Alexopoulos, L.G., Sorger, P.K., and Stolovitzky, G. (2011). Crowdsourcing network inference: the DREAM predictive signaling network challenge. Sci. Signal. *4*, mr7.

Purcell, R.V., Visnovska, M., Biggs, P.J., Schmeier, S., and Frizelle, F.A. (2017). Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Sci. Rep. *7*, 11590.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature *513*, 59–64.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. *41*, D590–D596.

Quinn, T.P. (2021). Stool Studies Don't Pass the Sniff Test: A Systematic Review of Human Gut Microbiome Research Suggests Widespread Misuse of Machine Learning.

Rhee, S.H., Pothoulakis, C., and Mayer, E.A. (2009). Principles and clinical implications of the brain–gut–enteric microbiota axis. Nat. Rev. Gastroenterol. Hepatol. *6*, 306–314.

Ridlon, J.M., Harris, S.C., Bhowmik, S., Kang, D.-J., and Hylemon, P.B. (2016). Consequences of bile salt biotransformations by intestinal bacteria. Gut Microbes *7*, 22–39.

Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., Sahin, I., Chandra, V., San Lucas, A., et al. (2019). Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. Cell *178*, 795–806.e12.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography *40*, 913–929.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Routy, B., Le Chatelier, E., Derosa, L., Duong, C.P.M., Alou, M.T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M.P., et al. (2018). Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science *359*, 91–97.

Rubinstein, M.R., Wang, X., Liu, W., Hao, Y., Cai, G., and Han, Y.W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell Host Microbe *14*, 195–206.

Scheperjans, F., Aho, V., Pereira, P.A.B., Koskinen, K., Paulin, L., Pekkonen, E., Haapaniemi, E., Kaakkola, S., Eerola-Rautio, J., Pohja, M., et al. (2015). Gut microbiota are related to Parkinson's disease and clinical phenotype. Mov. Disord. *30*, 350–358.

Schirmer, M., Smeekens, S.P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E.A., Horst, R.T., Jansen, T., Jacobs, L., Bonder, M.J., et al. (2016). Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell *167*, 1897.

Schirmer, M., Franzosa, E.A., Lloyd-Price, J., McIver, L.J., Schwager, R., Poon, T.W., Ananthakrishnan, A.N., Andrews, E., Barron, G., Lake, K., et al. (2018). Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nat Microbiol *3*, 337–346.

Schmidt, T.S.B., Raes, J., and Bork, P. (2018). The Human Gut Microbiome: From Association to Modulation. Cell *172*, 1198–1215.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nat. Methods *14*, 1063–1071.

Sears, C.L., and Garrett, W.S. (2014). Microbes, microbiota, and colon cancer. Cell Host Microbe *15*, 317–328.

Shah, M.S., DeSantis, T.Z., Weinmaier, T., McMurdie, P.J., Cope, J.L., Altrichter, A., Yamal, J.-M., and Hollister, E.B. (2018). Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. Gut *67*, 882–891.

Sheth, R.U., Li, M., Jiang, W., Sims, P.A., Leong, K.W., and Wang, H.H. (2019). Spatial metagenomic characterization of microbial biogeography in the gut. Nat. Biotechnol. *37*, 877–883.

Shi, H., Shi, Q., Grodner, B., Lenz, J.S., Zipfel, W.R., Brito, I.L., and De Vlaminck, I. (2020). Highly multiplexed spatial mapping of microbial communities. Nature *588*, 676–681.

Sinha, R., The Microbiome Quality Control Project Consortium, Abu-Ali, G., Vogtmann, E., Fodor, A.A., Ren, B.,

Amir, A., Schwager, E., Crabtree, J., Ma, S., et al. (2017). Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nature Biotechnology *35*, 1077–1086.

Smialowski, P., Frishman, D., and Kramer, S. (2010). Pitfalls of supervised feature selection. Bioinformatics *26*, 440–443.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. Science *348*.

TCGA Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., 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.

Thompson, S.G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. BMJ *309*, 1351–1355.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Series B Stat. Methodol. *58*, 267–288.

Tin Kam Ho (1995). Random decision forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition, (ieeexplore.ieee.org), pp. 278–282 vol.1.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

Tsilimigras, M.C.B., and Fodor, A.A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann. Epidemiol. *26*, 330–335.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. Nature *551*, 507–511.

Vieira-Silva, S., Falony, G., Darzi, Y., and Lima-Mendez, G. (2016). Species–function relationships shape ecological properties of the human gut microbiome. Nature.

Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A.Y., Hercog, R., Goedert, J.J., Shi, J., Bork, P., and Sinha, R. (2016). Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PLoS One *11*, e0155362.

Voigt, A.Y., Costea, P.I., Kultima, J.R., Li, S.S., Zeller, G., Sunagawa, S., and Bork, P. (2015). Temporal and technical variability of human gut metagenomes. Genome Biol. *16*, 73.

Vujkovic-Cvijin, I., Sklar, J., Jiang, L., Natarajan, L., Knight, R., and Belkaid, Y. (2020). Host variables confound gut microbiota studies of human disease. Nature *587*, 448–454.

Wang, J., Kurilshikov, A., Radjabzadeh, D., Turpin, W., Croitoru, K., Bonder, M.J., Jackson, M.A., Medina-Gomez, C., Frost, F., Homuth, G., et al. (2018). Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. Microbiome *6*, 101.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

WCRFI (2018). Diet, nutrition, physical activity and cancer: a global perspective: a summary of the Third Expert Report (World Cancer Research Fund International).

Weir, T.L., Manter, D.K., Sheflin, A.M., Barnett, B.A., Heuberger, A.L., and Ryan, E.P. (2013). Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. PLoS One *8*, e70803.

Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R.,

Vázquez-Baeza, Y., Birmingham, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome *5*, 27.

Westermann, A.J., and Vogel, J. (2021). Cross-species RNA-seq for deciphering host–microbe interactions. Nat. Rev. Genet. *22*, 361–378.

Wilson, M.R., Jiang, Y., Villalta, P.W., Stornetta, A., Boudreau, P.D., Carrá, A., Brennan, C.A., Chun, E., Ngo, L., Samson, L.D., et al. (2019). The human gut bacterial genotoxin colibactin alkylates DNA. Science *363*.

Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. *25*, 679–689.

Wirbel, J., Zych, K., Essex, M., Karcher, N., Kartal, E., Salazar, G., Bork, P., Sunagawa, S., and Zeller, G. (2021). Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol. *22*, 93.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. *74*, 5088–5090.

Wu, H., Esteve, E., Tremaroli, V., Khan, M.T., Caesar, R., Mannerås-Holm, L., Ståhlman, M., Olsson, L.M., Serino, M., Planas-Fèlix, M., et al. (2017). Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. Nat. Med. *23*, 850–858.

Wu, S., Rhee, K.-J., Albesiano, E., Rabizadeh, S., Wu, X., Yen, H.-R., Huso, D.L., Brancati, F.L., Wick, E., McAllister, F., et al. (2009). A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nat. Med. *15*, 1016–1022.

Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., Ward, K.J., Jackson, M.A., Xia, Y., Chen, X., et al. (2016). Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. Cell Syst *3*, 572–584.e3.

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat. Med. *25*, 968–976.

Yoshimoto, S., Loo, T.M., Atarashi, K., Kanda, H., Sato, S., Oyadomari, S., Iwakura, Y., Oshima, K., Morita, H., Hattori, M., et al. (2013). Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. Nature *499*, 97–101.

Yu, J., Feng, Q., Wong, S.H., Zhang, D., Liang, Q.Y., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut *66*, 70–78.

Zackular, J.P., Rogers, M.A.M., Ruffin, M.T., 4th, and Schloss, P.D. (2014). The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev. Res. *7*, 1112–1121.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized Nutrition by Prediction of Glycemic Responses. Cell *163*, 1079–1094.

Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. *10*, 766.

Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., Wu, X., Li, J., Tang, L., Li, Y., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat. Med. *21*, 895–905.

Zhang, Y., Patil, P., Johnson, W.E., and Parmigiani, G. (2021). Robustifying genomic classifiers to batch effects via ensemble learning. Bioinformatics *37*, 1521–1527.

Zhao, Q., Adeli, E., and Pohl, K.M. (2020). Training confounder-free deep learning models for medical

applications. Nat. Commun. *11*, 6010.

Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A.L. (2019). Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature *570*, 462–467.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. *67*, 301–320.

Appendix

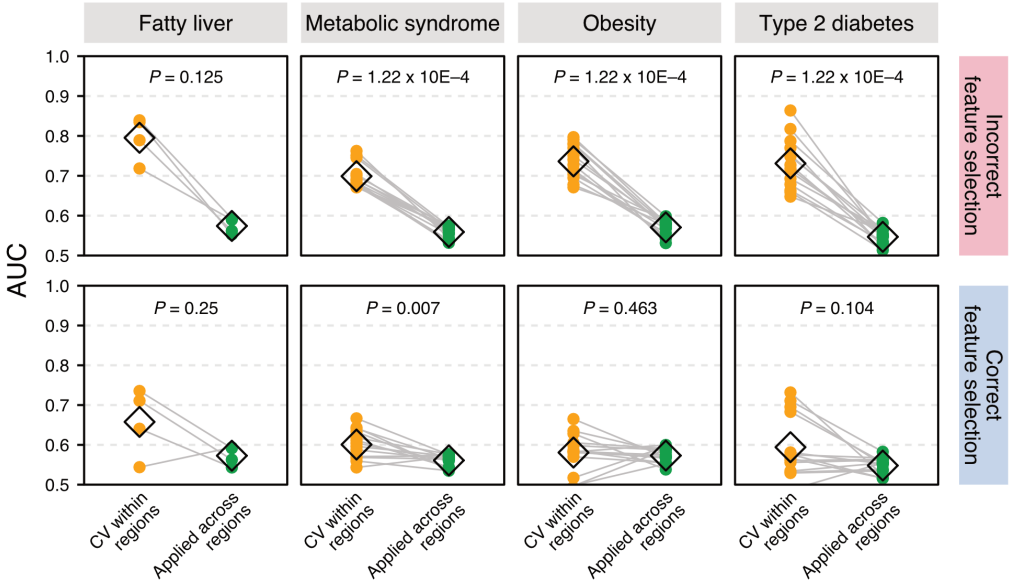
Additional Methods & Figures

Methods for the CRC meta-analysis using different data types

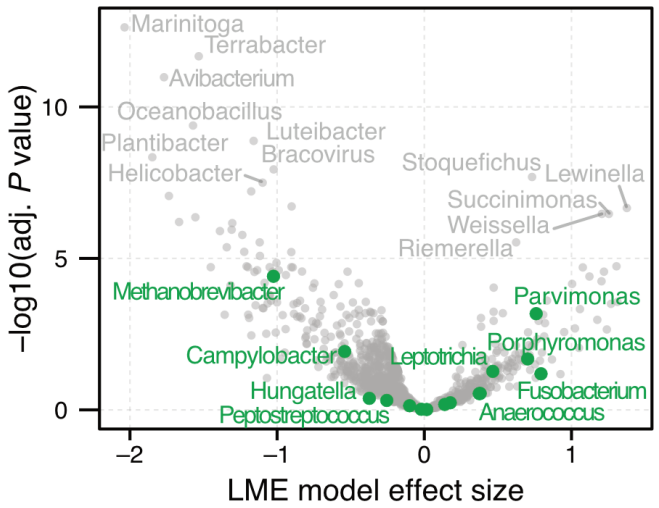
To investigate the link between CRC and the gut microbiome in different data types, taxonomic profiles were generated from different studies. The data from fecal shotgun metagenomic studies were processed as described in (Wirbel et al., 2019) and taxonomically profiled with mOTUs2.5 (Milanese et al., 2019). Data from fecal and tissue 16S studies were processed with DADA2 (Callahan et al., 2016) and the resulting amplicon sequence variants were taxonomically annotated with MAPseq v1.2.6 (Matias Rodrigues et al., 2017). Taxonomic profiles generated from TCGA whole exome sequencing (WXS) and total RNA-seq data were downloaded from the supplementary material from (Dohlman et al., 2021) and (Poore et al., 2020), respectively. The taxonomic profiles from mOTUs2 and DADA2 were combined at genus level according to the mOTUs2 taxonomy table and MAPSeq annotations. All data were converted first into relative and then into log-relative abundances, with a pseudocount of 1E-05.

After prevalence filtering (prevalence of at least 5% in at least two studies, except for the TCGA data), all remaining bacterial genera were tested for significant differences between controls and CRC samples using a linear mixed effect model implemented in the *lmerTest* package (Kuznetsova et al., 2017), adjusting for the study as random effect. For the tissue 16S studies, some studies included matched normal tissue and tumor biopsies and in this case, the LME model was additionally adjusted for the Participant ID as a random effect. The following studies were included in this meta-analysis:

- Shotgun fecal studies (Feng et al., 2015; Thomas et al., 2019; Vogtmann et al., 2016; Wirbel et al., 2019; Yachida et al., 2019; Yu et al., 2017; Zeller et al., 2014)
- 16S fecal studies (Baxter et al., 2016; Flemer et al., 2018; Zackular et al., 2014; Zeller et al., 2014)
- 16S tissue studies (Bullman et al., 2017; Burns et al., 2015; Flemer et al., 2018; Kostic et al., 2012; Nakatsu et al., 2015; Purcell et al., 2017; Zeller et al., 2014)
- TCGA (TCGA Research Network et al., 2013) projects COAD and READ, as profiled in (Dohlman et al., 2021) and (Poore et al., 2020)



Additional Figure 1: Incorrect feature selection procedure can lead to inaccurate conclusions
The upper row shows the AUC values for the cross-validation (CV) of random forest machine learning models within regions and when applied across regions, trained using the incorrect feature selection procedure as described in (He et al., 2018). All models were trained with SIAMCAT and yield very similar results to those reported in the original publication. The lower row shows the same setting, but using the correct feature selection procedure, nested into the cross-validation. The differences between CV and application to other regions is not significant anymore in most cases or the effect size is very small. In general, the AUC values for the CV setting are lower for the correct feature selection, highlighting the inflated performance through the incorrect procedure. Black diamonds show the mean for each group.



Additional Figure 2: Contamination and bioinformatic mis-annotations can lead to biologically implausible enrichments for CRC TCGA samples
Meta-analysis significance (FDR-adjusted *P* value) and LME model effect size for the microbial profiles from (Poore et al., 2020), derived from TCGA total RNA-seq samples. The top-enriched genera are in the majority of cases not human-associated and likely represent contaminations or mis-annotations from the Kraken2 pipeline (see also (Milanese et al., 2019)). The top 15 CRC-enriched genera (identified from the other data types, see **Figure 4**) are highlighted again in green, with *Campylobacter* and *Peptostreptococcus* showing enrichment in control samples -- contrary to the results obtained from other data types.

CRC meta-analysis publication:
Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer

Jakob Wirbel, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, Anita Y. Voigt, Albert Palleja, Ruby Ponnudurai, Shinichi Sunagawa, Luis Pedro Coelho, Petra Schrotz-King, Emily Vogtmann, Nina Habermann, Emma Niméus, Andrew M. Thomas, Paolo Manghi, Sara Gandini, Davide Serrano, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Levi Waldron, Alessio Naccarati, Nicola Segata, Rashmi Sinha, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork & Georg Zeller

Published as: Wirbel, J., Pyl, P.T., Kartal, E. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 25, 679–689 (2019).
https://doi.org/10.1038/s41591-019-0406-6

Abstract

Association studies have linked microbiome alterations with many human diseases. However, they have not always reported consistent results, thereby necessitating cross-study comparisons. Here, a meta-analysis of eight geographically and technically diverse fecal shotgun metagenomic studies of colorectal cancer (CRC, n = 768), which was controlled for several confounders, identified a core set of 29 species significantly enriched in CRC metagenomes (false discovery rate (FDR) < 1 × 10−5). CRC signatures derived from single studies maintained their accuracy in other studies. By training on multiple studies, we improved detection accuracy and disease specificity for CRC. Functional analysis of CRC metagenomes revealed enriched protein and mucin catabolism genes and depleted carbohydrate degradation genes. Moreover, we inferred elevated production of secondary bile acids from CRC metagenomes, suggesting a metabolic link between cancer-associated gut microbes and a fat- and meat-rich diet. Through extensive validations, this meta-analysis firmly establishes globally generalizable, predictive taxonomic and functional microbiome CRC signatures as a basis for future diagnostics.

Main

Metagenomic sequencing technologies have enabled the study of microbial communities that colonize the human body in a culture-independent manner[1]. They have yielded glimpses into the complex, yet incompletely understood, interactions between the gut microbiome—the microbial ecosystem residing primarily in the large intestine—and its host[2]. To explore microbiome–host interactions within a disease context, metagenome-wide association studies (MWAS) have begun to map gut microbiome alterations in diabetes, inflammatory bowel disease, CRC, and many other conditions[3–12]. However, due to the many biological factors that may influence gut microbiome composition in addition to the condition studied, a current challenge for MWAS are confounders, which can cause false associations[13,14]. This issue is further aggravated by a lack of standards in metagenomic data generation and processing, making it difficult to disentangle technical from biological effects[15].

The robustness of microbiome disease associations can be assessed through comparisons across multiple metagenomic case-control studies, that is, meta-analyses. The aim of meta-analyses is to identify associations that are consistent across studies and thus less likely to be attributable to biological or technical confounders. Most informative are meta-analyses of populations from diverse geographic and cultural regions. Previous microbiome meta-analyses based on 16S ribosomal RNA (rRNA) gene amplicon data found stark technical differences between studies; the reported taxonomic disease associations were either of low effect size or not well resolved[16–18]. In contrast, shotgun metagenomics have enabled analyses with higher taxonomic resolution as well as analyses of gene functions, which have improved the statistical power needed to fine-map disease-associated strains and aid in the interpretation of host-microbial co-metabolism. However, thus far, the meta-analyses of shotgun metagenomic data have either reported on the features of general dysbiosis in comparisons across multiple diseases[19], or have left it unclear how well microbiome signatures generalize across studies of the same disease when data are rigorously separated to avoid overoptimistic evaluations of their prediction accuracy[20].

In this study, we present a meta-analysis of eight studies of CRC, including fecal metagenomic data from 386 cancer cases and 392 tumor-free controls (CTRLs). After consistent data reprocessing, we examined an initial set of five studies for CRC-associated changes in the gut microbiome. First, we investigated potential confounders; then, we identified (univariate) microbial species associations, and inferred species co-occurrence patterns in CRC. Second, we trained multivariable classification models to recognize CRC status, from both taxonomic and functional microbiome profiles, and tested how accurately these models generalized to data from studies not used for training. Moreover, we evaluated the performance improvements achieved by pooling data across studies and the disease specificity of the resulting classification models. Third, the targeted investigation of virulence and toxicity genes as candidate functional biomarkers for CRC revealed several of these to be enriched in CRC metagenomes, which is indicative of their prevalence and potential relevance in CRC patients. Three additional, more recent studies were finally used to independently validate these taxonomic and functional CRC signatures.

Results

Consistent processing of published and new data for the meta-analysis of CRC metagenomes

In this meta-analysis, we included four published studies that used fecal shotgun metagenomics to characterize CRC patients compared to healthy CTRLs (see **Table M1.1**, **Supplementary Table 1**, and **Methods** for the inclusion criteria). For an additional fifth study population, we generated new fecal metagenomic data from samples collected in Germany; a subset of samples from this patient collective were published previously (see **Table M1.1** and **Methods**[8]). These five studies were conducted on three continents and differed in sampling procedures, sample storage, and DNA extraction protocols. Notably, the fecal specimens of the United States study were freeze-dried and stored at −80 °C for more than 25 years before DNA extraction and sequencing[10]. However, in all studies, samples were collected before treatment, thus excluding cancer therapy as a potential confounding effect[14,21]. Most samples were taken before bowel preparation for colonoscopy, with some exceptions in the Germany, China, and United States studies (**Supplementary Table 2**). To ensure consistency in the bioinformatic analyses, all raw sequencing data were reprocessed using the mOTUs2 tool for taxonomic profiling[22] and MOCAT2 (metagenomic analysis toolkit) for functional profiling[23].

Table M1.1: Fecal metagenomic studies of CRC included in this meta-analysis
See the Methods for the inclusion criteria and **Supplementary Table 2** for the extended metadata. For a detailed description of patient recruitment and data generation for the German study, see the **Methods**. The data for 38 samples from the German study has been published previously as part of an independent validation cohort in[8].

Country code	Reference	No. of cases	No. of controls
France	[8]	53	61
Austria	[9]	46	63
China	[11]	74	54
United States	[10]	52	52
Germany	The current study	60	60
External validation cohorts			
Italy 1	[24]	29	24
Italy 2	[24]	32	28
Japan	Courtesy of T. Yamada et al.	40	40

Univariate meta-analysis of species associated with CRC

The first aim of the meta-analysis was to determine the gut microbial species that are enriched or depleted in CRC metagenomes in a consistent manner across the five study populations. However, since these studies differed from one another in many biological and technical aspects, we first quantified the effect of study-associated heterogeneity on microbiome composition. We contrasted this with other potential confounders (patient age, body mass index (BMI), sex, sampling after colonoscopy, and library size; additionally, smoking status, type 2 diabetes comorbidity, and vegetarian diet where available; **Extended Data Fig. 1** and **Supplementary Table 3**). This analysis revealed the factor 'study' to have a predominant impact on species composition, which is supported by a recent comparison of DNA extraction protocols, since these typically differ between studies[15]. An analysis of microbial alpha and beta diversity showed that study heterogeneity also had a larger effect on overall microbiome composition than CRC in our data (**Extended Data Fig. 2**).

Parametric effect size measures are not well established for the identification of microbial taxa significantly differing in abundance in CRC because microbiome data is characterized

by non-Gaussian distributions with extreme dispersion; thus, we used a generalization of the fold change (**Extended Data Fig. 3**) and non-parametric significance testing. In this permutation test framework[25] (herein referred to as blocked (univariate) Wilcoxon tests), differential abundance in CRC can be assessed while accounting for 'study' as a confounding effect that is treated as a blocking factor; additionally, motivated by our confounder analysis, we also blocked for 'colonoscopy' in all analyses (**Methods** and **Extended Data Fig. 1**). To rule out spurious associations due to the compositional nature of microbial relative abundance data, we additionally compared the results of this test with a method[26] that employs log-ratio transformation and found highly correlated results (**Supplementary Fig. 1** and **Supplementary Table 4**).

At a meta-analysis FDR of 0.005, we identified 94 microbial species to be differentially abundant in the CRC microbiome out of 849 species consistently detected across studies (**Supplementary Table 4** and **Methods**). Among these, we focused on a core set of the 29 most significant markers (FDR < 1 × 10−5; **Fig. 1a**) for further analysis. The latter included members of several genera previously associated with CRC, such as Fusobacterium, Porphyromonas, Parvimonas, Peptostreptococcus, Gemella, Prevotella, and Solobacterium (**Fig. 1b**)[8–11] and 8 additional species without genomic reference sequences (meta-mOTUs[22]; see **Methods**) mostly from the Porphyromonas and Dialister genera and the Clostridiales order (see **Extended Data Fig. 4** and **Supplementary Table 4** for genus-level associations). Collectively, these 29 core CRC-associated species show a previously underappreciated diversity of 11 Clostridiales species to be enriched in CRC (**Fig. 1b**). In contrast to the majority of species that are more strongly affected by study heterogeneity than by CRC status, 26 out of the 29 CRC-associated species varied more according to disease status (**Fig. 1d**).

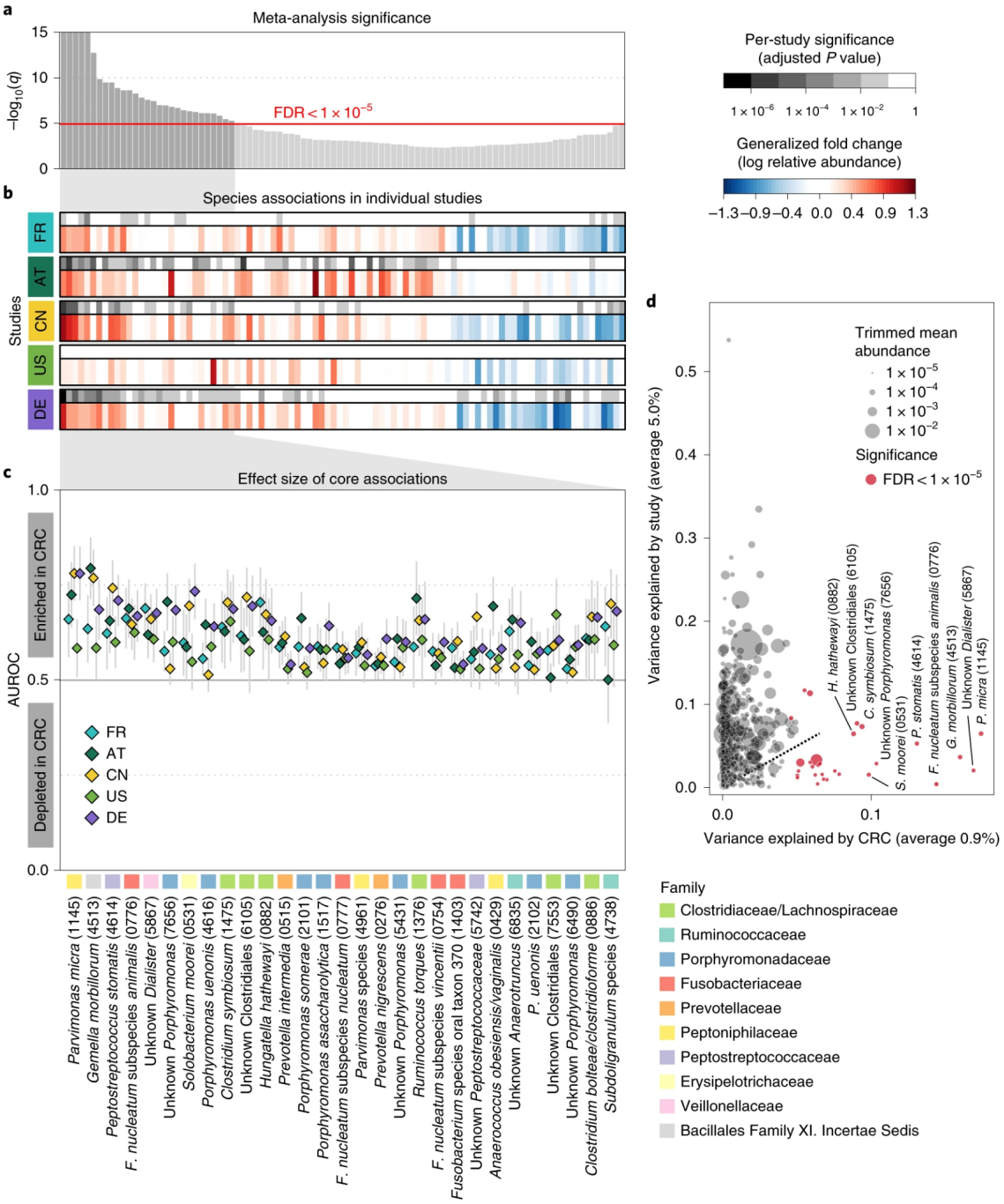


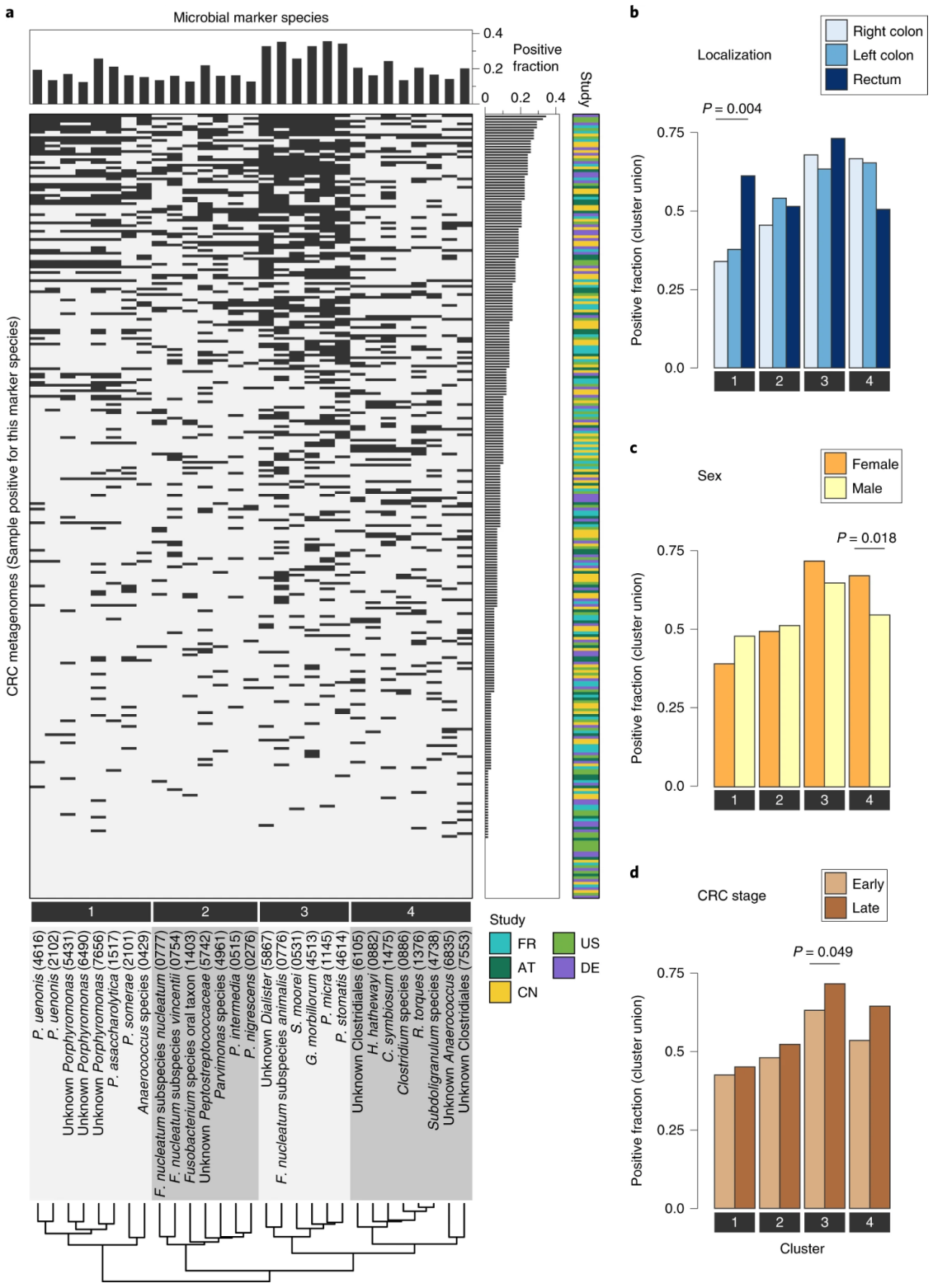
Fig. 1: Despite study differences, meta-analysis identifies a core set of gut microbes strongly associated with CRC.

a, The meta-analysis significance of gut microbial species derived from blocked Wilcoxon tests (n = 574 independent observations) is given by the bar height (FDR = 0.005). **b**, Underneath, species-level significance, as calculated with a two-sided Wilcoxon test (FDR-corrected *P* value), and the generalized fold change (**Methods**) within individual studies are displayed as heatmaps in gray and in color, respectively (see color bars and Table 1 for details on the studies included). Species are ordered by meta-analysis significance and direction of change. AT, Austria; CN, China; DE, Germany; FR, France; US, United States. **c**, For a core of highly significant species (meta-analysis FDR = 1 × 10−5), association strength is quantified by the AUROC across individual studies (color-coded diamonds), and the 95% confidence intervals are indicated by the gray lines. Family-level taxonomic information is color-coded above the species names (the numbers in brackets are mOTUs2 species identifiers; see Methods). **d**, Variance explained by disease status (CRC versus CTRLs) is plotted against variance explained by study effects for individual microbial species with dot size being proportional to abundance (see Methods); core microbial markers are highlighted in red.

All of the core CRC-associated species were enriched in patients and were often undetectable in metagenomes from non-neoplastic CTRLs. While previous studies were contradictory in the reported proportion of positive versus negative associations[8,9,17,20], our meta-analysis results are more easily reconciled with a model in which—potentially many—gut microbes contribute to or benefit from tumorigenesis than with the opposing model where a lack of protective microbes contributes to CRC development (**Fig. 1c**). Although these core taxonomic CRC associations were highly significant and consistent, individual studies showed marked discrepancies in the species identified as significant (**Fig. 1b**). Retrospective examination of the precision and sensitivity with which individual studies detected this core of CRC-associated species showed relatively low sensitivity for the United States study (consistent with the original report[10]) and low precision of the Austrian study due to associations that were not replicated in other studies (**Supplementary Fig. 2**).

Analyzing patient metagenomes for co-occurrences among the core set of 29 species that are strongly enriched in the CRC microbiome revealed four species clusters with distinct taxonomic composition (**Fig. 2a** and **Extended Data Fig. 5**; **Methods**). Two of them showed strong taxonomic consistency: cluster 1 exclusively comprised Porphyromonas species and cluster 4 only contained members of the Clostridiales order. In contrast, the other two clusters were taxonomically more heterogeneous, with cluster 3 grouping together the species with the highest prevalence in CRC cases (all among the ten most highly significant markers), consistent with a co-occurrence analysis of one of the data sets included here[11]. Cluster 2 contained species with intermediate prevalence.

Fig. 2 [next page]: Co-occurrence analysis of CRC-associated gut microbial species reveals four clusters preferentially linked to specific patient subgroups.
a, For all CRC patients (n = 285 independent samples), the heatmap shows whether the respective sample is positive for each of the core set of microbial marker species (see Methods for adjustment of positivity threshold). Samples are ordered according to the sum of positive markers, and marker species are clustered based on the Jaccard index of positive samples, resulting in four clusters (see Methods). **b–d**, The barplots in **b**, **c**, and **d** show the fraction of CRC samples that are positive for marker species clusters (defined as the union of positive marker species) broken down by patient subgroups based on differences in tumor location, sex, or CRC stage, respectively. Statistically significant associations between CRC subgroups and marker species clusters were identified using the Cochran–Mantel–Haenszel test blocked for 'study' and 'colonoscopy' effects and are indicated above the bars (P<0.1). Country codes as in Fig. 1b.



Investigating whether these four clusters were associated with different tumor characteristics, we found the Porphyromonas cluster 1 to be significantly enriched in rectal tumors (**Fig. 2b**), consistent with the presence of superoxide dismutase genes in Porphyromonas genomes possibly conferring tolerance to a more aerobic milieu in the

rectum (**Extended Data Fig. 5**). The Clostridiales cluster 4 was significantly more prevalent in female CRC patients. All species clusters showed a slight tendency toward late-stage CRC (that is, American Joint Committee on Cancer stages 3 and 4), but this was only significant for cluster 3. Associations with patient age and BMI were weaker and not significant (**Extended Data Fig. 5**). To rule out secondary effects due to differences in patient characteristics among studies, all of these tests were corrected for study effects (by blocking for 'study' and 'colonoscopy'; see **Methods**). At the level of individual species, significant stage-specific enrichments could not be detected, suggesting CRC-associated microbiome changes to be less dynamic during cancer progression than previously postulated[27], although fecal material may be less suitable to address this question than tissue samples.

Metagenomic CRC classification models

To establish metagenomic signatures for CRC detection across studies in the face of geographic and technical heterogeneity, we developed multivariable statistical modeling workflows with rigorous external validation to avoid prevailing issues of overfitting and overoptimistic reports of model accuracy[19]. As a precaution against overoptimistic evaluation, these workflows are independent of the differential abundance analysis described earlier. Instead, least absolute shrinkage and selection operator (LASSO) logistic regression classifiers were employed to select predictive microbial features and eliminate uninformative ones (see **Methods**).

In a first step, we used abundance profiles from five studies including the 849 most abundant microbial species and assessed how well classifiers trained in cross-validation on one study generalize in evaluations on the other four studies (study-to-study transfer of classifiers; **Fig. 3a**). Within-study cross-validation performance, as quantified by the area under the receiver operating characteristics curve (AUROC), ranged between 0.69 and 0.92 and was generally maintained in study-to-study transfer (AUROC dropping by 0.07 ± 0.12 on average) with two notable exceptions. First, in line with the univariate analysis of species associations, CRC detection accuracy in the United States study was lower than for the other studies, both in cross-validation and in study-to-study transfer. This could potentially be explained by the United States fecal specimens, unlike the other studies, being

freeze-archived for > 25 years before metagenomic sequencing[10]. Second, classifiers trained on the Austrian study did not generalize as well to the other studies, consistent with low study precision seen in univariate meta-analysis (**Supplementary Fig. 2**). Given the microbial co-occurrence clusters described earlier, we wondered whether species–species interactions would provide additional information relevant for CRC recognition that is not contained in the species abundance profiles. However, non-linear classifiers able to exploit such interactions did not yield significantly better accuracies (**Supplementary Fig. 3**; see also ref.[24]), suggesting that the linear model based on few biomarkers (on average 17 species accounted for more than 80% of the total classifier weights; **Extended Data Fig. 6**) is near-optimal for CRC prediction.

We further assessed if including data from all but one study in model training improves prediction on the remaining hold-out study (leave-one-study-out (LOSO) validation). The LOSO performance of species-level models ranged between 0.71 and 0.91; when the United States study was disregarded as an outlier, it was ≥0.83 (**Fig. 3b**). This corresponds to a LOSO accuracy increase of 0.076 ± 0.03 compared to study-to-study transfer. These results suggest that one can expect a CRC detection accuracy ≥0.8 (AUROC) for any new CRC study using similarly generated metagenomic data. Moreover, we verified that metagenomic CRC classification models trained on species composition were not biased for clinical subgroups. With the exception of slightly more sensitive detection of late-stage CRC (P=0.04, mostly originating from the United States study; **Extended Data Fig. 7**), we did not observe any classification bias by patient age, sex, BMI, or tumor location. Taken together, this suggests that these metagenomic classifiers are unlikely to be strongly confounded by the clinical parameters recorded.

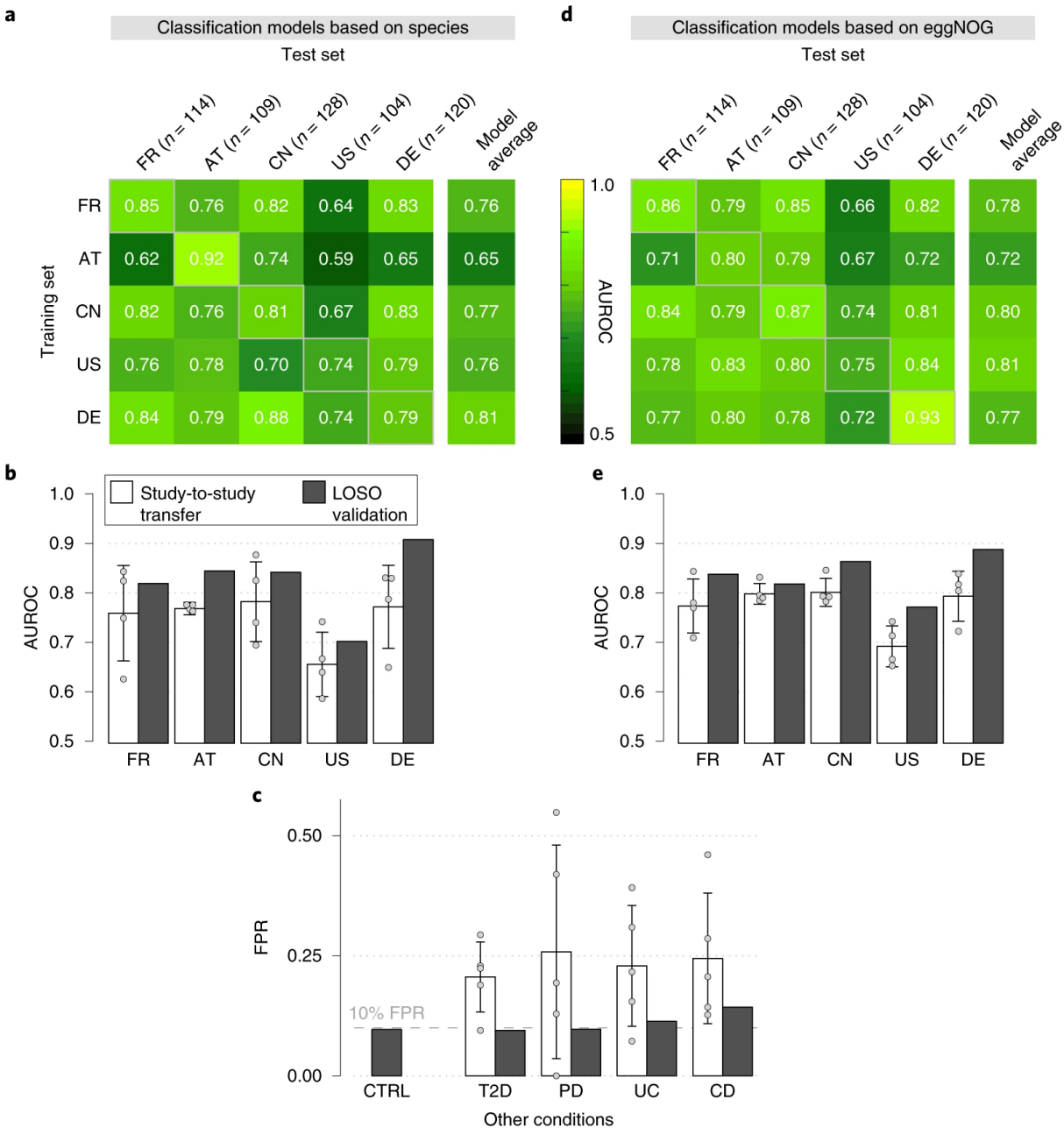


Fig. 3: Both taxonomic and functional metagenomic classification models generalize across studies, in particular when trained on data from multiple studies.

a–e, CRC classification accuracy resulting from cross-validation within each study (gray boxes along the diagonal) and study-to-study model transfer (external validations off-diagonal) as measured by the AUROC for classifiers trained on the species (a) and eggNOG gene family (d) abundance profiles. The last column depicts the average AUROC across the external validations. Classification accuracy, as evaluated by AUROC on a hold-out study, improves if taxonomic (b) or functional (e) data from all other studies are combined for training (LOSO validation) relative to models trained on data from a single study (study-to-study transfer, average and s.d. shown by bar height and error bars, respectively, n = 4). c, Combining training data across studies substantially improves CRC specificity of the (LOSO) classification models relative to models trained on data from a single study (depicted by bar color, as in c and d) as assessed by the FPR on fecal samples from patients with other conditions (see legend). The bar height for study-to-study transfer corresponds to the average FPR across classifiers (n = 5) with the error bars indicating the s.d. of the FPR values observed. T2D, type 2 diabetes; PD, Parkinson's disease; UC, ulcerative colitis; CD, Crohn's disease. Country codes as in Fig. 1b.

Several previous studies comparing microbiome changes across multiple diseases reported primarily general dysbiotic alterations and highlighted the need to examine the disease specificity of microbiome signatures[17,19]. Therefore, we assessed the false positive predictions of our metagenomic CRC classifiers on the fecal metagenomes of type 2 diabetes[4,5], Parkinson's disease[12], ulcerative colitis, and Crohn's disease[6,7] patients, reasoning that classifiers relying on biomarkers for general dysbiosis would yield an excess of false positives on these cohorts. However, our LOSO classification models calibrated to have a false positive rate (FPR) of 0.1 on CRC data sets in fact maintained similarly low FPRs on other disease data sets ranging from 0.09 to 0.13 (**Fig. 3c**). Interestingly, the disease specificity of LOSO models was significantly improved over that observed for classifiers trained on a single study, indicating that inclusion of multiple studies in the training set of a classifier can substantially improve its specificity for a given disease.

Functional metagenomic signatures for CRC

Since shotgun metagenomics data, unlike 16S rRNA gene amplicon data, allow for a direct analysis of the functional potential of the gut microbiome, we examined how predictive the metabolic pathways and orthologous gene families differing in abundance between CRC patients and CTRLs would be of CRC status. When applying the same classification workflow as stated earlier to eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) orthologous gene family abundances[28], CRC detection accuracy was very similar to that observed for the taxonomic models (**Fig. 3d,e**). AUROC values ranged from 0.70 to 0.81 for study-to-study transfer (per-study averages; see **Fig. 3e**) and from 0.78 to 0.89 in LOSO validation with a pattern of generalization across studies resembling that for taxonomic classifiers. The accuracy of functional signatures did not strongly depend on eggNOG as an annotation source, but was similar when based on other comprehensive functional databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)[29] (**Extended Data Fig. 8**). When using individual gene abundances from metagenomic gene catalogs as a classifier input[30], we observed higher within-study cross-validation AUROC values ≥ 0.96 in all studies, but lower generalization to other studies (AUROC between 0.60 and 0.79) (**Extended Data Fig. 8**).

To explore changes in the metabolic capacity of gut microbiomes from CRC patients more broadly, we quantified gut metabolic modules (defined in ref.[31]) and subjected these to the

same differential abundance analysis developed for microbial species. Gut metabolic modules with significantly higher abundance in CRC metagenomes (FDR < 0.01, Wilcoxon test blocked for 'study' and 'colonoscopy') predominantly belonged to pathways for the degradation of amino acids, mucins (glycoproteins), and organic acids. This clear trend was accompanied by a depletion of genes from carbohydrate degradation modules (**Fig. 4a,b**). The differences in all four high-level categories were highly significant ($P<1 \times 10-6$ in all cases, blocked Wilcoxon tests) and consistent across studies (**Fig. 4b**). Overall, these results establish a clear shift from dietary carbohydrate utilization in a healthy gut microbiome to amino acid degradation in CRC that is consistent with an earlier report based on a subset of the data[8]. Correlation analysis suggests that increased capacity for amino acid degradation is mostly contributed by CRC-associated Clostridiales (compare with cluster 4 in **Fig. 2** and **Supplementary Fig. 4**). Approximately one half of these metagenomic pathway enrichments are also in agreement with independent metabolomics data, suggesting increased availability of amino acids in the epithelial cells or feces of CRC patients (**Supplementary Table 5**)[32–36]. While the observed pathway enrichments could potentially result from many factors, including unmeasured ones[13], they are consistent with established dietary risk factors for CRC, which include red and processed meat consumption[37] and low fiber intake[38].

Fig. 4 [next page]: Meta-analysis identifies consistent functional changes in CRC metagenomes.
a, The meta-analysis significance of gut metabolic modules derived from blocked Wilcoxon tests (n = 574 independent samples) is indicated by the bar height (top panel, FDR = 0.01). Underneath, the generalized fold change (see **Methods**) for gut metabolic modules[31] within individual studies is displayed as a heatmap (see color key in **b**). Metabolic modules are ordered by significance and direction of change. A higher-level classification of the modules is color-coded below the heatmap for the four most common categories (colors as in **b**; white indicates other classes). **b**, Normalized log abundances for these selected functional categories is compared between CTRLs and CRC cases. Abundances are summarized as the geometric mean of all modules in the respective category and statistical significance determined using blocked Wilcoxon tests (n = 574 independent samples, see **Methods**). **c**, Normalized log abundances for virulence factors and toxins compared between metagenomes of CTRLs and CRC cases (significant differences, P < 0.05 was determined by blocked Wilcoxon test, n = 574 independent samples; see **Methods** for gene identification and quantification in the metagenomes). *fadA*, gene encoding *F. nucleatum* adhesion protein A; *bft*, gene encoding *B. fragilis* enterotoxin; *pks*, genomic island in *E. coli* encoding enzymes for the production of genotoxic colibactin; *bai*, bile acid-inducible operon present in some Clostridiales species encoding bile acid-converting enzymes.

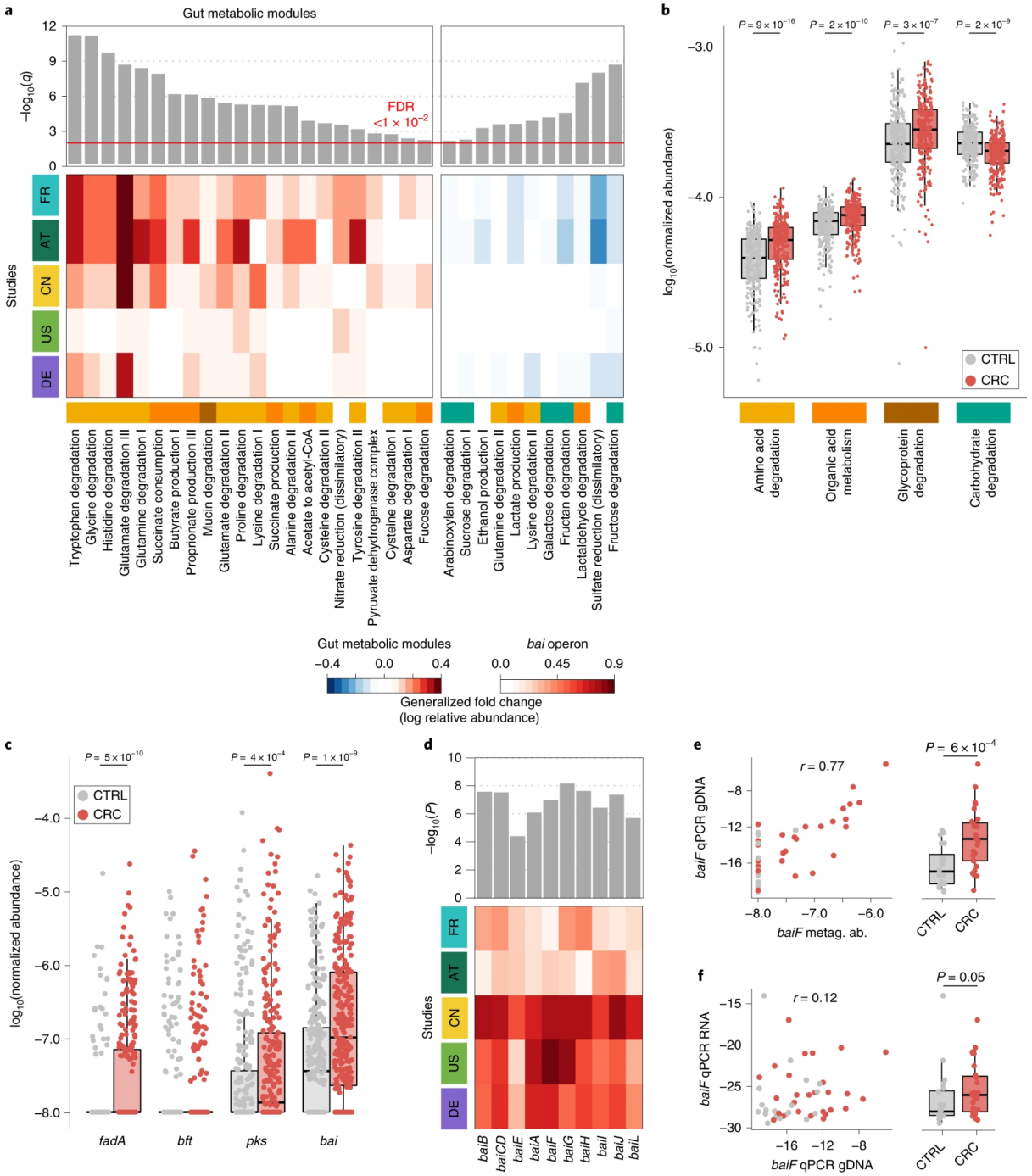


Fig. 4 [continued] d, The meta-analysis significance (uncorrected *P* value), as determined by blocked Wilcoxon tests (n = 574 independent samples), and generalized fold change within individual studies are displayed as bars and heatmap, respectively, for the genes contained in the *bai* operon. Due to high sequence similarity to *baiF*, *baiK* was not independently detectable with our approach. **e**, Metagenomic quantification of *baiF* (metagenomic abundance-normalized relative abundance) is plotted against qPCR quantification in gDNA extracted from a subset of German study samples (n = 47), with Pearson correlation (*r*) indicated (see **Methods**). **f**, Expression of *baiF* determined via qPCR on reverse-transcribed RNA from the same samples in contrast to gDNA (as in **e**). The boxplots on the right of **e** and **f** show the difference between CRC and CTRL samples in the respective qPCR quantification (the *P* values on top were calculated using a one-sided Wilcoxon test). All boxplots show the interquartile ranges (IQRs) as boxes, with the median as a black horizontal line and the whiskers extending up to the most extreme points within 1.5-fold IQR. Country codes as in **Fig. 1b**.

The large metagenomic data set analyzed in this study allowed us to quantify the prevalence of the gut microbial virulence and toxicity mechanisms thought to play a role in colorectal carcinogenesis. Prominent examples include the *Fusobacterium nucleatum* adhesion protein A (encoded by the *fadA* gene), the *Bacteroides fragilis* enterotoxin (*bft* gene) and colibactin produced by some *Escherichia coli* strains (from the *pks* genomic island)[39,40]. Moreover, intestinal Clostridium species are known to contribute to the conversion of primary to secondary bile acids using several metabolic pathways including 7α-dehydroxylation, encoded in the *bai* operon[41]. The products of this 7α-dehydroxylation pathway, deoxycholate and lithocholate, are known hepatotoxins associated with liver cancer[42] and hypothesized to also promote CRC[43]. Although intensely studied at a mechanistic level, these factors are not (well)-represented in general databases that can be used for metagenome annotation (**Supplementary Fig. 5**). Thus, we built a targeted metagenome annotation workflow based on Hidden Markov Models (HMMs) to identify and quantify the virulence factors and toxicity pathways of interest in CRC. Additionally, we used co-abundance clustering to infer operon completeness for factors encoded by multiple genes (see **Methods**, **Extended Data Fig. 9**, and **Supplementary Fig. 5**). While *fadA*, *bft*, the *pks* island, and the *bai* operon were clearly detectable in deeply sequenced fecal metagenomes, they varied broadly with respect to abundance, significance, and cross-study consistency of enrichment (**Fig. 4c**). *fadA* and *pks* were significantly enriched in CRC metagenomes (P=5.3 × 10−10 and 4.1 × 10−4, respectively), whereas no significant abundance difference could be detected for *bft* in fecal metagenomes, despite reports on its enrichment in the mucosa of CRC patients[44], its carcinogenic effect in mouse models[45], and synergistic action with *pks*[46]. Our quantification of the *bai* operon showed a highly significant enrichment in CRC metagenomes (P=1.6 × 10−9) observed across all five studies (**Fig. 4d**) at an average abundance that exceeded *fadA* and *pks* copy numbers (**Fig. 4c**). Metagenome analysis indicated that at least four Clostridiales species (including the well characterized *Clostridium scindens* and *Clostridium hylemonae*)[47,48] have a (near)-complete 7α-dehydroxylation pathway contributing to the observed enrichment of *bai* operon copies (**Extended Data Fig. 9**). To validate this finding and further explore its value toward diagnostic application, we developed a targeted quantification assay for the *baiF* gene based on quantitative PCR (qPCR; see **Methods**). Quantification of *baiF* by qPCR using genomic DNA (gDNA) from 47 fecal samples of the German study population was found to

be similar to, yet more sensitive than by metagenomics (**Fig. 4e**). Gut microbial *baiF* copy numbers clearly distinguished CRC patients from CTRLs (P = 0.001) at an AUROC of 0.77, which in this subset of samples is surpassed by only a single-species marker for CRC (**Extended Data Fig. 9**). Although consistent with the increased deoxycholate metabolite levels reported for serum and stool samples of CRC patients[49], this finding does not imply 7α-dehydroxylation pathway activity. Therefore, we quantified *baiF* expression using RNA extracts from the same set of fecal samples, and found transcript levels to be elevated in CRC patients also (**Fig. 4f**). The observed weak correlation of *baiF* expression with genomic abundance (**Fig. 4f**) might be explained by dynamic transcriptional regulation[50] and therefore *bai* expression in feces might not accurately reflect the tumor environment. Taken together, these data suggest gut microbial metabolic markers to be meaningful and highly predictive of CRC status.

Validation of CRC signatures in independent study populations

Even though CRC classification accuracy for both species and functions were evaluated on independent data, we nonetheless sought to confirm it using two additional study populations from Italy (Italy 1 and Italy 2, combined N = 61 CRC, N = 62 CTRLs; see **Methods** and **Table 1**) and one from Japan (N = 40 CRC, N = 40 CTRLs; see **Methods** and **Table 1**). The overlap of single- species associations detected in the Italy 2 study and those from the meta-analysis was found to vary within the range seen for the other studies, whereas for Italy 1 and Japan, the overlap was slightly lower (compare study precision in **Supplementary Fig. 2** and **Extended Data Fig. 10**). Nonetheless, the AUROC of LOSO classification models based on species ranged between 0.79 and 0.81; that for the classifiers based on eggNOG ranged from 0.71 to 0.92 (**Fig. 5a,b**). We also validated CRC enrichment of the *fadA*, *pks*, and *bai* genes in these three study populations (**Fig. 5c**). Altogether, these results highlight consistent alterations in the gut microbiome of CRC patients across eight study populations from seven countries in three continents.

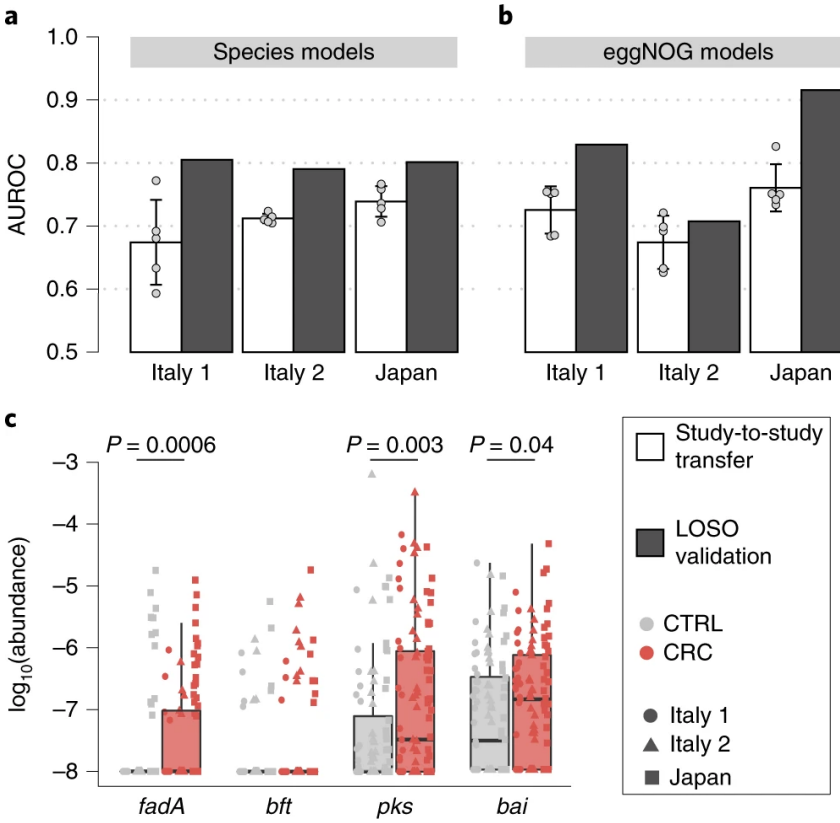


Fig. 5: Meta-analysis results are validated in three independent study populations.
a,b, CRC classification accuracy for independent data sets, two from Italy and one from Japan (see Table 1 and Supplementary Table 2), is indicated by the bar height for single-study (white) and LOSO (gray) models using either species (a) or eggNOG gene family (b) abundance profiles (see Fig. 3). Bar height for single-study models corresponds to the average of five classifiers (the error bars indicate the s.d., n = 5). **c**, Normalized log abundances for virulence factors and toxins (see Fig. 4c) compared between CTRLs and CRC cases. *P* values were determined by one-sided blocked Wilcoxon tests (n = 193 independent samples). The boxes represent the IQRs with the median as a black horizontal line and the whiskers extending up to the most extreme points within 1.5-fold IQR.

Discussion

Through extensive and statistically rigorous validation, where data from studies used for training is strictly separated from that for testing, our meta-analysis firmly establishes that gut microbial signatures are highly predictive of CRC (see also[24]). In particular, metagenomic classifiers trained on species profiles from multiple studies maintained an AUROC of at least 0.8 in seven out of eight data sets and achieved an accuracy similar to the fecal occult blood test, a standard non-invasive clinical test for CRC (**Supplementary Fig. 6**; see[8]). Thus, these results suggest that polymicrobial CRC classifiers are globally applicable and can overcome technical and geographical study differences, which we found to generally impact observed microbiome composition more than the disease itself (**Fig. 1c** and **Extended Data Figs. 1** and **2**). The generalization accuracy of classifiers across studies seen in this study is higher than that reported in 16S rRNA gene amplicon sequencing studies, which are characterized by even larger heterogeneity across studies[16,18] (**Supplementary Fig. 7**).

Previous microbiome meta-analyses suggested that the majority of gut microbial taxa differing in any given case-control study reflect general dysbiosis rather than disease-specific alterations, thereby illustrating the difficulty of establishing disease-specific microbiome signatures[17,19]. In the current study, by combining data across studies for training (LOSO), we developed disease-specific signatures that maintained false positive control on diabetes and inflammatory bowel disease metagenomes at a very similar level as for CRC (**Fig. 3c**), despite these diseases having shared effects on the gut microbiome[17,50] and an increased comorbidity risk[51].

Although for diagnostic purposes, unresolved causality between microbial and host processes during CRC development are not a central concern, elucidating the underlying mechanisms would greatly enhance our understanding of colorectal tumorigenesis. Toward this goal, we developed both broad and targeted annotation workflows for functional metagenome analysis. First, we found functional signatures based on the abundances of orthologous groups of microbial genes to yield accuracies as high as taxonomic signatures (**Fig. 3**), which raises the hope for future improvements in metagenome annotation that can be translated into microbiome signature refinements. Second, by investigating potentially

carcinogenic bacterial virulence and toxicity mechanisms using a targeted metagenome annotation approach, we confirmed highly significant enrichments of the colibactin-producing *pks* gene cluster and the *F. nucleatum* adhesin *FadA* in CRC metagenomes (**Fig. 4c**). Our results support the clinical relevance of these factors and add to the experimental evidence for their carcinogenic potential [46,52–54]. We further examined the *bai* operon, which encodes enzymes that produce secondary bile acids via 7α-dehydroxylation, as an example of toxic host-microbe co-metabolism (see [24] for another intriguing example). While α-dehydroxylated bile acids are established liver carcinogens [42], their contribution to CRC is less clear [43]. In the current study, we have, for the first time, shown *bai* to be highly enriched in stool from CRC patients (**Fig. 4c,d**) and confirmed this finding at both the genomic and transcriptomic level using qPCR (**Fig. 4e,f**). Since *bai* enrichment (and expression) is probably a consequence of a diet rich in fat and meat [55], it is intriguing to explore whether *bai* could be used as a surrogate microbiome marker for such difficult-to-measure dietary CRC risk factors.

To further unravel the molecular underpinning of dietary CRC risk factors, molecular pathological epidemiology studies that investigate the mucosal microbiome as part of the tumor microenvironment hold great potential [56,57]. However, they will require more comprehensive diet questionnaires, medical records, and molecular tumor characterizations than are available for the study populations analyzed in the current study. In this context, carcinogens possibly contained in the virome also warrant further investigation [58,59]; however, for this goal, metagenomic data need to be generated with protocols optimized for virus enrichment [60].

Taken together, our results and those by [24], strongly support the promise of microbiome-based CRC diagnostics. Both the taxonomic and metabolic gut microbial marker genes established in these meta-analyses could form the basis of future diagnostic assays that are sufficiently robust, sensitive, and cost-effective for clinical application. The targeted qPCR-based quantification of the baiF gene is a first step in this direction. Our metagenomic analysis of this and other virulence and toxicity markers bridge to existing mechanistic work in preclinical models and could enable future work that aims to precisely determine the contribution of gut microbiota to CRC development

Methods

Study inclusion and data acquisition

We used PubMed to search for studies that published fecal shotgun metagenomic data of human CRC patients and healthy CTRLs. The search term, all hits, and the justification for exclusion or inclusion are available in **Supplementary Table 1**. Raw FASTQ files were downloaded for the four included studies from the European Nucleotide Archive (ENA) using the following ENA identifiers: PRJEB10878 for[11], PRJEB12449 for [10], ERP008729 for [9], and ERP005534 for [8].

German study recruitment and sequencing

The German study population data consist of 60 fecal CRC metagenomes, 38 of which were sequenced and published in [8] under ENA accession no. ERP005534. The fecal metagenomes from an additional 22 CRC patients recruited for the same ColoCare study (German Cancer Research Center, Heidelberg [61,62]) were sequenced later as part of this work. All fecal samples were collected after colonoscopy. Sixty sex- and age-matched participants of the PRÄVENT study run by the same clinical investigators were included as healthy CTRLs; since these participants did not undergo colonoscopy, the presence of undiagnosed colorectal carcinomas cannot be completely ruled out but is expected to be unlikely due to the low prevalence of preclinical CRC in the general population [63].

Written informed consent was obtained from all additional 22 CRC patients and 60 CTRLs. The study protocol was approved by the institutional review board (European Molecular Biology Laboratory (EMBL) Bioethics Internal Advisory Board) and the ethics committee of the Medical Faculty at the University of Heidelberg. The study is in agreement with the World Medical Association Declaration of Helsinki (2008) and the Department of Health and Human Services, Belmont Report.

Genomic DNA was extracted from the fecal samples (preserved in RNALater, Sigma-Aldrich) and libraries were prepared as described previously [8]. Whole-genome shotgun sequencing was performed with the HiSeq 2000/2500/4000 systems (Illumina) at the Genomics Core Facility, EMBL, Heidelberg.

Independent validation cohorts

During the revision of this manuscript, we included three independent study populations for external validation. Two of them were recruited in Italy (Italy 1 and Italy 2) with informed consent from all participants and ethical approval by the ethics committees of Azienda Ospedaliera di Alessandria and the European Institute of Oncology of Milan. Fecal shotgun metagenomic data were generated as described in [24].

The third study population was recruited in Japan with informed consent and ethical approval of the institutional review boards of the National Cancer Center Japan— Research Institute and the Tokyo Institute of Technology. DNA was extracted from frozen fecal samples using a Gnome DNA Isolation Kit (MP Biomedicals) with an additional bead-beating step as described previously [64]. DNA quality was assessed with an Agilent 4200 TapeStation (Agilent Technologies). After final precipitation, the DNA samples were resuspended in Tris-EDTA buffer and stored at −80 °C before further analysis. Sequencing libraries were generated with the Nextera XT DNA Sample Preparation Kit (Illumina). Library quality was confirmed with an Agilent 4200 TapeStation.

Whole-genome shotgun sequencing was carried out on the HiSeq 2500 system (Illumina). All samples were paired-end sequenced with a 150-base pair (bp) read length to a targeted data set size of 5.0 Gb.

Taxonomic profiling and data preprocessing

The metagenomic samples were quality controlled using MOCAT2's '-rtf' procedure, which is based on the 'solexaQA' algorithm [23]. In particular, reads that map with at least a 95% sequence identity and an alignment length of at least 45 bp to the human genome hg19 were removed. In a second step, taxonomic profiles were generated with the mOTU profiler v.2.0.0 (refs. [22,65,66]; see https://motu-tool.org/ and GitHub v.2.0.0) using the following parameters: -l 75; -g 2; and -c. Briefly, this profiler is based on ten universal single-copy marker gene families (COG0012, COG0016, COG0018, COG0172, COG0215, COG0495, COG0525, COG0533, COG0541, and COG0552) [66]. These marker genes were extracted from > 25,000 reference genomes and > 3,000 metagenomic samples allowing us to profile prokaryotic species with a sequenced reference genome (ref-mOTUs) and ones without (meta-mOTUs). The read count for a mOTU was calculated as the median of the read count of the genes that belonged to that mOTU.

mOTU profiles were first converted to relative abundances to account for library size. Then, profiles were filtered to focus on a set of species that were confidently detectable in multiple studies. Specifically, microbial species that did not exceed a maximum relative abundance of 1 × 10−3 in at least three of the studies were excluded from further analysis, together with the fraction of unmapped metagenomic reads.

Functional metagenome profiling and data preprocessing

High-quality reads, with the same quality filtering as for taxonomic profiling, were aligned against a combined database (IGChg38 hereafter) consisting of the hg38 release of the human reference genome and the integrated gene catalog (IGC) containing 9.9 million non-redundant microbial genes [30] using the Burrows–Wheeler Aligner MEM algorithm [67] (v.0.7.15-r1140) with default parameters. The purpose of adding the human genome to the reference database was to filter out reads that mapped as well as or better to some human sequence than to any bacterial gene. Alignments were calculated separately for paired-end and single-read libraries. (Single reads could result from read pairs where one read was filtered out in the quality filtering procedure described earlier.) Alignments were then filtered to only retain those longer than 50 bp with > 95% sequence identity. Then, the highest scoring alignment(s) was/were kept for each read. As IGChg38 is a database of predominantly genes and not genomes, there will be a substantial proportion of read pairs where one end maps within the gene while the other end does not—it either maps to an adjacent gene or remains unmapped due to intergenic regions not contained in the database. Therefore, we counted a whole read pair aligning to a gene when (1) both ends from a read pair mapped to the same gene, (2) only one end from a read pair mapped to the gene, or (3) a read from the single-read library mapped to the gene. We then counted only the read pairs that mapped uniquely to one gene in the IGC, thus excluding ambiguous read pairs that mapped with similarly high scores to multiple genes in the database. For a given metagenomic sample, we further normalized the abundance of each IGC gene by the length of that gene. We then estimated the relative abundance of IGC genes by dividing gene abundances by the total abundance of all genes in the IGC (excluding the human chromosomes).

Because the metagenomes from CRC patients were not included when the IGC was constructed, we analyzed how well CRC-associated species as identified in this meta-analysis were represented in the IGC. Using a phylogenetic marker gene (COG0533), which is also used by the species profiling workflow on which the meta-analysis is based, for 24 out of the 29 core CRC-associated species, we found a match in the IGC with at least 90% nucleotide identity, indicating that a sequence from the same species (above 93.1% identity) or a slightly more distant relative is present in the IGC (**Supplementary Fig. 8**).

The relative abundance of eggNOG orthologous groups [28] was estimated by summing the relative abundances of genes annotated to belong to the same eggNOG orthologous group as of the most recent annotations provided by MOCAT2 [23]. To obtain the KEGG orthologous groups and pathway abundances, we applied the same procedure, but using the KEGG annotations for the IGC provided by MOCAT2 [29].

Overview of statistical analyses

For univariate association testing between the abundances of microbial taxa and gene functions, we used non-parametric tests throughout; all were two-sided Wilcoxon tests except where otherwise stated. To account for potential confounders and heterogeneity between data sets, we employed a stratified version of the Wilcoxon test [25]. Analysis of variance (ANOVA) was conducted on rank-transformed data. The significance of binary co-occurrence patterns was assessed using (stratified) Cochran–Mantel–Haenszel tests.

Multivariable analysis was done with strict separation between training and test data. Importantly, this also pertained to feature selection, which was either done via LASSO regression analysis [68] or by nested cross-validation procedures to avoid overoptimistic performance assessment [69]. All samples included in this meta-analysis came from distinct individuals to ensure that generalization across participants—rather than across time points within a given participant—is assessed.

Confounder analysis

To quantify the effect of potential confounding factors relative to that of CRC on single microbial species, we used an ANOVA-type analysis. The total variance within the abundance of a given microbial species was compared to the variance explained by disease status and the variance explained by the confounding factor akin to a linear model, including both CRC status and the confounding factor as explanatory variables for species abundance. Variance calculations were performed on ranks to account for non-Gaussian distribution of microbiome abundance data. Potential confounders with continuous values were transformed into categorical data either as quartiles or in the case of BMI into lean/obese/overweight according to conventional cutoffs (lean: <25; obese: 25–30; overweight: >30).

Univariate meta-analysis for the identification of CRC-associated gut microbial species

The significance of differential abundance was tested on a per species basis using a blocked Wilcoxon test implemented in the R 'coin' package [25]. Informed by the results of the preceding confounder analysis, we blocked for 'study' and 'colonoscopy' in the Chinese study. Within this framework, significance is tested against a conditional null distribution derived from permutations of the observed data. Notably, permutations are performed within each block to control for variations in block size and composition. To adjust for multiple hypothesis testing, *P* values were adjusted using the FDR method [70].

As non-parametric effect size measures, we used the AUROC with permutation-based confidence intervals calculated using the 'pROC' package in R [71]. We further developed a generalization of the (logarithmic) fold change that is widely used for other types of read abundance data. This generalization is designed to have better resolution for sparse microbiome profiles, where 0 entries can render median-based fold change estimates uninformative for a large portion of species with a prevalence below 0.5. The generalized fold change is calculated as the mean difference in a set of predefined quantiles of the logarithmic CTRL and CRC distributions (see **Extended Data Fig. 3** for further details). We used quantiles ranging from 0.1 to 0.9 in increments of 0.1.

For the retrospective analysis of study precision and recall in detecting microbial species associations from the meta-analysis, the true set was defined as the species that were associated at a given FDR in the meta-analysis. Then, we checked how well this set of species would be recovered using the single-study significance as determined by the Wilcoxon test. Study precision corresponds to the proportion of meta-analysis-significant species among those detected as significant in a single study. Similarly, recall (or sensitivity) corresponds to the proportion of species out of the true set of meta-analysis-significant species that were recovered in a given study.

Species co-occurrence and cluster analysis in CRC metagenomes

For the analysis of gut bacterial species co-occurring in CRC microbiomes, the relative abundances of the core set of associated species were discretized into binary values to determine whether a CRC (metagenomic) sample was 'positive' or 'negative' for a given microbial marker. To normalize for differences in prevalence (and therefore specificity) of these markers, we adjusted the threshold value above which a sample was labeled positive based on the abundance in healthy CTRLs. For each microbial species, the 95th percentile in healthy CTRLs was used as the threshold, which effectively results in adjusting the per marker FPR to 0.05. Based on the binarized species-by-sample matrix, species were then clustered using the Jaccard index as implemented in the 'vegan' package in R [72]. Associations between species clusters and meta-variables were tested as 2-by-n (where n is the number of categories in the meta-variable tested) contingency tables using a Cochran–Mantel–Haenszel test with 'study' and 'colonoscopy' as blocking factors, as implemented in the R 'coin' package [25].

Multivariable statistical modeling workflow and model evaluation

A main goal of our work is to assess the generalization accuracy of microbiome-based CRC classifiers across technical and geographic differences in patient populations; thus, we extensively validated classification models across studies taking the following two approaches.

In study-to-study transfer validation, metagenomic classifiers were trained on a single study and their performance was externally assessed on all other studies (off-diagonal cells in **Fig. 3a,d**). Effectively, we implemented a nested cross-validation procedure on the training study to calculate within-study accuracy (cells on the diagonal in **Fig. 3a,d**) and tune the model hyperparameters.

In LOSO validation, data from one study was set aside as an external validation set, while the data from the remaining four studies was pooled as a training set on which we implemented the same nested cross-validation procedure as for the study-to-study transfer (see [19] for a more detailed description of LOSO).

Data preprocessing, model building, and model evaluation was performed using the SIAMCAT R package v.1.1.0 (https://bioconductor.org/packages/SIAMCAT).

Preprocessing of taxonomic abundance profiles for statistical modeling

Relative abundances were first filtered to remove markers with low overall abundance and no variance (an artifact of single-study data arising from the joint data filtering described earlier), log10-transformed (after adding a pseudo-count of $1 \times 10{-}5$ to avoid non-finite values resulting from log10(0) [73]), and finally standardized as z-scores. Data were split into training and test sets for 10 times-repeated, tenfold stratified cross-validation (balancing class proportions across folds). For each split, an L1-regularized (LASSO) logistic regression model [68] was trained on the training set, which was then used to predict the test set. The lambda parameter, that is, regularization strength, was selected for each model to maximize the area under the precision-recall curve under the constraint that the model contained at least five non-zero coefficients. Models were then evaluated by calculating the AUROC based on the posterior probability for the CRC class.

In model transfer to a hold-out study, the hold-out data were normalized for comparability in the same way as the training data set by using the frozen normalization function in SIAMCAT, which retains the same features and reuses the same normalization parameters (for example, the mean of a feature for z-score standardization). Then, all 100 models derived from the cross-validation on the training data set (10 times-repeated tenfold cross-validation) were applied to the hold-out data set and predictions were averaged across all models.

In the LOSO setting, data from the four training studies were jointly processed as a single data set in the same way as described earlier using ten times-repeated tenfold stratified cross-validation.

Preprocessing of functional abundance profiles

Functional profiles, such as eggNOG gene family or KEGG module abundance profiles were preprocessed as described earlier for the species profiles, but using $1 \times 10{-}6$ as the maximum abundance cutoff and $1 \times 10{-}9$ as a pseudo-count during log transformation. Since these abundance tables contained several thousand input features, we implemented an additional feature selection step, which was nested properly into the cross-validation procedures described earlier. This nested approach is crucial to avoid over-optimistically biased performance estimates (see [74], Chapter 7.10). Specifically, features were filtered inside each training fold (without using any label information from the test fold) by selecting the 1,600 features with the highest single-feature AUROC values (for features depleted in CRC, 1 – AUROC was used for feature selection).

Preprocessing of gene abundance profiles

To ascertain the predictive power of a classifier based on the IGC gene abundances [30], we applied a series of filters to the abundance tables to reduce the number of genes that would be the input of LASSO modeling. These filters were applied once on a per study level and once in a LOSO mode, where they were applied jointly to all studies in the training set, with the remaining one being held out for external validation.

The following filters were applied in this order: (1) all genes with 0 abundance in ≥15% of samples (regardless of CRC status) were discarded; (2) the remaining data were discretized using the equal frequencies method implemented in the 'discretize' function of the 'sideChannelAttack' R package (v.1.0–6) as a preparation to the minimal-redundancy-maximal-relevance (mRMR) algorithm [75]; (3) as a feature selection procedure, the mRMR

(code version from 20 April 2009 downloaded from http://home.penglab.com/proj/mRMR/ on 3 December 2016) was run on the gene abundance table to retain the 100 top genes as output.

LASSO models were then built on log10-transformed abundances (pseudo-count of 10 × 10−9, centered and scaled) of the sets of the 100 top genes returned by mRMR. The whole process was repeated 10 times in a fivefold stratified cross-validation scheme to allow for an estimation of the confidence of the AUROC of the resulting models. We used the 'LiblineaR' package (v.2.10–8) to build the LASSO models in R and tested a sequence of 20 cost parameters (equivalent or the lambda parameter controlling the regularization strength) evenly spaced from 0.0012 to 0.22. The cost parameter was selected to maximize the AUROC within the training set.

External evaluation of disease specificity of the metagenomic classifiers

To assess how disease-specific the predictions of the CRC models were, we applied these to data from case-control studies investigating other human diseases. Fecal metagenomic data of patients with Parkinson's disease [12], type 2 diabetes [4,5], and inflammatory bowel disease [6,7] were taxonomically profiled as described earlier. The parameters for quality control with MOCAT2 and for mOTUs2 were the same as described earlier, except for the data from [6], where we used mOTUs2 with -l 50 to set the threshold for minimum alignment length to 50 since the read length is shorter (average read length 71) compared to the other more recently generated Illumina shotgun metagenomic data.

Relative abundance data were treated exactly as another hold-out data set for each model, that is, by applying the frozen normalization prediction routines as described earlier. For each CRC model applied to the external data sets, a cutoff on its prediction output was adjusted to yield an FPR of 0.1 on the CTRLs of its respective (CRC) training set. Subsequently, its FPR on metagenomes from patients suffering from the previously mentioned (non-CRC) conditions was assessed to evaluate its disease specificity. The rationale behind this is that a metagenomic classifier that recognizes the general features of dysbiosis would be expected to predict CRC patients and those suffering from other conditions at a similar rate; thus, in the evaluation described previously, such a classifier would display a much higher FPR than on the CTRLs of its training set. In contrast, maintaining a low FPR in this evaluation indicates that the classification model is based on CRC-specific features rather than the hallmarks of general dysbiosis or non-specific inflammation.

Functional profiling of gut metabolic modules

Gut metabolic modules were calculated as originally proposed [31], using the KEGG orthology profiles based on the IGC (see Functional metagenome profiling and data preprocessing) as input. Statistical analysis and generalized fold change calculations were performed analogously to species profiles (see earlier). Gut metabolic modules were summarized across functional groups (for example, amino acid degradation) as the geometric mean of all modules within the respective group.

Targeted functional analysis of virulence and toxicity pathways of potential relevance in CRC

To investigate the toxicity and virulence mechanisms that have previously been implicated in CRC [40], for each gene belonging to the respective virulence or toxicity pathway, we constructed an HMM. Each HMM was built from a multiple sequence alignment generated by MUSCLE (MUltiple Sequence Comparison by Log-

Expectation) [76], containing the respective reference sequences and close homologs identified using PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) [77]. Multiple sequence alignments are available together with the code for this study (https://github.com/zellerlab/crc_meta). Then, we screened the IGC metagenomic gene catalog [30] with each HMM using the HMMER software (v.3.1b2) [78]. Genes with an e-value below 1 × 10−10 were filtered for uniqueness since in some cases the HMMs would call different regions in the same gene. For single-gene virulence factors (that is, *fadA* and *bft*), potential IGC hits were aligned against the reference sequence using the Needleman–Wunsch algorithm in the European Molecular Biology Open Software Suite package [79]. Hits were then filtered based on the percentage of sequence identity (cutoff: 40%) and sequence similarly to the species relative abundance profiles based on maximum relative abundance (cutoff: 1 × 10−7) to exclude genes with limited relevance. Statistical analysis was performed on the sum of all genes.

For virulence pathways containing more than one gene, the IGC hits of each functional group within the pathway were aligned against the respective reference sequence and filtered for the percentage of sequence identity and maximum abundance. Then, all hits were clustered based on the Pearson correlation of the log abundances across all samples using the Ward algorithm as implemented in the 'hclust' function in R. The gene clusters were filtered based on operon completeness (that is, how many genes of the operon were present in the cluster) and average correlation within the cluster (**Extended Data Fig. 9**). For statistical analysis, the genes in the selected gene clusters were summed within each group or all together for the overall analysis.

Quantitative PCR for baiF

Real-time qPCR to quantify the abundance and expression of *baiF* was performed on a subset of samples in the German cohort (20 CTRL and 24 CRC samples; see **Supplementary Table 6**). For these samples, DNA and RNA extraction was done with the Allprep PowerFecal DNA/RNA Kit (QIAGEN) with additional RNAse and DNAse digestion steps, respectively, as described by the manufacturer. DNA and RNA concentrations were determined using a Qubit Fluorometer (Invitrogen); quality control of all RNA samples was done using an Agilent 2100 Bioanalyzer (Agilent Technologies) in combination with the RNA 6000 Nano and Pico LabChip kits (Agilent Technologies).

First-strand complementary DNA (cDNA) was synthesized using the SuperScript IV VILO Master Mix with the ezDNAse enzyme and random hexamer primers (Thermo Fisher Scientific), as recommended by the manufacturer. Reactions were performed as described in the protocol with one minor change of temperature. The incubation for the reverse transcription step was carried out at 55 °C.

To quantify *baiF* relative to the total bacterial RNA/DNA in a sample, qPCR was performed in triplicates for the 16S rRNA and *baiF* genes using both cDNA and gDNA as templates. We used the following primers for *baiF*: TTCAGYTTCTACACCTG (forward); GGTTRTCCATRCCGAACAGCG (reverse); standard primers F515 and R806 for 16S [80]. Real-time PCR reactions were prepared with a final primer concentration of 0.5 µM, including 5 ng of gDNA or 10 ng of cDNA in a 20 µl final reaction volume; reactions were performed with a SYBR Green qPCR Mix on a StepOne Real-Time PCR system (Thermo Fisher Scientific). Cycling conditions were as follows: initial denaturation at 95 °C for 10 min; 40 cycles of denaturation at 95 °C for 15 s; and annealing at 60 °C for 60 s followed by melt curve analysis.

Δ-Ct values were calculated as the difference between *baiF* and 16S Ct values. The significance of the comparison between CTRL and CRC samples was tested on the Δ-Ct values using a one-sided Wilcoxon test as confirmation of metagenomic enrichment.

Data availability

The raw sequencing data for the samples in the German study that have not been published before (see **Methods**) are available from the European Nucleotide Archive under study no. PRJEB27928. The metadata for these samples are available as **Supplementary Table 6**.

For the other studies included in the current study, the raw sequencing data can be found under the following European Nucleotide Archive identifiers: PRJEB10878 for [11]; PRJEB12449 for [10]; ERP008729 for [9]; and ERP005534 for [8]. The independent validation cohorts can be found in the Sequence Read Archive under the identifier no. SRP136711 for [24] and in the DNA Data Bank of Japan database under identification no. DRA006684.

The filtered taxonomic and functional profiles used as input for the statistical modeling pipeline are available in **Supplementary Data 1**.

The code and all analysis results can be found under https://github.com/zellerlab/crc_meta.

Author contributions

G.Z., M.A., and P.B. conceived and supervised the study. P.S.K., N.H., C.M.U., H.B., E.V., and R.S. recruited the participants and collected the samples. E.K., A.Y.V., S.Sunagawa, and P.B. generated the metagenomic data. A.M., P.T.P., J.S.F., A.P., S.Sunagawa, L.P.C., G.Z., and M.A. developed the metagenomic profiling workflows and/or performed the taxonomic and functional profiling. J.W., G.Z., K.Z., P.T.P., A.K., M.A., and N.S. performed the statistical analysis and/or developed the statistical analysis workflows. E.K. and R.P. designed and performed the validation experiments. A.M.T., P.M., S.G., D.S., S.M., H.S., S.Shiba, T.S., S.Y., T.Y., L.W., A.N., and N.S. provided additional validation data. J.W., G.Z., M.A., P.T.P., and P.B. designed the figures. G.Z., J.W., M.A., and P.B. wrote the manuscript with contributions from P.T.P., A.M., S.Sunagawa, L.P.C., E.K., A.Y.V., E.V., R.S., P.S.K., H.B., E.N., N.S. and L.W. All authors discussed and approved the manuscript.

Supplementary material

Extended Data figures and Supplementary material can be found online with the original article under https://doi.org/10.1038/s41591-019-0406-6

References

1. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6**, 805–814 (2005).
2. Tremaroli, V. & Bäckhed, F. Functional interactions between the gut microbiota and host metabolism. *Nature* **489**, 242–249 (2012).
3. Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N. Engl. J. Med.* **375**,

2369–2379 (2016).
4. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
5. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
6. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
7. Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol* **3**, 337–346 (2018).
8. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
9. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
10. Vogtmann, E. *et al.* Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One* **11**, e0155362 (2016).
11. 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).
12. Bedarf, J. R. *et al.* Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. *Genome Med.* **9**, 39 (2017).
13. Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **172**, 1198–1215 (2018).
14. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
15. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
16. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
17. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).
18. Shah, M. S. *et al.* Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* **67**, 882–891 (2018).
19. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).
20. Dai, Z. *et al.* Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* **6**, 70 (2018).
21. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
22. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
23. Kultima, J. R. *et al.* MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* **32**, 2520–2523 (2016).
24. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
25. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. A Lego System for Conditional Inference. *Am. Stat.* **60**, 257–263 (2006).
26. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
27. Tjalsma, H., Boleij, A., Marchesi, J. R. & Dutilh, B. E. A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects. *Nat. Rev. Microbiol.* **10**, 575–582 (2012).
28. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–93 (2016).

29. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–205 (2014).
30. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
31. Vieira-Silva, S., Falony, G., Darzi, Y. & Lima-Mendez, G. Species–function relationships shape ecological properties of the human gut microbiome. *Nature* (2016).
32. Hirayama, A. *et al.* Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. *Cancer Res.* **69**, 4918–4925 (2009).
33. Denkert, C. *et al.* Metabolite profiling of human colon carcinoma--deregulation of TCA cycle and amino acid turnover. *Mol. Cancer* **7**, 72 (2008).
34. Mal, M., Koh, P. K., Cheah, P. Y. & Chan, E. C. Y. Metabotyping of human colorectal cancer using two-dimensional gas chromatography mass spectrometry. *Anal. Bioanal. Chem.* **403**, 483–493 (2012).
35. Weir, T. L. *et al.* Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One* **8**, e70803 (2013).
36. Goedert, J. J. *et al.* Fecal metabolomics: assay performance and association with colorectal cancer. *Carcinogenesis* **35**, 2089–2096 (2014).
37. Aykan, N. F. Red meat and colorectal cancer. *Oncol. Rev.* (2015).
38. International, W. C. R. F. *Diet, nutrition, physical activity and cancer: a global perspective: a summary of the Third Expert Report*. (World Cancer Research Fund International, 2018).
39. Dutilh, B. E., Backus, L., van Hijum, S. A. F. T. & Tjalsma, H. Screening metatranscriptomes for toxin genes as functional drivers of human colorectal cancer. *Best Pract. Res. Clin. Gastroenterol.* **27**, 85–99 (2013).
40. Sears, C. L. & Garrett, W. S. Microbes, microbiota, and colon cancer. *Cell Host Microbe* **15**, 317–328 (2014).
41. Ridlon, J. M., Harris, S. C., Bhowmik, S., Kang, D.-J. & Hylemon, P. B. Consequences of bile salt biotransformations by intestinal bacteria. *Gut Microbes* **7**, 22–39 (2016).
42. Yoshimoto, S. *et al.* Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature* **499**, 97–101 (2013).
43. Ajouz, H., Mukherji, D. & Shamseddine, A. Secondary bile acids: an underrecognized cause of colon cancer. *World J. Surg. Oncol.* **12**, 164 (2014).
44. Boleij, A. *et al.* The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* **60**, 208–215 (2015).
45. Wu, S. *et al.* A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat. Med.* **15**, 1016–1022 (2009).
46. Dejea, C. M. *et al.* Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592–597 (2018).
47. Ridlon, J. M., Kang, D.-J. & Hylemon, P. B. Isolation and characterization of a bile acid inducible 7alpha-dehydroxylating operon in Clostridium hylemonae TN271. *Anaerobe* **16**, 137–146 (2010).
48. Mallonee, D. H., White, W. B. & Hylemon, P. B. Cloning and sequencing of a bile acid-inducible operon from Eubacterium sp. strain VPI 12708. *J. Bacteriol.* **172**, 7011–7019 (1990).
49. Ocvirk, S. & O'Keefe, S. J. Influence of Bile Acids on Colorectal Cancer Risk: Potential Mechanisms Mediated by Diet - Gut Microbiota Interactions. *Curr. Nutr. Rep.* **6**, 315–322 (2017).
50. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
51. Viennot, S. *et al.* Colon cancer in inflammatory bowel disease: recent trends, questions and answers. *Gastroenterol. Clin. Biol.* **33 Suppl 3**, S190–201 (2009).
52. Rubinstein, M. R. *et al.* Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe* **14**, 195–206 (2013).
53. Kostic, A. D. *et al.* Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).
54. Arthur, J. C. *et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
55. Reddy, B. S. Diet and excretion of bile acids. *Cancer Res.* **41**, 3766–3768 (1981).

56. Ogino, S., Nowak, J. A., Hamada, T., Phipps, A. I. & Peters, U. Integrative analysis of exogenous, endogenous, tumour and immune factors for precision medicine. *Gut* (2018).
57. Ogino, S., Chan, A. T., Fuchs, C. S. & Giovannucci, E. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. *Gut* **60**, 397–411 (2011).
58. Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., 4th, Koumpouras, C. C. & Schloss, P. D. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *MBio* **9**, (2018).
59. zur Hausen, H. Red meat consumption and cancer: reasons to suspect involvement of bovine infectious factors in colorectal cancer. *Int. J. Cancer* **130**, 2475–2483 (2012).
60. Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).
61. Böhm, J. *et al.* Discovery of novel plasma proteins as biomarkers for the development of incisional hernias after midline incision in patients with colorectal cancer: The ColoCare study. *Surgery* **161**, 808–817 (2017).
62. Liesenfeld, D. B. *et al.* Metabolomics and transcriptomics identify pathway differences between visceral and subcutaneous adipose tissue in colorectal cancer patients: the ColoCare study. *Am. J. Clin. Nutr.* **102**, 433–443 (2015).
63. Pox, C. P. *et al.* Efficacy of a nationwide screening colonoscopy program for colorectal cancer. *Gastroenterology* **142**, 1460–7.e2 (2012).
64. Furet, J.-P. *et al.* Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. *FEMS Microbiol. Ecol.* **68**, 351–362 (2009).
65. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
66. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
68. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
69. Smialowski, P., Frishman, D. & Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics* **26**, 440–443 (2010).
70. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
71. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
72. Oksanen, J. *et al.* Package 'vegan'. *Community ecology package, version* **2**, 1–295 (2013).
73. Costea, P. I., Zeller, G., Sunagawa, S. & Bork, P. A fair comparison. *Nature methods* vol. 11 359 (2014).
74. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer Science & Business Media, 2009).
75. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).
76. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
77. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
78. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
79. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
80. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4516–4522 (2011).

SIAMCAT publication:
Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox

Jakob Wirbel, Konrad Zych, Morgan Essex, Nicolai Karcher, Ece Kartal, Guillem Salazar, Peer Bork, Shinichi Sunagawa & Georg Zeller

Published as: Wirbel, J., Zych, K., Essex, M. et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol 22, 93 (2021).
https://doi.org/10.1186/s13059-021-02306-1

Abstract

The human microbiome is increasingly mined for diagnostic and therapeutic biomarkers using machine learning (ML). However, metagenomics-specific software is scarce, and overoptimistic evaluation and limited cross-study generalization are prevailing issues. To address these, we developed SIAMCAT, a versatile R toolbox for ML-based comparative metagenomics. We demonstrate its capabilities in a meta-analysis of fecal metagenomic studies (10,803 samples). When naively transferred across studies, ML models lost accuracy and disease specificity, which could however be resolved by a novel training set augmentation strategy. This reveals some biomarkers to be disease-specific, with others shared across multiple conditions. SIAMCAT is freely available from siamcat.embl.de.

Introduction

The study of microbial communities through metagenomic sequencing has begun to uncover how communities are shaped by—and interact with—their environment, including the host organism in the case of gut microbes[1,2]. Especially within a disease context, differences in human gut microbiome compositions have been linked to many common disorders, for example, colorectal cancer[3], inflammatory bowel disease[4,5], or arthritis[6,7]. As the microbiome is increasingly recognized as an important factor in health and disease, many possibilities for clinical applications are emerging for diagnosis[8,9], prognosis, or prevention of disease[10].

The prospect of clinical applications also comes with an urgent need for methodological rigor in microbiome analyses in order to ensure the robustness of findings. It is necessary to

assess the clinical value of biomarkers identified from the microbiome in an unbiased manner—not only by their statistical significance, but more importantly also by their prediction accuracy on independent samples (allowing for external validation). Machine learning (ML) models—ideally interpretable and parsimonious ones—are crucial tools to identify and validate such microbiome signatures. Setting up ML workflows however poses difficulties for novices. In general, it is challenging to assess their performance in an unbiased way, to apply them in cross-study comparisons, and to avoid confounding factors, for example, when disease and treatment effects are intertwined[11]. For microbiome studies, additional issues arise from key characteristics of metagenomic data such as large technical and inter-individual variation[12], experimental bias[13], compositionality of relative abundances, zero inflation, and non-Gaussian distribution, all of which necessitate data normalization in order for ML algorithms to work well.

While several statistical analysis tools have been developed specifically for microbiome data, they are generally limited to testing for differential abundance of microbial taxa between groups of samples and do not allow users to evaluate their predictivity as they do not comprise full ML workflows for biomarker discovery[14–16]. To overcome the limitations of testing-based approaches, several researchers have explicitly built ML classifiers to distinguish case and control samples[17–24]; however, the software resulting from these studies is generally not easily modified or transferred to other classification tasks or data types. To our knowledge, a powerful yet user-friendly computational ML toolkit tailored to the characteristics of microbiome data has not yet been published.

Here, we present SIAMCAT (Statistical Inference of Associations between Microbial Communities And host phenoTypes), a comprehensive toolbox for comparative metagenome analysis using ML, statistical modeling, and advanced visualization approaches. It also includes functionality to identify and visually explore confounding factors. To demonstrate its versatile applications, we conducted a large-scale ML meta-analysis of 130 classification tasks from 50 gut metagenomic studies (see **Table 1**) that have been processed with a diverse set of taxonomic and functional profiling tools. Based on this large-scale application, we arrive at recommendations for sensible parameter choices for the ML algorithms and preprocessing strategies provided in SIAMCAT. Moreover, we illustrate how several common pitfalls of ML applications can be avoided using the statistically rigorous approaches implemented in SIAMCAT. When considering the

cross-study application of ML models, we note prevailing problems with type I error control (i.e., elevated false-positive rate, abbreviated as FPR) as well as disease specificity for ML models naively transferred across datasets. To alleviate these issues, we propose a strategy based on sampling additional external controls during cross-validation (which we call control augmentation). This enables cross-disease comparison of gut microbial biomarkers. Lastly, we showcase how SIAMCAT facilitates meta-analyses in an application to fecal shotgun metagenomic data from five independent studies of Crohn's disease. SIAMCAT is implemented in the R programming language and freely available from siamcat.embl.de or Bioconductor.

Table 1: Overview of diseases and datasets included in the ML meta-analysis

Disease	Disease abbr.	Datasets	Data type
Ankylosing spondylitis	AS	[7]	Shotgun
Rheumatoid arthritis	ART	[25]	16S rRNA
		[6]	Shotgun
Type 1 diabetes	T1D	[26]	16S rRNA
Crohn's disease	CD	[5,27–30]	Shotgun
Ulcerative colitis	UC	[5,30,31]	Shotgun
Inflammatory bowel disease	IBD	[4,32–34]	16S rRNA
Colorectal cancer	CRC	[35–41]	Shotgun
		[35,42–44]	16S rRNA
Advanced colorectal adenoma(s)	ADA	[35,36,40,41]	Shotgun
Atherosclerotic cardiovascular disease	ACVD	[45]	Shotgun
Hypertension Pre-hypertension	HT pHT	[46]	Shotgun
Clostridioides difficile infection	CDI	[47,48]	16S rRNA
enteric diarrheal disease	EDD	[49]	16S rRNA
HIV infection	HIV	[50–52]	16S rRNA
Liver cirrhosis	LIV	[53]	Shotgun
		[54]	16S rRNA
Non-alcoholic fatty liver disease	NAFLD	[55,56]	Shotgun
		[57,58]	16S rRNA
Parkinsons' disease	PAR	[59]	Shotgun
		[60]	16S rRNA
Autism spectrum disorder	ASD	[61,62]	16S rRNA
Obesity	OB	[63]	Shotgun
		[64–67]	16S rRNA
Metabolic syndrome	MS	[68]	Shotgun
Type 2 diabetes	T2D	[69,70]	Shotgun
Impaired glucose tolerance	IGT	[69]	Shotgun

Results

Machine learning and statistical analysis workflows implemented in SIAMCAT

The SIAMCAT R package is a versatile toolbox for analyzing microbiome data from case-control studies. The default workflows abstract from and combine many of the complex steps that these workflows entail and that can be difficult to implement correctly for non-experts. To increase ease of use, SIAMCAT interfaces with the popular phyloseq package[71], and design and parameter choices are carefully adapted to metagenomic data analysis. In addition to functions for statistical testing of associations, SIAMCAT workflows include ML procedures, also encompassing data preprocessing, model fitting, performance evaluation, and visualization of the results and models (**Fig. 1a**). Core ML functionality is based on the mlr package[72]. The input for SIAMCAT consists of a feature matrix (abundances of microbial taxa, genes, or pathways across all samples), a group label (case-control information for all samples), and optional meta-variables (such as demographics, lifestyle, and clinical records of sample donors or technical parameters of data acquisition).

To demonstrate the main workflow and primary outputs of the SIAMCAT package (see the "Methods" section and SIAMCAT vignettes), we analyzed a representative dataset[73] consisting of 128 fecal metagenomes from patients with ulcerative colitis (UC) and non-UC controls (**Fig. 1**). UC is a subtype of inflammatory bowel disease (IBD), a chronic inflammatory condition of the gastrointestinal tract that has been associated with dramatic changes in the gut microbiome[5,74]. As input, we used species-level taxonomic profiles available through the curatedMetagenomicsData R package[75].

After data preprocessing (unsupervised abundance and prevalence filtering, **Fig. 1a** and the **Methods** section), univariate associations of single species with the disease are computed using the non-parametric Wilcoxon test (which has been shown for metagenomic data to reliably control the false discovery rate in contrast to many other tests proposed[76]), and the results are visualized (using the *check.associations* function). The association plot displays the distribution of microbial relative abundance, the significance of the association, and a generalized fold change as a non-parametric measure of effect size[39] (**Fig. 1b**).

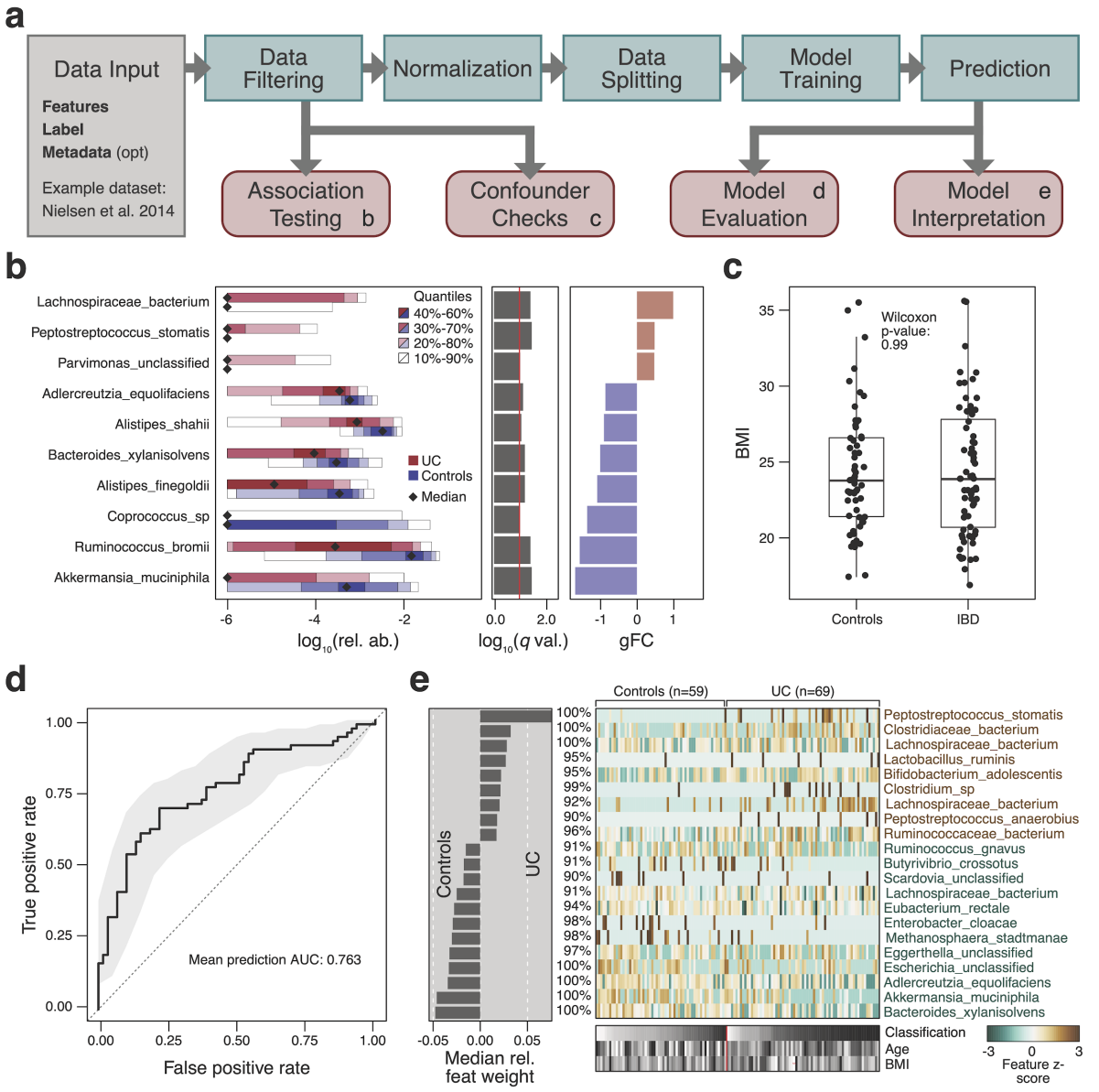


Fig. 1: Despite study differences, meta-analysis identifies a core set of gut microbes strongly associated with CRC.

a, The meta-analysis significance of gut microbial species derived from blocked Wilcoxon tests (n = 574 independent observations) is given by the bar height (FDR = 0.005). **b**, Underneath, species-level significance, as calculated with a two-sided Wilcoxon test (FDR-corrected *P* value), and the generalized fold change (Methods) within individual studies are displayed as heatmaps in gray and in color, respectively (see color bars and Table 1 for details on the studies included). Species are ordered by meta-analysis significance and direction of change. AT, Austria; CN, China; DE, Germany; FR, France; US, United States. **c**, For a core of highly significant species (meta-analysis FDR = 1 × 10−5), association strength is quantified by the AUROC across individual studies (color-coded diamonds), and the 95% confidence intervals are indicated by the gray lines. Family-level taxonomic information is color-coded above the species names (the numbers in brackets are mOTUs2 species identifiers; see Methods). **d**, Variance explained by disease status (CRC versus CTRLs) is plotted against variance explained by study effects for individual microbial species with dot size being proportional to abundance (see Methods); core microbial markers are highlighted in red.

The central component of SIAMCAT consists of ML procedures, which include a selection of normalization methods (*normalize.features*), functionality to set up a cross-validation scheme (*create.data.split*), and interfaces to different ML algorithms, such as LASSO, Elastic Net, and random forest (offered by the *mlr* package[72])[77–79]. As part of the cross-validation procedure, models can be trained (*train.model*) and applied to make predictions (*make.predictions*) on samples not used for training. Based on these predictions, the performance of the model is assessed (*evaluate.predictions*) using the area under the receiver operating characteristic (ROC) curve (AUROC) (**Fig. 1d**). SIAMCAT also provides diagnostic plots for the interpretation of ML models (*model.interpretation.plot*) which display the importance of individual features in the classification model, normalized feature distributions as heatmaps, next to sample meta-variables (optionally, see **Fig. 1c, e**).

Expert users can readily customize and flexibly recombine the individual steps in the described workflow above. For example, filtering and normalization functions can be combined or omitted before ML models are trained or association statistics calculated. To demonstrate its versatility beyond the workflow presented in **Fig. 1a**, we used SIAMCAT to reproduce two recent ML meta-analyses of metagenomic datasets[19,20]. By implementing the same workflows as described in the respective papers, we could generate models with very similar accuracy (within the 95% confidence interval) for all datasets analyzed (**Additional file 1: Figure S1**).

Confounder analysis using SIAMCAT

As many biological and technical factors beyond the primary phenotype of interest can influence microbiome composition[1], microbiome association studies are often at a high risk of confounding, which can lead to spurious results[11,80–82]. To minimize this risk, SIAMCAT provides a function to optionally examine potential confounders among the provided meta-variables. In the example dataset from[73], control samples were obtained from both Spanish and Danish subjects, while UC samples were only taken from Spanish individuals (**Fig. 2a**). Here, the meta-variable “country” could be viewed as a surrogate variable for other (often difficult-to-measure) factors, which can influence microbiome composition, such as diet, lifestyle, or technical differences between studies. The strong association of the “country” meta-variable with the disease status (SIAMCAT computes such associations

using Fisher's exact test or the Wilcoxon test for discrete and continuous meta-variables, respectively; see **Fig. 2a**) hints at the possibility that associations computed with the full dataset could be confounded by the country of the sample donor.

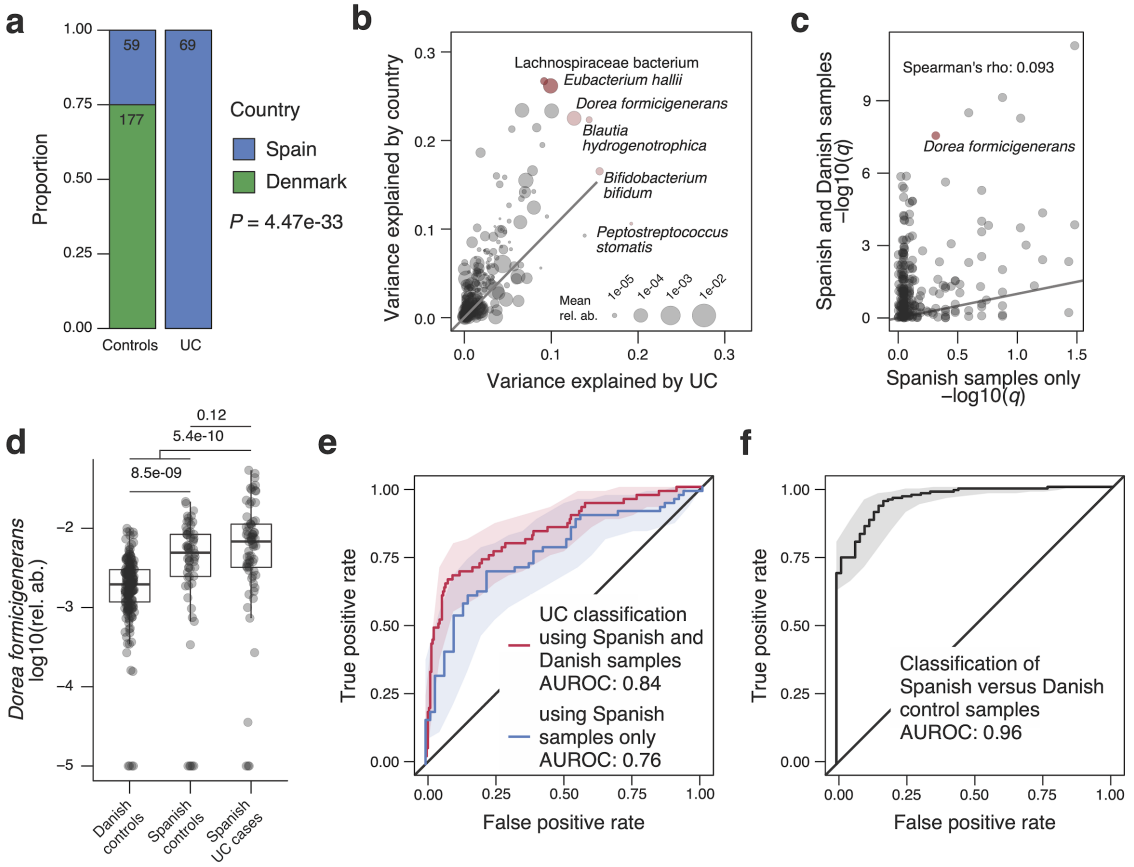


Fig. 2: Analysis of covariates that potentially confound microbiome-disease associations and classification models.
The UC dataset[73] contains fecal metagenomes from subjects enrolled in two different countries and generated using different experimental protocols (data shown is from *curatedMetagenomicData* with CD cases and additional samples per subject removed). **a**, Visualizations from the SIAMCAT confounder checks reveal that only control samples were taken from Denmark suggesting that any (biological or technical) differences between Danish and Spanish samples might confound a naive analysis for UC-associated differences in microbial abundances. **b**, Analysis of variance (using ranked abundance data) shows many species differ more by country than by disease, with several extreme cases highlighted. **c**, When comparing (FDR-corrected) *P* values obtained from SIAMCAT's association testing function applied to the whole dataset (y-axis) to those obtained for just the Danish samples (x-axis), only a very weak correlation is seen and strong confounding becomes apparent for several species including *Dorea formicigenerans* (highlighted). **d**, Relative abundance differences for *Dorea formicigenerans* are significantly larger between countries than between Spanish UC cases and controls (*P* values from Wilcoxon test) (see Fig. 1c for the definition of boxplots). **e**, Distinguishing UC patients from controls with the same workflow is possible with lower accuracy when only samples from Spain are used compared to the full dataset containing Danish and Spanish controls. This implies that in the latter case, the machine learning model is confounded as it exploits the (stronger) country differences (see c and f), not only UC-associated microbiome changes. **f**, This is confirmed by the result that control samples from Denmark and Spain can be very accurately distinguished with an AUROC of 0.96 (using SIAMCAT classification workflows)

To quantify this confounding effect on individual microbial features, SIAMCAT additionally provides a plot for each meta-variable that shows the variance explained by the label in comparison with the variance explained by the meta-variable for each individual feature (**Fig. 2b**, implemented in the *check.confounder* function). In our example case, several microbial species are strongly associated with both the disease phenotype (UC vs control) and the country, indicating that their association with the label might simply be an effect of technical and/or biological differences between samples taken and data processed in the different countries.

To further investigate this confounder, we used SIAMCAT to compute statistical association for the full dataset (including the Danish control samples) and the reduced dataset containing only samples from Spanish individuals (using the *check.association* function). The finding that *P* values were uncorrelated between the two datasets (**Fig. 2c**) directly quantified the effect of confounding by country on the disease-association statistic. The potential severity of this problem is highlighted by a comparison of the relative abundance of *Dorea formicigenerans* across subjects: the differences between UC cases and controls are only significant when Danish control samples are included, but not when restricted to Spanish samples only (**Fig. 2d**), exemplifying how confounders can lead to spurious associations.

Finally, confounding factors can not only bias statistical association tests, but can also impact the performance of ML models. A model trained to distinguish UC patients from controls seemingly performs better if the Danish samples are included (AUROC of 0.84 compared to 0.76 if only using Spanish samples), because the differences between controls and UC samples are artificially inflated by the differences between Danish and Spanish samples (**Fig. 2e**). How these overall differences between samples taken in different countries can be exploited by ML models can also be directly quantified using SIAMCAT workflows. The resulting model trained to distinguish between control samples from the two countries can do so with almost perfect accuracy (AUROC of 0.96) (**Fig. 2f**). This analysis demonstrates how confounding factors can lead to exaggerated performance estimates for ML models.

In summary, SIAMCAT can help to detect influential confounding factors that have the potential to bias statistical associations and ML model evaluations (see **Additional file 1: Figure S2** for additional examples).

Advanced machine learning workflows

When designing more complex ML workflows involving feature selection steps or applications to time series data, it becomes more challenging to set up cross-validation procedures correctly. Specifically, it is important to estimate how well a trained model would generalize to an independent test set, which is typically the main objective of microbial biomarker discovery. An incorrect ML procedure, in which information leaks from the test to the training set, can result in overly optimistic (i.e., overfitted) performance estimates. Two pitfalls that can lead to overfitting and poor generalization to other datasets (**Fig. 3a**) are frequently encountered in ML analyses of microbiome and other biological data, even though the issues are well described in the statistics literature[83–85]. These issues, namely supervised feature filtering and naive splitting of dependent samples, can be exposed by testing model performance in an external validation set, which has not been used during cross-validation at all (**Fig. 3b**).

The first issue arises when feature selection taking label information into account (supervised feature selection) is naively combined with subsequent cross-validation on the same data[84]. This incorrect procedure selects features that are associated with the label (e.g., by testing for differential abundance) on the complete dataset leaving no data aside for an unbiased test error estimation of the whole ML procedure. To avoid overfitting, correct supervised feature selection should always be nested into cross-validation (that is, the supervised feature selection has to be applied to each training fold of the cross-validation separately). To illustrate the extent of overfitting resulting from the incorrect approach, we used two datasets of colorectal cancer (CRC) patients and controls and performed both the incorrect and correct ways of supervised feature selection. As expected, the incorrect feature selection led to inflated performance estimates in cross-validation but lower generalization to an external dataset, whereas the correct procedure gave a better estimate of the performance in the external test set; the fewer features were selected, the more the performance in the external datasets dropped (see

Fig. 3c). SIAMCAT readily provides implementations of the correct procedure and additionally takes care that the feature filtering and normalization of the whole dataset are blind to the label (therefore called unsupervised), thereby preventing accidental implementation of the incorrect procedure.

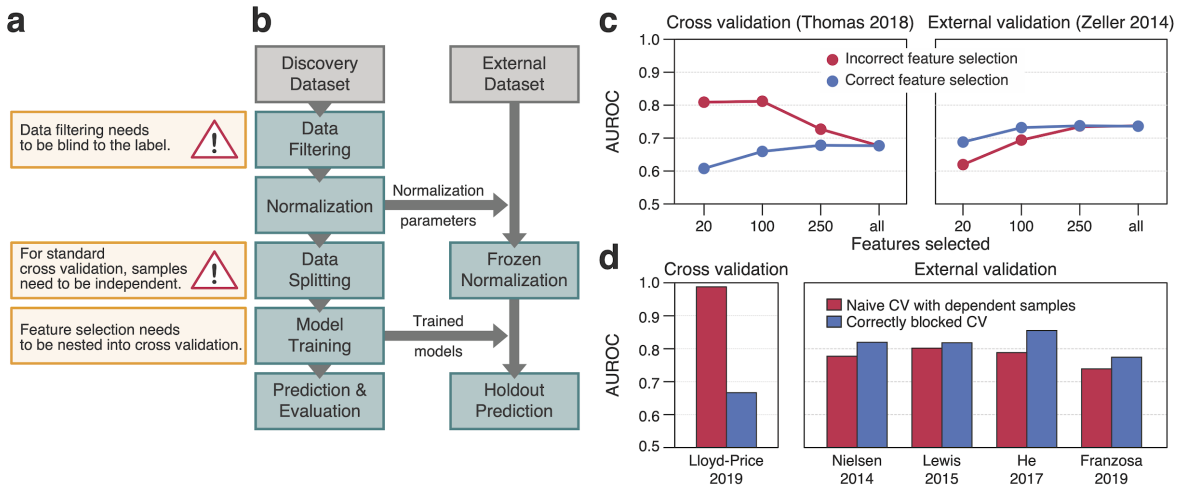


Fig 3: SIAMCAT aids in avoiding common pitfalls leading to a poor generalization of machine learning models.
a, Incorrectly setup machine learning workflows can lead to overoptimistic accuracy estimates (overfitting): the first issue arises from a naive combination of feature selection on the whole dataset and subsequent cross-validation on the very same data[83]. The second one arises when samples that were not taken independently (as is the case for replicates or samples taken at multiple time points from the same subject) are randomly partitioned in cross-validation with the aim to assess the cross-subject generalization error (see the main text). **b**, External validation, for which SIAMCAT offers analysis workflows, can expose these issues. The individual steps in the workflow diagram correspond to SIAMCAT functions for fitting a machine learning model and applying it to an external dataset to assess its external validation accuracy (see SIAMCAT vignette: holdout testing with SIAMCAT). **c**, External validation shows overfitting to occur when feature selection and cross-validation are combined incorrectly in a sequential manner, rather than correctly in a nested approach. The correct approach is characterized by a lower (but unbiased) cross-validation accuracy, but better generalization accuracy to external datasets (see header for datasets used). The fewer features are selected, the more pronounced the issue becomes, and in the other extreme case ("all"), feature selection is effectively switched off. **d**, When dependent observations (here by sampling the same individuals at multiple time points) are randomly assigned to cross-validation partitions, effectively the ability of the model to generalize across time points, but not across subjects, is assessed. To correctly estimate the generalization accuracy across subjects, repeated measurements need to be blocked, all of them either into the training or test set. Again, the correct procedure shows lower cross-validation accuracy, but higher external validation accuracy

The second issue tends to occur when samples are not independent[85]. For example, microbiome samples taken from the same individual at different time points are usually a lot more similar to each other than those from different individuals (see ref.[12] and **Additional file 1: Figure S3**). If these dependent samples are randomly split in a standard cross-validation procedure, so that some could end up in the training set and others in the test set, it is effectively estimated how well the model generalizes across time points (from the same individual) rather than across individuals. To avoid this, dependent measurements need to be blocked during cross-validation, ensuring that measurements of the same

individual are assigned to the same test set. How much the naive procedure can overestimate the performance in cross-validation and underperform in external validation compared to the correctly blocked procedure is demonstrated here using the iHMP dataset, which contains several samples per subject[30]. Although the cross-validation accuracy appears dramatically lower in the correct compared to the naive procedure, generalization to other datasets of the same disease is higher with the correctly blocked model (**Fig. 3d**). SIAMCAT offers the possibility to block the cross-validation according to meta-variables by simply providing an additional argument to the respective function call (see also SIAMCAT vignettes).

Large-scale machine learning meta-analysis

Previous studies that applied ML to microbiome data[17–20] have compared and discussed the performance of several learning algorithms. However, their recommendations were based on the analysis of a small number of datasets which were technically relatively homogeneous. To overcome this limitation and to demonstrate that SIAMCAT can readily be applied to various types of input data, we performed a large-scale ML meta-analysis of case-control gut metagenomic datasets. We included taxonomic profiles obtained with the RDP taxonomic classifier[86] for 26 datasets based on 16S rRNA gene sequencing[20]; additionally, taxonomic profiles generated from 12 and 24 shotgun metagenomic datasets using either MetaPhlAn2[87] or mOTUs2[88], respectively, as well as functional profiles obtained with HUMAnN2[89] or with eggNOG 4.5[90] for the same set of shotgun metagenomic data were included (in total 130 classification tasks, see **Table 1** and **Additional file 2: Table S1** for information about included datasets).

Focusing first on intra-study results, we found that given a sufficiently large input dataset (with at least 100 samples), SIAMCAT models are generally able to distinguish reasonably well between cases and controls: the majority (58%) of these datasets in our analysis could be classified with an AUROC of 0.75 or higher—compared to only 36% of datasets with fewer than 100 samples (**Fig. 4a–c**, **Additional file 1: Figures S4** and **S5** and the **Methods** section). Of note, accurate ML-based classification was possible even for datasets in which cases and controls could not easily be separated using beta-diversity analyses (**Additional file 1: Figure S6**), indicating that a lack of separation in ordination analysis does not

preclude ML-based workflows to extract accurate microbiome signatures. In the datasets for which a direct comparison of mOTUs2 and MetaPhlAn2 profiles was possible, we did not find any consistent trend towards either profiling method (paired Wilcoxon P=0.41, see **Additional file 1: Figure S7**). When comparing taxonomic and functional profiles derived from the same dataset, we found a high correlation between AUROC values (Pearson's r = 0.92, $P<2 \times 10-16$), although on average taxonomic profiles performed slightly better than functional profiles (**Additional file 1: Figure S7**). Taken together, this indicates that SIAMCAT can extract accurate microbiome signatures (model cross-validation AUROC > 0.75 in 64 of 130 classification tasks) from a range of different input profiles commonly used in microbiome research.

SIAMCAT provides various methods for data filtering and normalization and interfaces to several ML algorithms through mlr[72]. This made it easy to explore the space of possible workflow configurations in order to arrive at recommendations about sensible default parameters. To test the influence of different parameter choices within the complete data analysis pipeline, we performed an ANOVA analysis to quantify their relative importance on the resulting classification accuracy (**Fig. 4d** and the **Methods** section). Whereas the choice of filtering method and feature selection regime has little influence on the results, the normalization method and ML algorithm explained more of the observed variance in classification accuracy. Analysis of the different normalization methods shows that most of the differences can be explained by a drop in performance for naively normalized data (only total sum scaling and no further normalization) in combination with LASSO or Elastic Net logistic regression (**Additional file 1: Figure S8**). In contrast, the random forest classifier depended much less on optimal data normalization. Lastly, we compared the best classification accuracy for each classification task across the different ML algorithms. Interestingly, in contrast to a previous report[19], this analysis indicates that on average Elastic Net logistic regression outperforms LASSO and random forest classifiers when considering the optimal choice of ML algorithm (P=0.001 comparing Elastic Net to LASSO and $P=4 \times 10-14$ comparing it to random forest, **Fig. 4e**). In summary, this large-scale analysis demonstrates the versatility of the ML workflows provided by SIAMCAT and validates its default parameters as well as the robustness of classification accuracy to deviations from these.

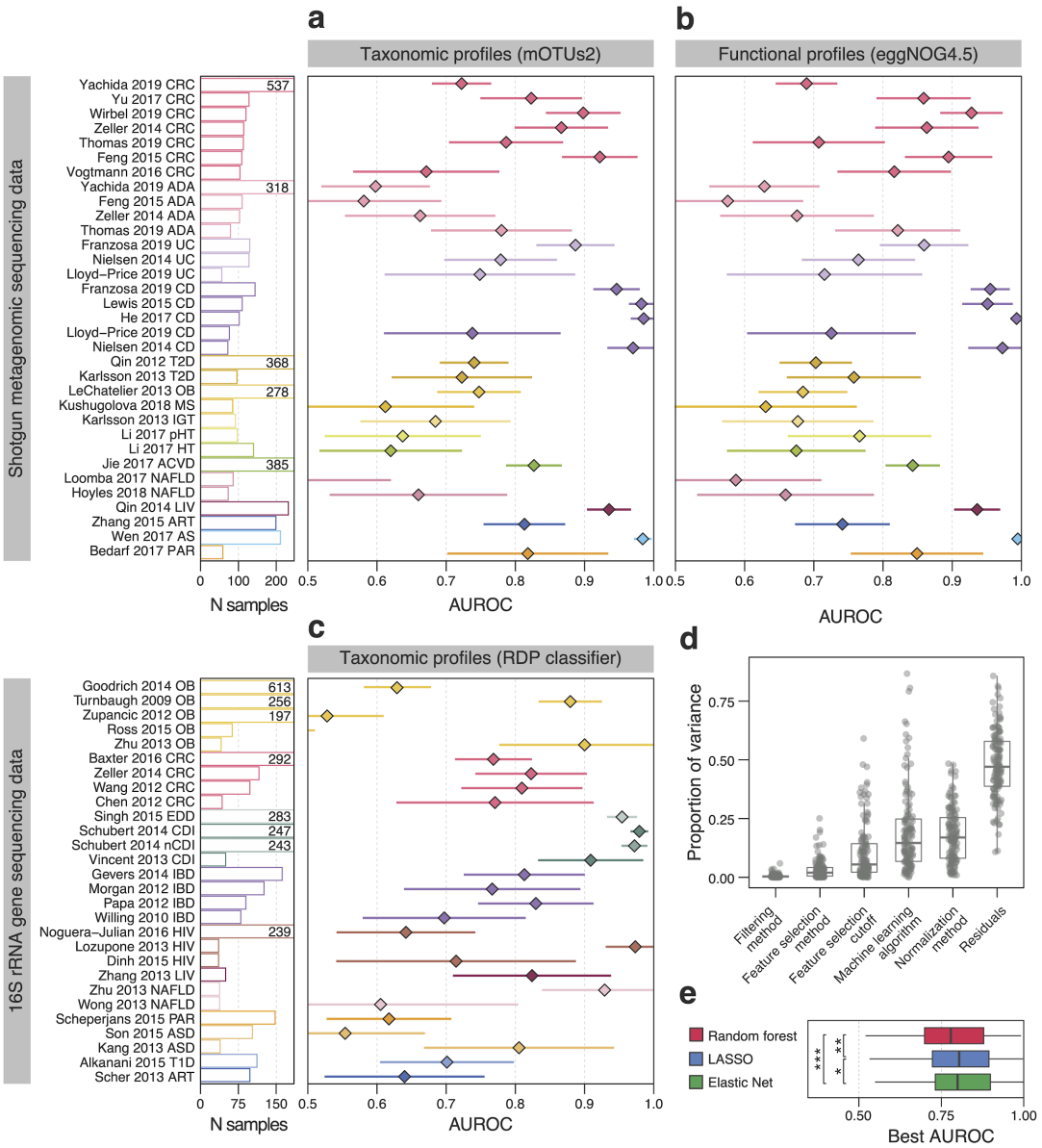


Fig.4: Large-scale application of the SIAMCAT machine learning workflow to human gut metagenomic disease association studies.

a, Application of SIAMCAT machine learning workflows to taxonomic profiles generated from fecal shotgun metagenomes using the mOTUs2 profiler. Cross-validation performance for discriminating between diseased patients and controls quantified by the area under the ROC curve (AUROC) is indicated by diamonds (95% confidence intervals denoted by horizontal lines) with sample size per dataset given as additional panel (cut at $N = 250$ and given by numbers instead) (see Table 1 and Additional file 2: Table S1 for information about the included datasets and key for disease abbreviations). **b**, Application of SIAMCAT machine learning workflows to functional profiles generated with eggNOG 4.5 for the same datasets as in a (see Additional file 1: Figure S4, S7 for additional types of and comparison between taxonomic and functional input data). **c**, Cross-validation accuracy of SIAMCAT machine learning workflows as applied to 16S rRNA gene amplicon data for human gut microbiome case-control studies [20] (see a for definitions). **d**, Influence of different parameter choices on the resulting classification accuracy. After training a linear model to predict the AUROC values for each classification task, the variance explained by each parameter was assessed using an ANOVA (see the Methods section) (see Fig. 1 for the definition of boxplots). **e**, Performance comparison of machine learning algorithms on gut microbial disease association studies. For each machine learning algorithm, the best AUROC values for each task are shown as boxplots (defined as in d). Generally, the choice of algorithm only has a small effect on classification accuracy, but both the Elastic Net and LASSO performance gains are statistically significant (paired Wilcoxon test: LASSO vs Elastic Net, $P = 0.001$; LASSO vs random forest, $P = 1e{-}08$; Elastic Net vs random forest, $P = 4e{-}14$)

Cross-study evaluation of microbiome signatures is crucial to establish their validity across patient populations. However, such comparisons are potentially hindered by inter-study differences in sample handling and data generation, with technical variation observed to often dominate over biological factors of interest[91–93]. Additionally, biological and clinical factors can contribute to inter-study differences. These not only include influences of geography, ethnicity, demographics, and lifestyle, but also how clinical phenotypes are defined and controls selected for each study[94].

Up to now, it has not been systematically explored how well microbiome-based ML models transfer across a range of diseases. To close this gap, we used our large-scale ML meta-analysis and trained ML models for each task using mOTUs2 taxonomic profiles as input (based on the previously established best-performing parameter set). We subsequently focused on models with reasonable cross-validation accuracy (AUROC > 0.75) and applied these to all remaining datasets to make predictions.

Cross-study application of ML models is straightforward within the same disease, since the model predictions on external datasets can easily be evaluated by an AUROC (**Additional file 1: Figure S9**, ref.[39,40]) under the assumption that case and control definitions are comparable between studies. However, when applying an ML model to a dataset from another disease, ROC analysis cannot be directly applied, since the cases the model was originally trained to detect are from another disease than those of the evaluation dataset. For this cross-disease application of ML models, we conducted extended evaluations, which specifically addressed the following two questions (see **Additional file 1: Figure S10** and the **Methods** section). First, we asked to which extent the separation between cases and controls (in terms of prediction scores) would be maintained when control samples of a different study are used. We therefore employed a modified ROC analysis (comparing true-positive rates from cross-validation to external FPRs via AUROC) as a newly defined measure of cross-study portability of an ML model. For convenience, we rescaled it to range between 0 (indicating a complete loss of discriminatory power on external data) and 1 (meaning that the ML model could be transferred to another dataset without loss of discrimination accuracy). Second, we asked how specific an ML model would be to the

disease it was trained to recognize, or whether its FPR would be elevated when presented with cases from a distinct condition. This is of interest in the context of an ongoing debate on whether there is a general gut microbial dysbiosis or distinct compositional changes associated with each disease[19,20,95]. Disease-specific classifiers would also be of clinical relevance when applied to a general population: due to large differences in disease prevalence, a model for CRC (a condition with low prevalence) misclassifying many type 2 diabetes (T2D) patients (high prevalence) would in the general population detect many more (false) T2D cases than true CRC cases, and thus have very low precision. To quantify the prediction rate for other diseases of an ML model, i.e., its disease specificity, we assessed how many samples from a distinct disease would be mispredicted as positive for the disease the ML model was trained on at a cutoff adjusted to maintain a FPR of 10% on the cross-validation set.

These extended evaluations showed low cross-study portability on the majority of external datasets (apparent also from a more than twofold increase in false positives on average) for most models (**Additional file 1: Figure S11**). Similarly, (false-positive) predictions for other diseases were elevated for most models (by a factor of 2.8 on average), with the extreme case of the ankylosing spondylitis (AS) model predicting more than 90% of cases from other diseases to be AS positive (median across studies, **Additional file 1: Figure S12**). These evaluations indicate that naive ML model transfer is substantially impacted—if not rendered impossible—by biological and technical study heterogeneity, apparent from loss of general accuracy and disease specificity.

Fig. 5 [next page]: Control augmentation improves ML model disease specificity and reveals shared and distinct predictors.
a, Schematic of the control augmentation procedure: control samples from external cohort studies are added to the individual cross-validation folds during model training. Trained models are applied to external studies (either of a different or the same disease) to determine cross-study portability (defined as maintenance of type I error control on external control samples) and cross-disease predictions (i.e., false detection of samples from a different disease). **b**, Cross-study portability was compared between naive and control-augmented models showing consistent improvements due to control augmentation. **c**, Boxplots depicting cross-study portability (left) and prediction rate for other diseases (right) of naive and control-augmented models (see Fig. 1 for the definition of boxplots). **d**, Heatmap showing prediction rates for other diseases (red color scheme) and for the same disease (green color scheme) for control-augmented models on all external datasets. True-positive rates of the models from cross-validation on the original study are indicated by boxes around the tile. Prediction rates over 10% are labeled.

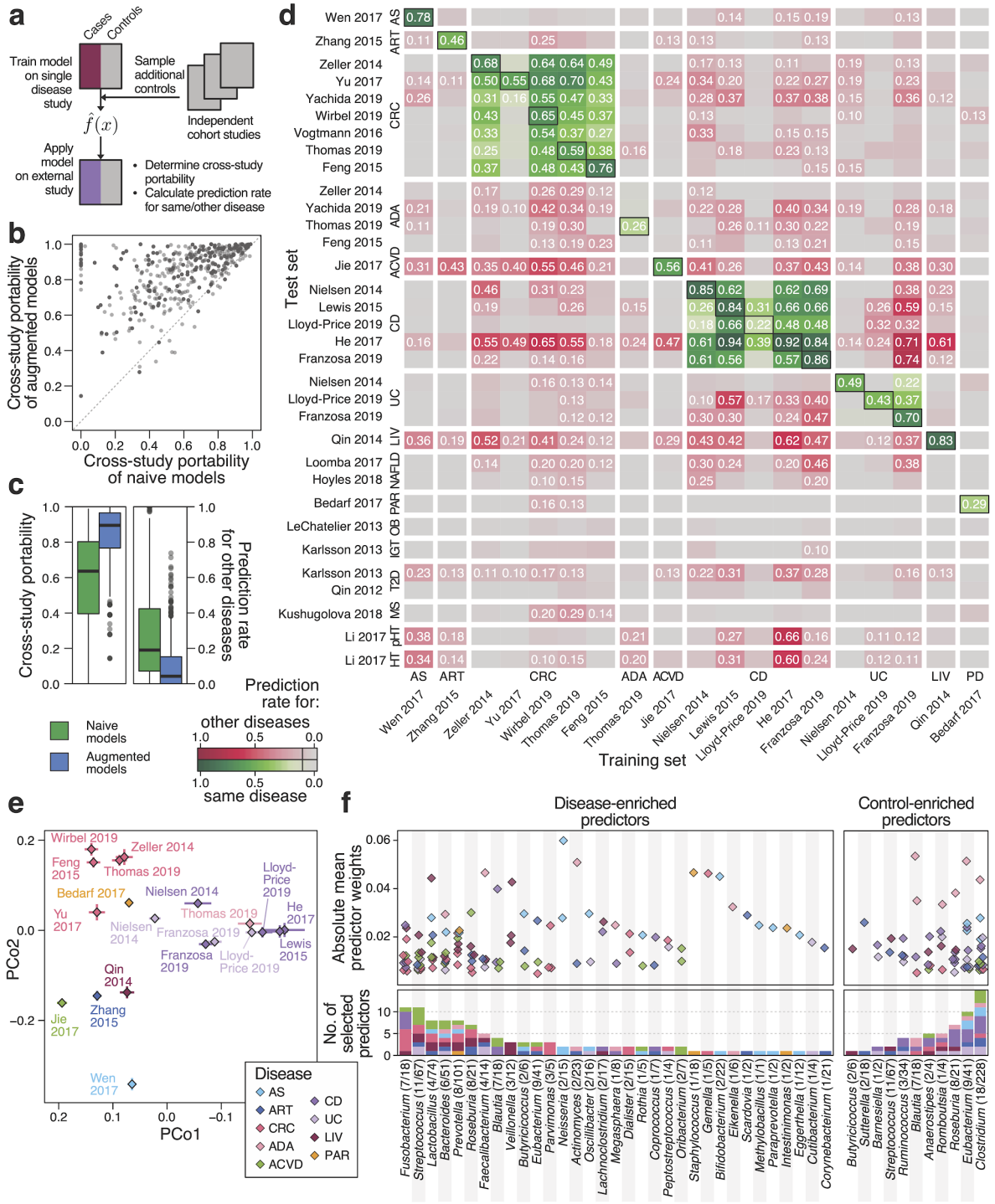


Fig. 5 [continued] e, Principal coordinate (PCo) analysis between models based on Canberra distance on model weights. Diamonds represent the mean per dataset in PCo space across cross-validation splits, and lines show the standard deviation. **f**, Visualization of the main selected model weights (predictors corresponding to mOTUs, see the Methods section for the definition of cutoffs) by genus and disease. Absolute model weights are shown as a dot plot on top, grouped by genus (including only genera with unambiguous NCBI taxonomy annotation). Below, the number of selected weights per genus is shown as a bar graph, colored by disease (see e for color key). Genus labels at the bottom include the number of mOTUs with at least one selected weight followed by the number of mOTUs in the complete model weight matrix belonging to the respective genus.

In order to improve the cross-study portability of ML models, we devised a strategy we call control augmentation, in which randomly selected control samples from independent microbiome population cohort studies[96–98] are added to the training set during model fitting (**Fig. 5a**, see the **Methods** section). This was motivated by the hypothesis that additional variability from a greater control pool comprising heterogeneous samples from multiple studies would enable classifiers to more specifically recognize disease signals while at the same time minimizing overfitting on peculiarities of a single dataset. However, a theoretical limitation of this approach is that the definition of controls can vary greatly across studies. In spite of this, in practice, we found control augmentation to greatly enhance cross-study portability uniformly across all ML models, both in cross-study analysis within the same condition and across different diseases (**Fig. 5b**, **c**, **Additional file 1: Figure S9**, **S11**). At the same time, cross-disease predictions decreased (**Fig. 5c**, **d**, **Additional file 1: Figure S12**) implying that it is an effective strategy to increase disease specificity of ML models.

The control augmentation strategy did not strongly depend on the set of controls used. We found large (> 250 samples) cohort studies to work well as a pool for control augmentation (allowing us to add five times the amount of control samples to each dataset). However, augmentation with fewer controls or with other datasets improved cross-study portability and disease specificity to almost the same effect (**Additional file 1: Figure S13**).
With cross-study portability greatly improved, we expect the remaining cross-disease predictions to be largely due to biological similarities between diseases rather than due to technical influences. In support of this, we show that CRC signatures have a tendency to cross-predict samples from patients with intestinal adenomas (ADA) or inflammatory bowel disease (CD), both of which are risk factors for CRC development[99]. Similarly, UC models cross-predict CD cases and vice versa, reflecting more general gut microbial changes, i.e., loss of beneficial commensal bacteria, that are shared across both types of inflammatory bowel disease[100]. In summary, we demonstrate that control augmentation is an effective strategy to broadly enable the validation of microbiome disease signatures across different studies, since it can overcome study-specific biases, which preclude the naive transfer of ML models.

When comparing microbiome signatures across diseases in more detail, we also revisited the question of whether microbiome alterations are specific to a disease, or signs of a general dysbiotic state[20]. As many ML algorithms, in particular (generalized) linear models, such as LASSO or Elastic Net logistic regression models, allow for model introspection, microbiome biomarkers can easily be extracted and their weight in the model directly quantified by (normalized) coefficient values. The model weights of the control-augmented models showed a clear clustering by disease in principal coordinate space revealing broad disease similarity patterns in terms of microbiome predictors that may reflect etiological similarities (**Fig. 5e**, not apparent from naively transferred ML models, **Additional file 1: Figure S14**). To obtain a more nuanced view of the gut bacterial taxa underlying these disease similarities, we analyzed individual mOTUs (grouped by genus membership) that were selected as predictors in disease models (**Fig. 5f**, to minimize bias from multiple studies of the same disease, we used the mean model for each disease and extracted those features whose weights accounted for more than 50% of the model, see the **Methods** section for details). We found some disease-enriched predictors to be very specific for a single disease, such as Veillonella spp. for LIV, Bifidobacteria and Neisseria mOTUs for AS, or Gemella and Parvimonas mOTUs for CRC. In contrast, species from other genera, for example, Lactobacillus, Bacteroides, or Fusobacteria, appear predictive of several diseases, although species and subspecies belonging to these vary in terms of their disease specificity (**Additional file 1: Figure S15**). Regarding control-enriched predictors, species from some genera are frequently depleted across multiple diseases (Anaerostipes and Romboutisa) while some diseases are marked by broad depletion of beneficial microbes, e.g., CD (consistent with ref.[100]).

Overall, enabled by control augmentation as an effective strategy to improve cross-study portability of ML models, our cross-disease meta-analysis reveals both shared and disease-specific predictors as a basis for further development of microbiome-based diagnostic biomarkers.

Meta-analysis of Crohn's disease gut microbiome studies

Microbiome disease associations being reported at an ever-increasing pace have also provided opportunities for comparisons across multiple studies of the same disease to assess the robustness of associations and the generalizability of ML models[19,20,39,40]. To

demonstrate SIAMCAT's utility in single-disease meta-analyses, we analyzed five metagenomic datasets[5,28–30,73], all of which included samples from patients with Crohn's disease (CD) as well as controls not suffering from inflammatory bowel diseases (IBD). Raw sequencing data were consistently processed to obtain genus abundance profiles with mOTUs2[88].

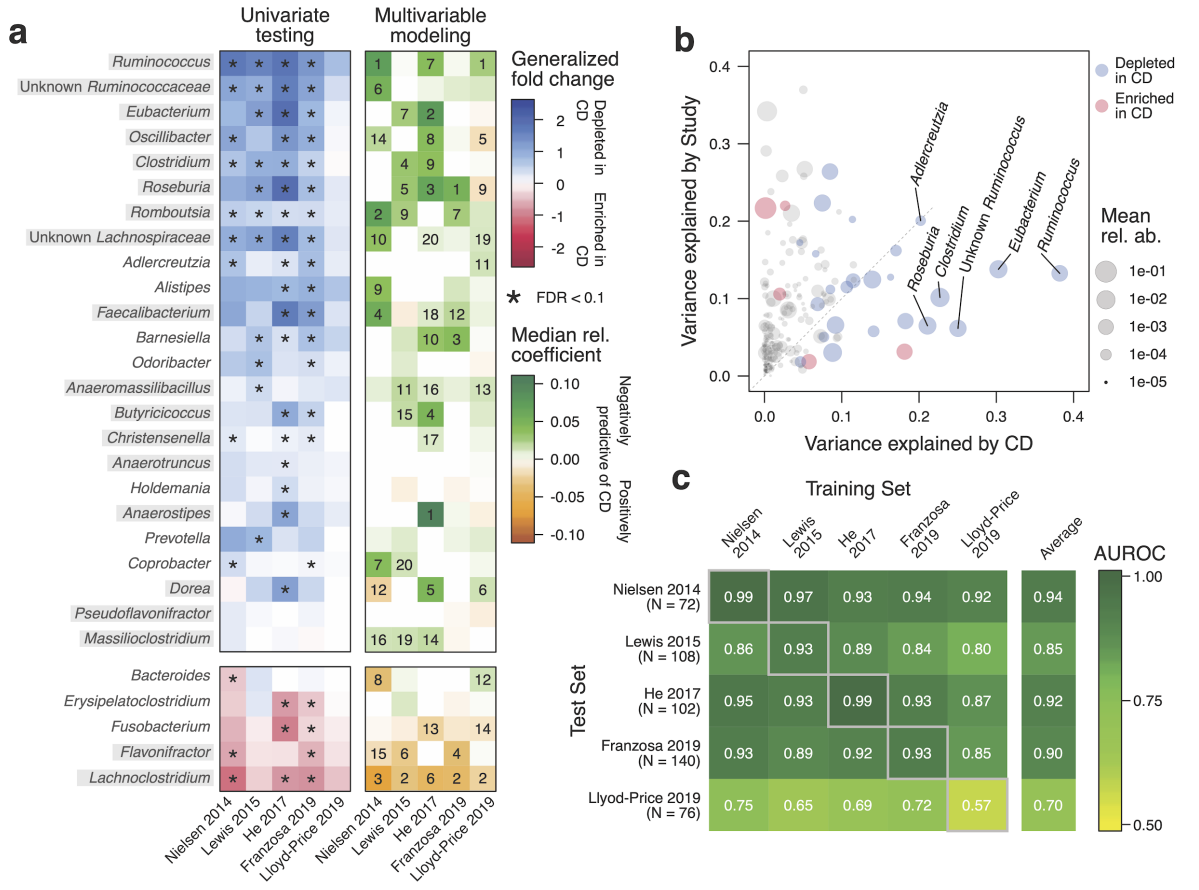


Fig. 6: Meta-analysis of CD studies based on fecal shotgun metagenomic data.
a, Genus-level univariate and multivariable associations with CD across the five included metagenomic studies. The heatmap on the left side shows the generalized fold change for genera with a single-feature AUROC higher than 0.75 or smaller than 0.25 in at least one of the studies. Associations with a false discovery rate (FDR) below 0.1 are highlighted by a star. Statistical significance was tested using a Wilcoxon test and corrected for multiple testing using the Benjamini-Hochberg procedure. Genera are ordered according to the mean fold change across studies, and genera belonging to the Clostridiales order are highlighted by gray boxes. The right side displays the median model weights for the same genera derived from Elastic Net models trained on the five different studies. For each dataset, the top 20 features (regarding their absolute weight) are indicated by their rank. **b**, Variance explained by disease status (CD vs controls) is plotted against the variance explained by differences between studies for individual genera. The dot size is proportional to the mean abundance, and genera included in a are highlighted in red or blue. **c**, Classification accuracy as measured by AUROC is shown as a heatmap for Elastic Net models trained on genus-level abundances to distinguish controls from CD cases. The diagonal displays values resulting from cross-validation (when the test and training set are the same), and off-diagonal boxes show the results from the study-to-study transfer of models

Based on SIAMCAT's *check.associations* function, we identified microbial genera that are significantly associated with CD in each study and visualized their agreement across studies (**Fig. 6a**, left panel). In line with previous findings[4], the gut microbiome of CD patients is characterized by a loss of diversity and many beneficial taxa. Though our re-analysis of the data from [30] could not identify any statistically significant genus-level associations, possibly due to the relatively small number of individuals or the choice of control samples obtained from patients with non-IBD gastrointestinal symptoms, the other four studies showed remarkable consistency among the taxa lost in CD patients, in particular, for members of the Clostridiales order.

We further investigated variation due to technical and biological differences between studies as a potential confounder using SIAMCAT's *check.confounder* function following a previously validated approach[39]. For many genera, variation can largely be attributed to heterogeneity among studies; the top five associated genera (cf. **Fig. 6a**), however, vary much more with disease status, suggesting that their association with CD is only minimally confounded by differences between studies (**Fig. 6b**).

Next, we systematically assessed cross-study generalization of ML models trained to distinguish CD patients from controls using SIAMCAT workflows. To this end, we trained an Elastic Net model for each study independently and evaluated the performance of the trained models on the other datasets (**Fig. 6c** and the **Methods** section). Most models maintained very high classification accuracy when applied to the other datasets for external validation (AUROC > 0.9 in most cases); again with the exception of the model cross-validated on the data from [30], which exhibited substantially lower accuracy in both cross-validation and external validation.

We lastly assessed the importance of individual microbial predictors in the CD models. The LASSO, and to some extent also the Elastic Net, are sparse models, in which the number of influential predictors (with non-zero coefficients) is kept small. As a consequence, these ML methods tend to omit statistically significant features when they are correlated to each other in favor of a smaller subset of features with optimal predictive power. Nonetheless, in our meta-analysis of CD, the feature weights derived from multivariable modeling

corresponded well to the univariate associations, and also showed some consistency across the four studies in which clear CD associations could be detected and an accurate ML model trained (**Fig. 6a**, right panel). Taken together, these results demonstrate that SIAMCAT could be a tool of broad utility for consolidating microbiome-disease associations and biomarker discovery by leveraging a large amount of metagenomic data becoming available for ML-based analyses.

Discussion

The rising interest in clinical microbiome studies and microbiome-derived diagnostic, prognostic, and therapeutic biomarkers also calls for more robust analysis procedures. An important step in that direction is the development of freely available, comprehensive, and extensively validated analysis workflows that make complex ML procedures available to non-experts, ideally while safeguarding against statistical analysis pitfalls. Designed with these objectives in mind, SIAMCAT provides a modular analysis framework that builds on the existing R-based microbiome analysis environment: data integration from DADA2[101] or phyloseq[71] is straightforward since SIAMCAT internally uses the phyloseq object. ML algorithms and procedures in SIAMCAT interface to the mlr package[72], a general-purpose ML library. Since the multitude of ML algorithms, workflow options, and design choices within such a general package can make ML workflow design challenging for non-experts, SIAMCAT mainly aims to enable users to apply robust and validated ML workflows to their data with preprocessing and normalization options tailored to the characteristics of microbiome data. At the same time, SIAMCAT allows advanced users to flexibly set up and customize more complex ML procedures, including non-standard cross-validation splits for dependent measurements and supervised feature selection methods that are properly nested into cross-validation (**Fig. 3**). Further developments of the package are planned to accommodate the rapidly changing needs of the microbiome research community, and updates will be published in accordance with the established Bioconductor release schedule.

To showcase the power of ML workflows implemented in SIAMCAT, we performed a meta-analysis of human gut metagenomic studies at a considerably larger scale than previous efforts[17–22] (see **Fig. 4**). It importantly encompassed a large number of diseases as

well as different taxonomic and functional profiles as input that were derived from different metagenomic sequencing techniques (16S rRNA gene and shotgun metagenomics sequencing) and profiling tools. Consequently, these benchmarks are expected to yield much more robust and general results than those from previous studies[17–22]. In our exploration of more than 7000 different parameter combinations per classification task (see the **Methods** section), we found the Elastic Net logistic regression algorithm to yield the highest cross-validation accuracies on average, albeit requiring the input data to be appropriately normalized (see **Fig. 4** and **Additional file 1: Figure S8**). Compared with the choice of normalization method and classification algorithm, other parameters had a considerably lower influence on the resulting classification accuracy. SIAMCAT's functionality to robustly fit statistical microbiome models and evaluate their performance will enable comparison to established diagnostic biomarkers[8] as an important prerequisite for further translation of microbiome research into the clinic.

To help resolve the debate about spurious associations and reproducibility issues in microbiome research[102], meta-analyses are crucial for the validation of microbiome biomarkers[39,40]. However, we found that ML models have substantial problems with type I error control (> 2-fold increase in FPR) and disease specificity (> 2.5-fold elevated FPR) when naively transferred across studies. We propose measures to detect these issues, which, if more widely adopted, could help to more precisely characterize them and their underlying causes. To address them, we introduce the control augmentation strategy, which greatly improved the cross-study portability of ML models. Being the first attempt to overcome study heterogeneity for improved cross-study model application, our work will hopefully stimulate further developments, which could easily be evaluated on the provided datasets. However, all such ML meta-analyses are limited by biological and clinical differences between studies[94], which will have to be addressed by better reporting standards[103]. Within these limitations, our ML meta-analysis datasets could become a valuable community resource for method development, systematic assessment of disease similarities, and further exploration of globally applicable microbiome biomarkers including validation of their disease specificity.

Using model introspection after control augmentation, we could revisit the question if microbiome alterations are specific to a given disease or more general hallmarks of dysbiosis[20]. In general, we found depletion of beneficial bacteria to be more often shared across several diseases (e.g., Anaerostipes or Romboutisa), in particular, in the subtypes of IBD. Conversely, disease-enriched bacteria were more often specific to a given disease. This could mean that some disease-specific microbiome alterations may reflect pathogens or pathobionts acting either as etiological agents or exploiting specific disease-related changes in the intestinal milieu. As examples of disease-specific markers, Parvimonas spp. are predictive for colorectal cancer, which is consistent with mechanistic work demonstrating this species to accelerate proliferation and cancer development both in vitro and in vivo[104]. Similarly, a putative link between oral Veillonella spp. and liver cirrhosis severity has been reported in the context of proton-pump inhibitor therapy[105], potentially enabled by increased transmission from the oral to the gut microbiome[81]. Other taxa showing a broader disease spectrum, such as Fusobacterium spp., have been extensively studied both in the context of CRC[106] and in IBD[107] using cellular and animal models. However, firmly establishing disease specificity or disease spectra for microbial biomarkers will be difficult to achieve in preclinical studies but require large patient cohorts. Nonetheless, our analyses generated candidates of both shared and disease-specific gut microbial biomarkers to guide further investigations of specific hypotheses on their ecological roles.

Although the analyses presented here are focused on human gut metagenomic datasets with disease prediction tasks, SIAMCAT is not restricted to these. It can also be applied to other tasks of interest in microbiome research, e.g., for investigating the effects of medication (see **Additional file 1: Figure S2**). Metagenomic or metatranscriptomic data from environmental samples can also be analyzed using SIAMCAT, e.g., to understand the associations between community composition and transcriptional activity of the ocean microbiome with physicochemical environmental properties (see **Additional file 1: Figure S16** for an example[108]) highlighting that SIAMCAT could be of broad utility in microbiome research.

Methods

Implementation

SIAMCAT is implemented as an R package with a modular architecture, allowing for a flexible combination of different functions to build ML and statistical analysis workflows (see the **Code box** section). The output of the functions (for example, the feature matrix after normalization) is stored in the SIAMCAT object, which is an extension of the phyloseq object that contains the raw feature abundances, meta-variables about the samples, and other optional information (for example, a taxonomy table or a phylogenetic tree) [71]. The label defining the sample groups for comparison is then derived from a user-specified meta-variable or an additional vector. ML models are trained using the mlr infrastructure as an interface to the implementations of different ML algorithms in other R packages [72]. SIAMCAT is available under the GNU General Public License, version 3.

Code box

Given two R objects called `feat` (relative abundance matrix) and `meta` (meta-variables about samples as a dataframe, containing a column called disease which encodes the label), the entire analysis can be conducted with a few commands (more extensive documentation can be found online in the SIAMCAT vignettes).

```
sc.obj <- siamcat(feat=feat, meta=meta, label='disease')
sc.obj <- filter.features(sc.obj, filter.method = 'abundance')
sc.obj <- check.associations(sc.obj,
    fn.plot = 'associations_plot.pdf')) # produces Fig. 1b
check.confounders(sc.obj,
    fn.plot = 'confounder_plot.pdf') # produces Fig. 1c
sc.obj <- normalize.features(sc.obj, norm.method = 'log.std')
sc.obj <- create.data.split(sc.obj)
sc.obj <- train.model(sc.obj, method='lasso')
sc.obj <- make.predictions(sc.obj)
sc.obj <- evaluate.predictions(sc.obj)
model.evaluation.plot(sc.obj,
    fn.plot = 'evaluation.pdf') # produces Fig. 1d
model.interpretation.plot(sc.obj, consens.thres = 0.8,
    fn.plot = 'interpretation.pdf') # produces Fig. 1e
```

Included datasets and microbiome profiling

In this study, we analyzed taxonomic and functional profiles derived with different profiling tools from several metagenomic datasets (see **Additional file 2: Table S1**). Taxonomic profiles generated using the RDP classifier [86] on the basis of 16S rRNA gene sequencing data were downloaded from a recent meta-analysis by [20] and summarized at the genus level. MetaPhlAn2 [87] and HUMAnN2 [89] taxonomic and functional profiles were obtained from the *curatedMetagenomicsData* R package [75] for all human gut datasets within the package that contained at least 20 cases and 20 controls. MetaPhlAn2 profiles were filtered to contain only species-level microbial taxa.

Additional datasets were profiled in-house with the following pipeline implemented in NGless [109]: after preprocessing with MOCAT2 [110] and filtering for human reads, taxonomic profiles were generated using the mOTUsv2 profiler [88], and functional profiles were calculated by first mapping reads against the integrated gene catalog [111] and then aggregating the results by eggNOG orthologous groups [90].
Additionally, genus-level taxonomic profiles from the TARA Oceans microbiome project [108] were used for two different classification tasks: to classify samples from polar and non-polar ocean regions and to classify samples based on their iron concentration at a depth of 5 m (high vs low iron content).

Primary package outputs and confounder analysis
To illustrate the main outputs of SIAMCAT, we analyzed the taxonomic profiles from a metagenomic study of IBD [73] included in the *curatedMetagenomicsData* R package [75]. For the analyses presented in **Fig. 1**, we restricted the dataset to control samples from Spain and cases with UC, since the two IBD subtypes included in the dataset (ulcerative colitis and Crohn's disease) are very different from one another in terms of the associated changes in the gut microbiome composition (see the SIAMCAT vignettes for more information or the **Code box** section for an outline of the basic SIAMCAT workflow.
To demonstrate how SIAMCAT can aid in confounder detection, we used the same dataset but this time included the Danish control samples in order to explore potential confounding by differences between samples collected and processed in these two countries. The analyses presented in **Fig. 2** have all been conducted with the respective functions of SIAMCAT (see SIAMCAT vignettes).

Machine learning hyperparameter exploration
To explore suitable hyperparameter combinations for ML workflows, we trained an ML model for each classification task and each hyperparameter combination. By hyperparameter, we mean configuration parameters of the workflow, such as normalization parameters, tuning parameters controlling regularization strength, or properties of the external feature selection procedure in contrast to model parameters fitted during the actual training of the ML algorithms. Specifically, we varied the filtering method (no data filtering; prevalence filtering with 1%, 5%, 10% cutoffs; abundance filtering with 0.001, 0.0001, and 0.0001 as cutoffs; and a combination of abundance and prevalence filtering), the normalization method (no normalization beyond the total sum scaling, log-transformation with standardization, rank-transformation with standardization, and centered log-ratio transformation), the ML algorithm (LASSO, Elastic Net, and random forest classifiers), and feature selection regimes (no feature selection and feature selection based on generalized fold change or based on single-feature AUROC; cutoffs were 25, 50, 100, 200, and 400 features for taxonomic profiles and 100, 500, 1000, and 2000 features for functional profiles). To cover the full hyperparameter space, we therefore trained 7488 models for taxonomic and 3168 models for functional datasets for each classification task.
To determine the optimal AUROC across input types (shown in **Fig. 4**), we calculated for each individual parameter combination the mean AUROC across all classification tasks with a specific type of input. Different feature filtering procedures could lead to cases in which the feature selection cutoffs were larger than the number of available features after filtering, therefore terminating the ML procedure. For that reason, we only considered those parameter combinations that did produce a result for all classification tasks with the specific type of input data.

To compare the importance of feature filtering, feature selection, normalization method, and ML algorithm on classification accuracy, we trained one linear model per classification task predicting the AUROC values from those variables. We then partitioned the variance attributable to each of these variables by calculating type III sums of squares using the Anova function from the car package in R [112].

In order to contrast the class separation of samples in distance space with the classification performance achieved by ML algorithms (see **Additional file 1: Figure S6**), we designed a distance-based measure of separation. For each dataset, we determined the distances between all pairs of samples within a class as well as all pairs of samples between classes and then calculated an AUROC value based on these two distributions. This distance-based measure effectively quantifies to what extent samples are closest to other samples from the same class (i.e., cluster together) and hence corresponds well to the visual separation of classes in ordination space (see **Additional file 1: Figure S6**).

Model transfer, cross-study portability, and prediction rate for other diseases

To assess cross-study portability and prediction rate for other diseases, ML models were applied to external datasets using the *make.predictions* function in SIAMCAT. In short, the function uses the normalization parameters of the discovery dataset to normalize the external data in a comparable way and then makes predictions by averaging the results of the application of all models of the repeated cross-validation folds to the normalized external data.

Cross-study portability is then calculated by comparing the predictions for cases in the discovery datasets and controls in the external dataset. First, the AUROC between these two prediction vectors is calculated, and values below 0.5 (when the predictions on controls in the external dataset are higher than predictions on cases in the discovery dataset) are set to 0.5. Cross-study portability is then defined as (|0.5 - AUROC|)*2 so that it afterwards ranges from 0 (no separation between cases and external controls or higher predictions on external controls) to 1 (perfect separation between cases and external controls).

To calculate the prediction rate for other diseases (or the same disease) on external datasets, a cutoff on the (real-valued) predictions is chosen so that the FPR in the discovery dataset is 0.1. Based on this cutoff, the external predictions are evaluated as positive (diseased) or negative predictions, and a detection rate corresponding to the fraction of positive predictions is determined.

Training Elastic Net models with control augmentation

To train models with the control augmentation strategy, we used the data from cohort microbiome studies as additional control samples [96–98]. Repeated measurements for the same individual were removed in the case of [96]. For each training set in the repeated cross-validation, we increased the number of control samples 5-fold by randomly sampling the appropriate number of controls (in a balanced manner between datasets to avoid overrepresentation of the larger external cohorts). Before addition, the additional control samples were normalized using the normalization parameters of the discovery set. Due to the introduction of additional variability, the control-augmented Elastic Net models were trained with a pre-set alpha value of 0.5 to ensure the stability of the model size.

To compare the predictors across different diseases, model weights of the control-augmented models were transformed into relative weights by dividing by the sum of absolute coefficient values. Then, models from the

same disease were averaged. Predictors (that is, mOTUs) were selected for display in **Fig. 5f**, if they (i) cumulatively contributed more than 50% of the mean relative disease model, (ii) their individual weights were bigger than 1%, and (iii) the genus annotation had an unambiguous NCBI taxonomy.

Illustration of common pitfalls in machine learning procedures

To demonstrate how naive sequential application of supervised feature selection and cross-validation might bias performance estimations, we trained LASSO ML models to distinguish colorectal cancer cases from controls based on MetaPhlAn2-derived species abundance profiles using the dataset with the handle ThomasAM_2018a [40] obtained through the *curatedMetagenomicsData* R package [75]. For the incorrect procedure of feature selection, single-feature AUROC values were calculated using the complete dataset (inverted for negatively associated features). Then, the features with the highest AUROC values were selected for model training (number depending on the cutoff). In contrast, the correct procedure implemented in SIAMCAT excludes the data in the test fold when calculating single-feature AUROC values; instead, AUROC values are calculated on the training fold only. To test generalization to external data, the models were then applied to another colorectal cancer metagenomic study [8] available through the *curatedMetagenomicsData* R package (also see the SIAMCAT vignette: holdout testing).

To illustrate the problem arising when combining naive cross-validation with dependent data, we used the Crohn's disease (CD) datasets used in the meta-analysis described below. We first subsampled the iHMP dataset [30] to five repeated measurements per subject, as some subjects had been sampled only five times and others more than 20 times. Then, we trained LASSO models using both a naive cross-validation and a cross-validation procedure in which samples from the same individual were always kept together in the same fold. External generalization was tested on the other four CD datasets described below.

Meta-analysis of Crohn's disease metagenomic studies

For the meta-analysis of Crohn's disease gut microbiome studies, we included five metagenomic datasets [5,28–30,73] that had been profiled with the mOTUs2 profiler [88] on the genus level. While some datasets contained both UC and CD patients [5,30,73], other datasets contained only CD cases [28,29]. Therefore, we restricted all datasets to a comparison between only CD cases and control samples, since the two subtypes of IBD are very different from each other.

For training of ML models, we blocked repeated measurements for the same individual when applicable [28,30,73]; specifically for the iHMP dataset [30], we also subsampled the dataset to five repeated measurements per individual to avoid biases associated with differences in the number of samples per individual. For external validation testing, we completely removed repeated measurements in order not to bias the estimation of classification accuracy.

To compute association metrics and to train and evaluate ML models, each dataset was encapsulated in an individual SIAMCAT object. To produce the plot showing the variance explained by label vs the variance explained by study, all data were combined into a single SIAMCAT object. The code to reproduce the analysis can be found in the SIAMCAT vignettes.

Author contributions

G.Z. conceived the study and prototyped the software. GZ., SS., and P.B. supervised the work. K.Z., J.W., and G.Z. implemented the software package with contributions from M.E., N. K, and E.K. J.W. and G.S. acquired the metagenomic data and/or performed the taxonomic and functional profiling. J.W., G.Z., and N.K. designed and performed the statistical analyses. J.W. and G.Z. designed the figures with help from N.K., M.E., and E.K. J.W., G.Z., and S.S. wrote the manuscript with contributions from P.B., M.E., N.K., G.S., E.K., and K.Z. All authors discussed and approved the final manuscript.

Supplementary material

Additional Files and Supplementary material can be found online with the original article under https://doi.org/10.1186/s13059-021-02306-1.

References

1. Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **172**, 1198–1215 (2018).
2. Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N. Engl. J. Med.* **375**, 2369–2379 (2016).
3. Garrett, W. S. The gut microbiota and colon cancer. *Science* **364**, 1133–1135 (2019).
4. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
5. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293–305 (2019).
6. Zhang, X. *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* **21**, 895–905 (2015).
7. Wen, C. *et al.* Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* **18**, 142 (2017).
8. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
9. Tilg, H., Cani, P. D. & Mayer, E. A. Gut microbiome and liver diseases. *Gut* **65**, 2035–2044 (2016).
10. Zitvogel, L., Ma, Y., Raoult, D., Kroemer, G. & Gajewski, T. F. The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. *Science* **359**, 1366–1370 (2018).
11. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
12. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16**, 73 (2015).
13. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *Elife* **8**, e46923 (2019).
14. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
15. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
16. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
17. Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C. & Knight, R. Human-associated microbial signatures: examining their predictive value. *Cell Host Microbe* **10**, 292–296 (2011).
18. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359 (2011).

19. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).
20. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).
21. Wang, J. *et al.* Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. *Microbiome* **6**, 101 (2018).
22. Bang, S. *et al.* Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. *Sci. Rep.* **9**, 10189 (2019).
23. Zhou, Y.-H. & Gallins, P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Front. Genet.* **10**, 579 (2019).
24. Le Goallec, A. *et al.* A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. *PLoS Comput. Biol.* **16**, e1007895 (2020).
25. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2**, e01202 (2013).
26. Alkanani, A. K. *et al.* Alterations in Intestinal Microbiota Correlate With Susceptibility to Type 1 Diabetes. *Diabetes* **64**, 3510–3520 (2015).
27. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
28. Lewis, J. D. *et al.* Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host Microbe* **18**, 489–500 (2015).
29. He, Q. *et al.* Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* **6**, 1–11 (2017).
30. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
31. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
32. Willing, B. P. *et al.* A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* **139**, 1844–1854.e1 (2010).
33. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
34. Papa, E. *et al.* Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS One* **7**, e39242 (2012).
35. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, (2014).
36. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
37. 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).
38. Vogtmann, E. *et al.* Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One* **11**, e0155362 (2016).
39. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
40. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
41. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
42. Wang, T. *et al.* Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* **6**, 320–329 (2012).
43. Chen, W., Liu, F., Ling, Z., Tong, X. & Xiang, C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS One* **7**, e39743 (2012).

44. Baxter, N. T., Ruffin, M. T., 4th, Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 37 (2016).
45. Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.* **8**, 845 (2017).
46. Li, J. *et al.* Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome* **5**, 14 (2017).
47. Schubert, A. M. *et al.* Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. *MBio* **5**, e01021–14 (2014).
48. Vincent, C. *et al.* Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. *Microbiome* **1**, 18 (2013).
49. Singh, P. *et al.* Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome* **3**, 45 (2015).
50. Lozupone, C. A. *et al.* Alterations in the gut microbiota associated with HIV-1 infection. *Cell Host Microbe* **14**, 329–339 (2013).
51. Dinh, D. M. *et al.* Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *J. Infect. Dis.* **211**, 19–27 (2015).
52. Noguera-Julian, M. *et al.* Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* **5**, 135–146 (2016).
53. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
54. Zhang, Z. *et al.* Large-scale survey of gut microbiota associated with MHE Via 16S rRNA-based pyrosequencing. *Am. J. Gastroenterol.* **108**, 1601–1611 (2013).
55. Loomba, R. *et al.* Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. *Cell Metab.* **25**, 1054–1062.e5 (2017).
56. Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat. Med.* **24**, 1070–1080 (2018).
57. Wong, V. W.-S. *et al.* Molecular characterization of the fecal microbiota in patients with nonalcoholic steatohepatitis--a longitudinal study. *PLoS One* **8**, e62885 (2013).
58. Zhu, L. *et al.* Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH. *Hepatology* **57**, 601–609 (2013).
59. Bedarf, J. R. *et al.* Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. *Genome Med.* **9**, 39 (2017).
60. Scheperjans, F. *et al.* Gut microbiota are related to Parkinson's disease and clinical phenotype. *Mov. Disord.* **30**, 350–358 (2015).
61. Kang, D.-W. *et al.* Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. *PLoS One* **8**, e68322 (2013).
62. Son, J. S. *et al.* Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons Simplex Collection. *PLoS One* **10**, e0137725 (2015).
63. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
64. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
65. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
66. Zupancic, M. L. *et al.* Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. *PLoS One* **7**, e43052 (2012).
67. Ross, M. C. *et al.* 16S gut community of the Cameron County Hispanic Cohort. *Microbiome* **3**, 7 (2015).
68. Kushugulova, A. *et al.* Metagenomic analysis of gut microbial communities from a Central Asian population. *BMJ Open* **8**, e021682 (2018).
69. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
70. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
71. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).

72. Bischl, B., Lang, M., Kotthoff, L. & Schiffner, J. mlr: Machine Learning in R. *The Journal of Machine* **17**, 1–5 (2016).
73. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
74. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13780–13785 (2007).
75. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
76. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20**, 210–221 (2019).
77. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
78. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).
79. Tin Kam Ho. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* vol. 1 278–282 vol.1 (ieeexplore.ieee.org, 1995).
80. Deloris Alexander, A. *et al.* Quantitative PCR assays for mouse enteric flora reveal strain-dependent differences in composition that are influenced by the microenvironment. *Mamm. Genome* **17**, 1093–1104 (2006).
81. Imhann, F. *et al.* Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2016).
82. Jackson, M. A. *et al.* Proton pump inhibitors alter the composition of the gut microbiota. *Gut* **65**, 749–756 (2016).
83. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: Data mining, inference, and prediction, second edition*. (Springer, 2009).
84. Smialowski, P., Frishman, D. & Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics* **26**, 440–443 (2010).
85. Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).
86. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
87. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
88. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
89. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
90. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–93 (2016).
91. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
92. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nature Biotechnology* **35**, 1077–1086 (2017).
93. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
94. Thompson, S. G. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* **309**, 1351–1355 (1994).
95. Olesen, S. W. & Alm, E. J. Dysbiosis is not an answer. *Nature Microbiology* **1**, 16228 (2016).
96. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
97. Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1897 (2016).
98. Xie, H. *et al.* Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the

Gut Microbiome. *Cell Syst* **3**, 572–584.e3 (2016).
99. Bernstein, C. N., Blanchard, J. F., Kliewer, E. & Wajda, A. Cancer risk in patients with inflammatory bowel disease: a population-based study. *Cancer* **91**, 854–862 (2001).
100. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205–211 (2006).
101. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
102. Cani, P. D. Gut microbiota - at the intersection of everything? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 321–322 (2017).
103. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
104. Yu, J. *et al.* The role of Parvimonas micra in intestinal tumorigenesis in germ-free and conventional APCmin/+ mice. *J. Clin. Orthod.* **37**, 531–531 (2019).
105. Horvath, A. *et al.* Biomarkers for oralization during long-term proton pump inhibitor therapy predict survival in cirrhosis. *Sci. Rep.* **9**, 12000 (2019).
106. Rubinstein, M. R. *et al.* Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe* **14**, 195–206 (2013).
107. Ohkusa, T. *et al.* Induction of experimental ulcerative colitis by Fusobacterium varium isolated from colonic mucosa of patients with ulcerative colitis. *Gut* **52**, 79–83 (2003).
108. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).
109. Coelho, L. P. *et al.* NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. *Microbiome* **7**, 84 (2019).
110. Kultima, J. R. *et al.* MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* **32**, 2520–2523 (2016).
111. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
112. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (SAGE Publications, 2018).

Benchmarking manuscript:
Evaluation of microbiome association models under realistic and confounded conditions

Jakob Wirbel[1,*], Morgan Essex[2,*], Sofia Kirke Forslund[1,2,3,4,5,†], Georg Zeller[1,†]

1. Structural and Computational Biology Unit (SCB), EMBL Heidelberg, Germany
2. Experimental and Clinical Research Center, a cooperation of Charité-Universitätsmedizin and the Max-Delbrück Center, Berlin, Germany
3. Max Delbrück Center for Molecular Medicine (MDC), Berlin, Germany
4. Charité – Universitätsmedizin Berlin, Berlin, Germany
5. DZHK (German Centre for Cardiovascular Research), partner site Berlin

* both authors contributed equally to the work
† correspondence should be addressed to Sofia.Forslund@mdc-berlin.de or zeller@embl.de

Unpublished manuscript; an updated version of this manuscript has been uploaded to bioRxiv as a preprint: https://doi.org/10.1101/2022.05.09.491139

Abstract

Background: Metagenome-wide association studies have uncovered characteristic changes in gut microbiome composition across various diseases. The correct statistical methodology to test for differential abundance of microbial features is however still debated, since unbiased and realistic benchmarks are missing.

Methods: Here, we developed a framework for benchmarking differential abundance testing methods based on implanting signals into a real baseline dataset that combines the advantages of a ground truth with the statistical attributes of real metagenomic data. We extensively validated the realism of the resulting samples and additionally included realistic patterns of confounding in our benchmark.

Results: We demonstrated a dramatic issue with elevated false discovery rates for the majority of methods with the exception of limma, linear models, and the Wilcoxon test. Even these methods recorded an abundance of false discoveries under confounded conditions, behavior which could be alleviated by linear mixed effect models or the blocked Wilcoxon test.

Conclusion: The results from our benchmark show that real metagenomic data represent major challenges for method development which have thus far been insufficiently addressed. We provide the accompanying software to enable other researchers to more

rigorously validate newly developed differential abundance testing methods, and we hope that our results contribute to more consensus in the field.

Introduction

The human gut microbiome is increasingly understood to carry out critical roles in host physiology and immunity, and has therefore been extensively mined for biomarkers related to host health and disease states. The composition of gut microbial communities is highly variable in time and between individuals[1], yet clinical metagenome-wide association studies (MWAS) strive to associate taxonomic and functional features with grouping variables such as disease phenotypes or lifestyle factors. MWAS typically perform statistical tests for (mean) differential abundance (DA) on each microbial feature independently, borrowing both nomenclature and entire methods from differential gene expression analysis. DA methods applied in MWAS loosely fall into four categories: a) methods borrowed or adapted from RNA-Seq analysis, b) linear models, c) rank-based statistical tests, and finally d) methods developed specifically for microbiome data. Researchers have performed MWAS in the context of numerous diseases, including but not limited to inflammatory bowel diseases[2,3], gastrointestinal cancers[4,5], and cardiometabolic diseases[6], while multi-omics and experimental validation studies have shed further light on specific hypotheses spawned from MWAS[7,8]. To identify group-specific microbiome alterations, MWAS typically perform statistical tests for (mean) differential abundance (DA) on each microbial feature independently, borrowing both nomenclature and entire methods from differential gene expression analysis. DA methods applied in MWAS loosely fall into four categories: a) methods adapted from RNA-Seq analysis, b) generalized linear models, c) rank-based statistical tests, and finally d) methods developed specifically for microbiome data.

While significant microbiome disease associations have been reported in many studies, some meta-analyses and cross-disease comparisons have suggested many of these associations to be unspecific or confounded[9–11]. Microbiome composition varies not only with host health and disease states, but also with myriad other host and environmental factors (covariates) collectively estimated to explain nearly 20% of microbiome variation[12]. Lifestyle and physiological covariates demonstrating the largest effects on microbial communities include medication regimens[13,14], stool quality, geography, and alcohol consumption frequency[15]. Technical differences such as stool sample collection and DNA

extraction method often outweigh biological factors of interest in terms of explained variation and can therefore hamper meta-analyses if unaccounted for[5,16,17].

Although the unique statistical challenges intrinsic to high-throughput metagenomic data are well-described by now[18–20], there is no consensus about the most appropriate DA procedures in the literature[21–28]. In principle, this is the purpose of benchmarking studies, which try to assert how methods perform under varying yet controlled conditions in order to establish points of reference for method behavior that cannot be discerned from single applications. To achieve such conditions, simulated data is typically generated from parametric models, which cleanly specify the differentially abundant features required for performance evaluation as a ground truth.

A high-level concern with benchmarking is the phenomenon of over-optimistic performance evaluation observed in newly introduced bioinformatic methods, whereby data, results, and competitor tools in a benchmarking study (frequently published in the same manuscript as the new method) suffer from selection and reporting bias[29]. The lack of consensus on how to simulate data in microbiome research, however, is a much more fundamental problem; an evaluation of simulation methods on the basis of their resemblance to experimental data and impact on downstream applications has not yet been conducted.

Here, we quantitatively assess the degree to which parametric simulations employed in previous benchmarks lack biological realism, and show that the choice of simulation framework can explain the divergent recommendations regarding DA methods. To address these shortcomings, we propose a novel simulation technique using real data to implant calibrated differential abundance between two groups (imitating a case-control MWAS), and extend it to additionally include confounded effects. Based on these more realistic simulations, we perform a comprehensive benchmarking study of widely used DA methods revealing an alarmingly common inability to control the false discovery rate (FDR). Under confounded conditions, FDR control is unattainable for all methods, but we show that some methods can be adjusted to suffer less from these issues.

Results

Assessment of realism for parametric simulations of microbiome data

As a first step, we aimed to explicitly evaluate how data generated from previous simulation frameworks compare with real metagenomic data. To do so, we simulated taxonomic

profiles using the source code employed in previous benchmarks[22,24,25,30], whereby case-control datasets were repeatedly generated with differentially abundant features introduced under varying effect sizes in order to quantify the uncertainty of the stochastic simulation process (see **Methods**). Simulation parameters were estimated in each case from the same baseline dataset of healthy adults[31]. We observed the data simulated with every one of the tested parametric models to be very different from real data as visualized by principal coordinate analysis (see **Fig. 1a**). Additionally, there was a large discrepancy between the feature variance and sparsity of simulated profiles and what is observed in real metagenomic data (see **Fig. 1b** and **SFig. 1**), with especially the multinomial method underestimating feature variance. Finally we trained machine learning classifiers to distinguish between real and simulated samples and could do so with almost perfect accuracy in nearly all cases, except for data generated from sparseDOSSA[30] (**Fig. 1c**). This was motivated by the fact that machine learning classification, commonly employed in MWAS to detect biomarkers, can reveal even subtle differences between groups and is generally more sensitive than ordination-based analyses[11]. Overall, all of the assessed parametric simulation frameworks failed to produce realistic metagenomic data.

Feature implantation yields realistic benchmarking datasets

To devise a simulation framework which would generate data that closely recapitulates key characteristics of metagenomic data, we opted to manipulate real baseline data as little as possible, by implanting a known signal with pre-defined effect size into a small number of differentially abundant features using randomly selected groups (see **Methods**). As the baseline dataset, we chose a cohort consisting of healthy adults without obvious biological groupings, into which we repeatedly implanted DA features by multiplying the counts in one group with a constant (abundance scaling) and/or by shuffling a certain percentage of non-zero entries across groups (prevalence shift, see **Methods**). The main advantage of this proposed signal implantation approach as a foundation for benchmarking is that it generates a clearly defined ground truth of DA features while retaining key characteristics of real data. In particular, feature variance and sparsity (see **Fig. 1b**) are preserved, which is reflected in both the principal coordinate projection (**Fig. 1a**) and in the more sensitive machine learning classification task (**Fig. 1c**).

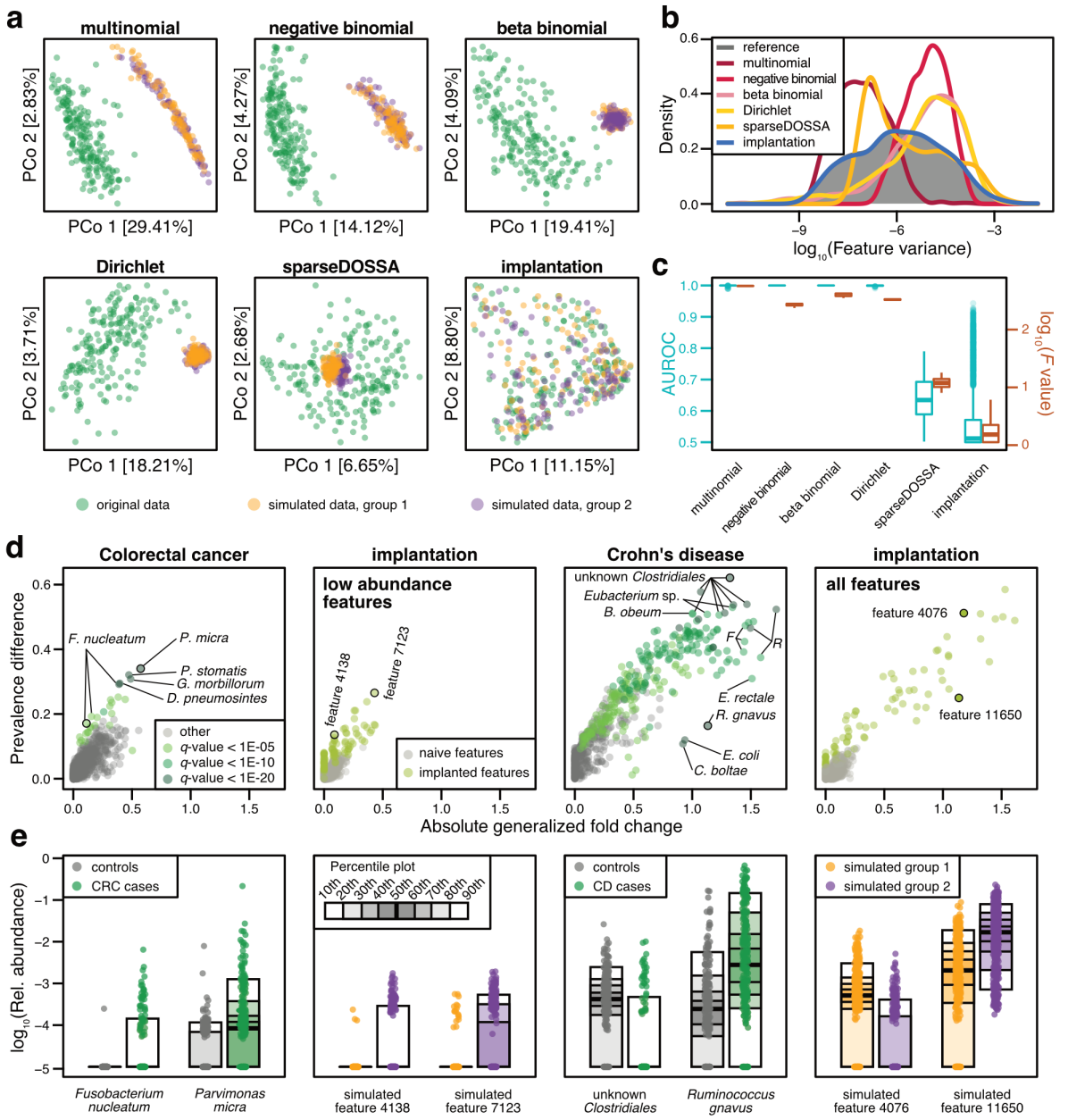


Figure 1: Feature implantation, but not parametric simulations, can reproduce key characteristics of metagenomic data and realistic disease effects

a) Principal coordinate projections on log-Euclidean distances for real samples (from Zeevi et al.[31], which served as a baseline data set) and representative samples of data simulated in a case-control setting (group 1 and 2) using different parametric models or feature implantation. For each method, the results from a single repetition and a fixed effect size are shown (see **Methods**, abundance scaling factor of 2 and prevalence shift of 0.3, if applicable). **b)** The log-transformed feature variance is shown for the real and simulated data from the same simulated data as in a). **c)** The AUROC values from machine learning models to distinguish between real and simulated samples are shown across all simulated data sets in cyan. As complementary information, the log-transformed F values from PERMANOVA tests are shown in brown. **d)** The absolute generalized fold change[5] and the absolute difference in prevalence across groups is shown for all features in colorectal cancer (CRC) and Crohn's disease (CD). As a comparison, the same values are displayed for two data sets simulated using feature implantation (abundance scaling factor of 2, prevalence shift of 0.1), with implantations either into all features or only low-abundance features. Well-described disease-associated features are highlighted (F = Faecalibacterium, R=Ruminococcus) and selected bacterial taxa and simulated features are shown as percentile plot in **e)**.

Implanted DA features are similar to real-world disease effects

To compare implanted DA features to those observed in real MWAS data in terms of their effect sizes, we focused on two diseases with well-established microbiome alterations, namely colorectal cancer (CRC)[4,5] and Crohn's disease (CD)[2,3]. In two separate meta-analyses (see **Methods**), we calculated the generalized fold change as well as the difference in prevalence between controls and the respective cases for each microbial feature (see **Fig. 1de**). The effect sizes in CRC were generally found to be much lower than in CD, which is consistent with machine learning results in both diseases (mean AUROC for the distinction between cases and controls: 0.92 in CD and 0.81 in CRC, see ref[11]). For instance, the well-described CRC marker Fusobacterium nucleatum exhibits a rather moderate increase in abundance in CRC, but a strongly increased prevalence. This observation, generalizable also to many other established microbial disease biomarkers, motivates the inclusion of the prevalence shift as an additional type of effect size for the proposed implantation framework.

Depending on the type and strength of effect size used to implant DA features, the simulated datasets included effects that closely resemble those observed in the CRC and CD case-control datasets (**Fig. 1ed**). In particular, simulated abundance shifts with a scaling factor between groups <=10 were the most realistic and thus used for subsequent analyses (**SFig. 2**).

Performance evaluation of differential abundance testing methods

To benchmark the performance of various DA testing methods under realistic conditions, DA tests (see **Methods** for a list) were applied to each feature across all simulated datasets including repeated sampling with varying effect sizes. Different sample sizes were created by repeatedly selecting random samples from each simulated group in a balanced manner, and each test was applied to the exact same sets of samples (see **Methods**). In general, we aimed to use the recommended data preprocessing steps for each method, but for some tests (such as the linear model (LM) or the Wilcoxon test), different normalization techniques were also explored (see **SFig. 3**).

The *P* values resulting from each of the included DA methods were adjusted for multiple hypothesis testing with the Benjamini-Hochberg procedure to obtain false discovery rate (FDR) estimates[32]. Additionally, a receiver operating characteristic (ROC) analysis was

carried out to evaluate how accurately the returned *P* values could distinguish between ground truth and background features. If *P* values for all ground truth features are smaller than for any of the background features, the area under the ROC curve (AUROC) will be one; for random *P* values an AUROC of 0.5 is expected.

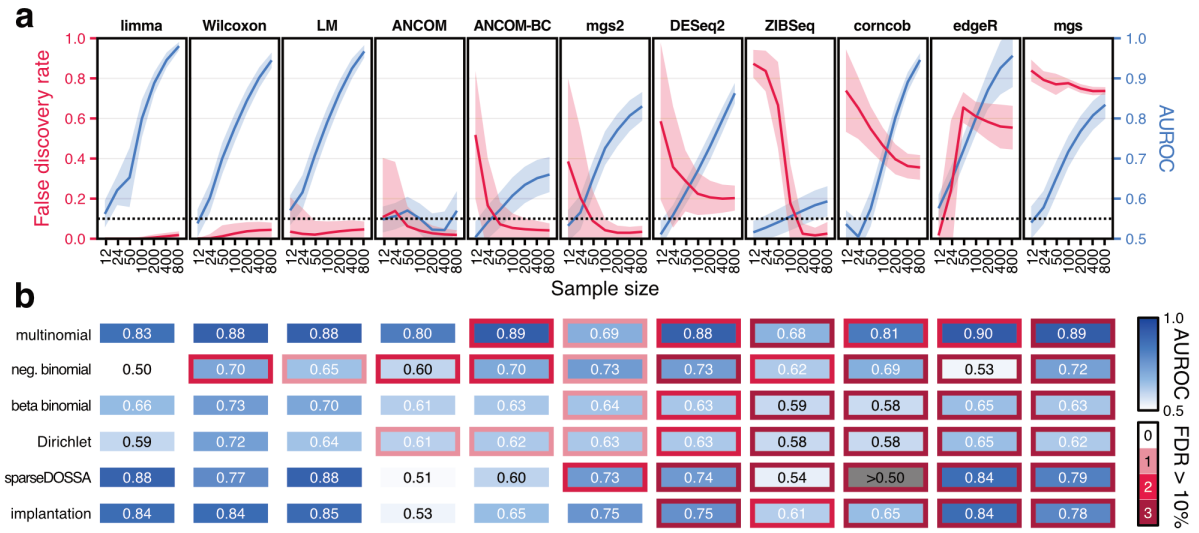


Figure 2: Most differential abundance testing methods fail to control the false discovery rate.
a) The mean false discovery rate across all repetitions of a signal implantation simulation with a single, moderate effect size combination (abundance scaling factor of 2, no prevalence scaling, all features eligible for implantation) is shown as a red line, with the shaded area indicating the standard deviation. Similarly, the blue line and shaded area denote the mean AUROC (and standard deviation) for the detection of implanted signals based on the *P* values (*mgs* - *fitZig* function and *mgs2* - *fitFeatureModel* function from the metagenomeSeq package, LM - linear model, see also **Methods**). **b)** The mean AUROC values across the subset sizes 50, 100, and 200, all repetitions, and all effect sizes are depicted in the heatmap for the different simulation strategies. The cell borders are colored according to the number of subsets in which the mean FDR exceeded 10%.

In this benchmark, the majority of methods failed to consistently control the FDR at the nominal 5% level, with several methods exceeding this value manyfold (displayed for a single representative effect size in **Fig. 2a** and **SFig. 4** for other effect sizes). In the most extreme case, the *fitZig* method from metagenomeSeq (*mgs*), only 20% of features identified as significantly differentially abundant between groups were correctly predicted (FDR of 0.8), regardless of sample size and observed consistently across many effect sizes (see **SFig. 4**). Theoretically, a possible explanation for these high FDR values could be that the methods are not well-calibrated for microbiome data and report universally low *P* values, but are still able to distinguish between ground truth and background features. Such cases could be readily diagnosed using ROC analysis, and theoretically changing the *P* value cutoff for significance could alleviate the problem. However, empirically most methods with

elevated FDRs recorded lower AUROC values when compared to other methods as well, indicating that among these methods do not efficiently enrich true DA features in their results. As a consequence, these DA methods cannot simply be improved by recalibration. These findings were largely independent from the implanted effect size, with larger effect sizes facilitating easier detection and therefore resulting in higher AUROC values on average (see **SFig. 4**).

Distributional assumptions made by the simulation method have a large influence on the performance estimates for DA testing methods and can in turn lead to biased conclusions (see **Fig. 2b** and **SFig. 5**). For example, the benchmarking study by Weiss et al.[24] proposed using the ANCOM method for DA testing. While this method shows seemingly superior performance when multinomial data distributions are assumed (as done in Weiss et al.[24] and reproduced here), this is in stark contrast to benchmarking with the realistic feature implantation setting, where ANCOM did not perform significantly better than random guessing for the identification of ground truth DA features (P=0.34, one-sample t-test for sample sizes of 50, 100, and 200). Similarly, the publication by Hawinkel et al.[25] concluded that almost all methods fail to control the FDR. Being based on the assumption that metagenomic data follow a negative binomial distribution, these results could be faithfully reproduced here (**Fig. 2b**), but are not supported by the results of benchmarking based on more realistic simulations.

Limma, the Wilcoxon test, and the LM were found to be the only methods in the feature implantation benchmark that consistently controlled the FDR, even at smaller sample sizes. For sample sizes over 200, more methods (including ANCOM, ANCOM-BC, *mgs2* or ZIBSeq) exhibit acceptable FDR control. However, none of these methods resulted in AUROC values comparable to limma, LM, or the Wilcoxon test, indicating limited sensitivity of those methods for the detection of true DA features.

Simulating confounding through batch effects

Awareness of confounding being a prevalent issue in MWAS is increasing[33]. This arises when covariates other than the main variable of interest are associated with microbial composition or individual bacterial taxa. When not accounted for, confounding can lead to spurious associations that do not replicate in independent datasets. For example, associations between gut microbiome composition and type 2 diabetes from two different

studies were later identified to be mainly caused by metformin treatment in a subset of type 2 diabetes patients[9].

To study how suitable existing DA methods are for confounder adjustment, we simulated additional confounders in the feature implantation framework. To do so, we mainly relied on technical variation differences between datasets, also called study or batch effects, which are prominent in metagenomic data due to non-standardized experimental protocols[34,35]. To simulate confounding by study effects, we combined the data from two baseline datasets of healthy adults[31,36] and varied the proportion of samples from each data set in the two groups used for implantation of DA features (see **Methods**). In this setup, the degree of confounding by study effects can be modulated by disproportionately sampling one group from one data set and the other group from the other data set. This is opposed to proportional sampling of both groups from both data sets, where there is minimal confounding (see **Fig. 3a**). With increasing simulated confounder strength, study differences will become more aligned with the DA features implanted into the two groups, making their identification increasingly challenging (see **Fig. 3b**).

For this confounder benchmark, we also assessed biological realism by comparison to real metagenomic datasets, as they would be encountered in meta-analyses of CRC and CD. Contrasting the generalized fold changes associated with study differences and those associated with the disease label indicated that moderate or even strong study confounding as simulated in our benchmark does indeed reflect real effects observed in some cross-study comparisons of relevant gastrointestinal diseases (see **SFig. 6**). In most pairwise comparisons between studies, though, the experimental design resulted in roughly equal proportions of cases and controls, thereby mitigating strong study confounding.

Performance evaluation under confounded conditions

To evaluate how DA testing methods perform under confounded conditions, we only retained those methods which were found to control the FDR in the previous benchmark and which could be explicitly adjusted for confounders . We found that without such an adjustment, that is when applied naively, almost all DA tests exhibited an increased FDR, already in the presence of moderate study confounding (median FDR around 20% for Wilcoxon test, LM, and ANCOM-BC; see **Fig. 3c**); their FDR increased further to around 75% in simulations with strong study confounding.

We next adjusted these DA methods by including 'study' as a single covariate in the respective test formula (see **Methods** for details). Interestingly, when confounding was explicitly modeled, we found three out of the four included methods to perform nearly as well as under non-confounded conditions, even in situations with strong confounding (see **Fig. 3c**). One notable exception was limma, which showed a relatively high median FDR (~30%) in the strongly confounded case, even after including the study covariate in the model formula. Overall this result however suggests that measured confounders can be effectively adjusted for in MWAS when explicitly modeled as such.

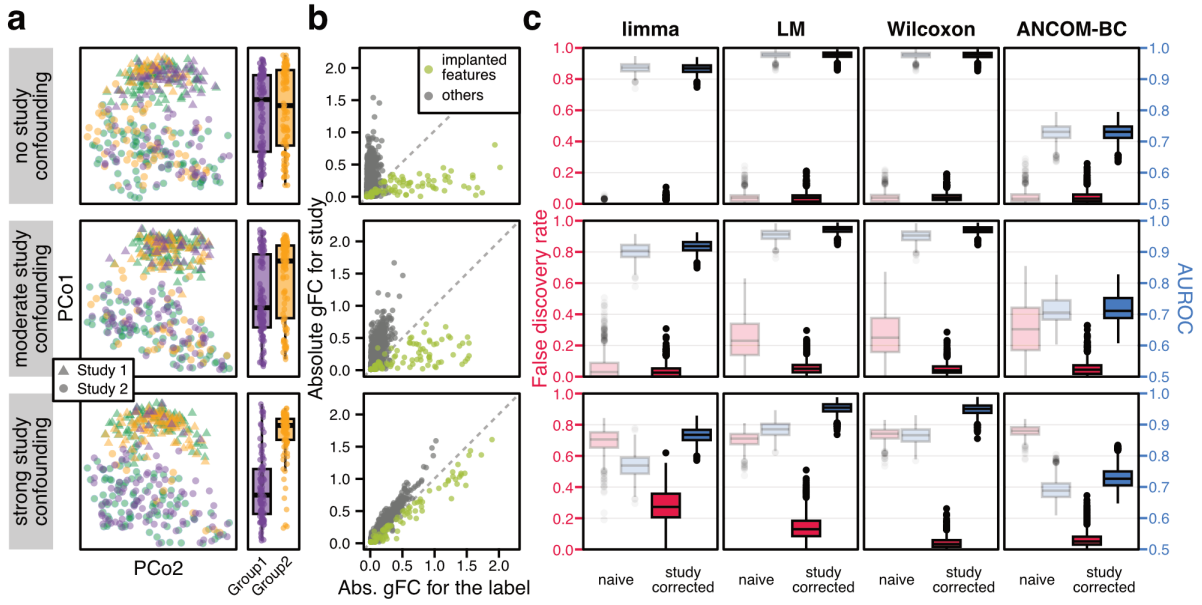


Figure 3: Strong confounding leads to spurious associations which can be alleviated by methods able to model confounders.
a) Principal coordinate projections for simulated data with no, moderate, and strong study confounding (abundance scaling factor of 2, prevalence scaling factor of 0.3, all features eligible for implantation, a single representative repeat shown). The difference between studies is most prominent in the first principal coordinate. On the right, the first principal coordinate values are shown across the two simulated groups. In the top row, samples are proportionally selected from both studies, leading to a setup with minimal confounding. In contrast, the proportion of samples of Study 1 to be selected into group2 increases in the other rows, leading to stronger study confounding. **b)** Generalized fold change (gFC) calculated for the label is contrasted to the gFC calculated for differences between studies across all bacterial taxa for the same repeat as shown in a). **c)** FDR and AUROC across all repeats with the same effect sizes as shown in a) were computed for all included DA methods, for both a naive test and one in which the study covariate was included in the test formula. The boxplots show the results for a sample size of 200.

Discussion

Clinical interest in the microbiome has produced myriad studies which apply differential abundance tests to detect associations with host phenotypes, including many common diseases. Although DA testing is a fundamental statistical task in MWAS, a surprisingly large number of different methods have been employed in various studies[27], necessitating the empirical evaluation of their performance using simulated data in which the ground truth is known. For simulations to generalize, it is key to assess how well the underlying models mimic reality by comparing data they generate to real samples, which crucially was lacking in previous benchmarks[21–26,28]. To address this, we propose a novel implantation framework for the generation of simulated taxonomic profiles based on minimal modifications to real metagenomic data. We empirically verified that our framework, but not previously employed simulation methods based on parametric distributions, retains essential data properties, most importantly feature sparsity and variance. This was also verified by both PERMANOVA (P=0.743 compared to P<0.01 for all other simulations) and machine learning analyses (median AUROC of 0.5 compared to 0.62 for sparseDOSSA and 1 for other simulations). Yet, our framework provides the flexibility to specify effect and sample sizes needed for an extensive evaluation, which was not the case in previous benchmarks built upon real datasets[25,26,28].

To resolve the question of which DA methods are best suited for microbiome data, we performed an unbiased benchmarking study of widely used DA methods using the feature implantation framework. Evaluating each DA test on nearly one million simulated data sets, we found that the majority of methods yielded and excess of false-positives (mean FDR > 20% for 6/11 methods at a data set size of N=100), and this was generally worse for smaller sample sizes (N<100). Notable exceptions were the well-established non-parametric Wilcoxon test (also implemented in the LEfSe[37] and SIAMCAT[11] packages that are tailored to microbiome data), limma, and LMs, all of which were found to properly control the FDR while retaining high sensitivity across a range of sample and effect size. These results strongly suggest that these methods should be preferred over the other tests evaluated here. DA methods borrowed from RNA-seq analysis, with the exception of limma, were among those with the highest FDRs. These methods were originally developed for few replicates with much lower dispersion than what is observed for fecal metagenomes of

different human individuals. Our conclusion that these methods are unsuitable for microbiome data directly contradicts the results of a previous benchmark[22], a discrepancy that can be explained by the use of the multinomial simulation in that study, which strongly underestimated the variance of real microbiome data (see **Fig. 1b**). Surprisingly, most methods developed in recent years with the characteristics of metagenomic data specifically in mind were found to have comparably low power and high false discovery rates (with the exception of the *mgs2*) across the range of dataset sizes most commonly seen in MWAS (see **Fig. 2b**). On a positive note, more DA tests (including both ANCOM versions, ZIBseq and *mgs2*) controlled the FDR at the nominal level when applied to larger samples (N=200 per group). Overall, however, our finding that most DA tests evaluated here would return up to ten times more false positives than expected across sample sizes typically seen in MWAS (N=100 per group) indicates that many of these studies reported a substantial fraction of spurious microbiome-disease associations.

This issue is further exacerbated by confounding factors, for which awareness is growing with the various factors revealed to shape microbiome composition[9,13,15,33,38]. Confounding remains difficult to identify and address post-hoc in most clinical studies. To model confounding in multi-center trials or meta-analyses, we extended our signal implantation framework to take two studies as input and simulated a range of study effects (see **Methods**), which can represent an impediment to combined analysis[5,39]. To assess DA method performance, we selected the subset of methods that performed reasonably well in our first benchmark (mean FDR across effect sizes not exceeding 10% for sample sizes 50, 100, and 200) and allowed inclusion of a covariate to correct for confounding (see **Methods**). To our knowledge, this represents the first benchmark of DA methods under confounded conditions. As expected, application of the unadjusted methods resulted in strongly elevated FDRs under confounded conditions for all methods (median FDR between 23 and 30% under moderate study confounding as compared to 4 to 5% in the absence of confounders), except for limma, which showed lower increases in FDR (1% compared to 3%). Reassuringly, inclusion of the study covariate in the DA models mostly restored unconfounded performance. When explicitly adjusted for the covariate, the blocked Wilcoxon test most tightly controlled the FDR while retaining high power even under strong confounding; however, as it is limited to blocking a single discrete covariate, this test is less

flexible than linear mixed-effect models, which can accommodate multiple covariates and be configured to handle nested or longitudinal study designs. In many areas of biomedicine, mixed-effect models have long been a solution to explicitly deal with study heterogeneity in meta-analyses[4,40].

In our benchmark, we opted to include a simple implementation of linear mixed-effect models, available in the *lmerTest* R package[41]. A statistically analogous implementation is found in MaAsLin2, a microbiome DA analysis pipeline which also includes several count-based linear models[42]. Similarly, metadeconfoundR[38] employs the same types of linear models we used but applies an iterative nested model testing procedure to identify which features are subject to confounding, which is a similar approach as recently employed to find robust microbial biomarkers[43]. Both tools may include multiple covariates in their feature models, a practice whose performance we did not assess here, but which merits further exploration[44].

In our assessment, the unsatisfactory performance from a wide range of DA methods warrants a community effort to develop a more robust methodology, both for testing and assessing their results. Towards this goal, both the signal implantation framework and the benchmarking analysis are designed to be easily extensible and available as open source code (see **Methods**), with the hopes of assisting researchers wishing to develop and validate new DA methods or aiming to establish benchmarks beyond what we have presented here. Ultimately, the consolidation of statistical methodology in the microbiome field might be accelerated by a community-driven benchmarking project similar to DREAM challenges crowdsourcing tasks such as the inference of signaling networks[45] or the critical assessment of metagenome interpretation (CAMI)[46] project. Ideally, such efforts would be neutral and evidence-based[47] to avoid sources of biases that have contributed to the current problem.

Methods

The codebase for the presented results is split into two projects. The first one, an R package called SIMBA (Simulation of Metagenomic data with Biological Accuracy), provides the modular functionality to i) simulate metagenomic data for a benchmarking project, ii) perform reality checks on the simulated data, iii) run differential abundance (DA) testing methods, and finally iv) evaluate the results of the tests. The second project, BAMBI (Benchmarking Analysis of MicroBiome Inference methods), is a collection of scripts that produce the presented analyses, consisting mostly of functions to automate and parallelize the execution of SIMBA functions by making use of the batchtools package[48]. Both projects are available through Gitlab and will enable other researchers to explore a similar benchmarking setting for other baseline datasets, other biomes, and additional DA testing methods.

Data preprocessing

The dataset from Zeevi et al.[31] was used as a baseline for the simulations in most cases. Additionally, the TwinsUK dataset[36] was included in some of the study-confounded simulations as well. Raw data were downloaded from ENA (PRJEB11532 for Zeevi and ERP010708 and TwinsUK) and profiled using the mOTUs2 profiler, version 2.5[49]. The resulting taxonomic profiles were filtered within SIMBA for prevalence (at least 5% in the complete dataset) and abundance (relative abundance of at least 1e-04). In the case of repeated samples per patient, SIMBA selects only the first time point for each patient.

Parametric methods for the simulation of metagenomic data

To simulate metagenomic data on the basis of parametric methods, the implementations employed in previous benchmarking efforts were adapted into SIMBA. Data were simulated under multinomial distributions using code from both McMurdie and Holmes[22] and Weiss et al.[24], since the functions to include differentially abundant features differed slightly between the two benchmarks. If not indicated otherwise, results for multinomial simulations were based on the implementation from Weiss et al., since the effect sizes were closer to real effects (see **SFig. 2**). The publication from Hawinkel et al.[25] included simulations based on the negative binomial, the beta binomial, and the Dirichlet distribution, which were likewise included in SIMBA. As in the original publication[25], correlations across bacterial taxa were estimated using SPIEC-EASI[19], since the correlation structure was needed for the beta binomial and could optionally be considered for the negative binomial simulations. Lastly, to simulate data as described in Ma et al.[30], SIMBA relies on the dedicated functions in the sparseDOSSA R package.

For each of the parametric simulation methods, the required parameters were estimated on the filtered Zeevi dataset. A dataset of equal size was simulated to include two different groups into which differentially abundant features were added as described in the respective original publications. For the multinomial simulations from McMurdie and Holmes as well as for the sparseDOSSA approach, features were scaled in abundance after the simulation was completed. In the case of the other simulation methods, the underlying parameters were adjusted with a scaling factor before the simulation. A range of effect sizes (abundances scaled by multipliers of 1, 1.25, 1.5, 2, 5, 10, and 20) was explored and for each effect size, a total of 20 repetitions were simulated per

simulation method. At an abundance scaling factor of 1, no effects were introduced into the data and therefore those repeats can serve as internal negative controls.

Implantation of differentially abundant features into real data

To create benchmarking datasets through minimally adjusting the original data, differentially abundant features were implanted into the Zeevi dataset as a baseline. In each repetition, the original samples were randomly split into two groups, which served as the positive and negative groups. Differential abundance effects were implanted into a set of randomly selected features both via scaling abundances (same effect sizes as the parametric simulations) as well as by shifting prevalences (0.0, 0.1, 0.2, and 0.3).

For the abundance scaling, the count values in one group were multiplied with the scaling factor to increase the abundance. The prevalence shifts were implemented by identifying non-zero counts in one group and exchanging a specific percentage of those with occurrences of zero abundances in the other group (if possible), thereby creating a difference in prevalence across the groups. The feature implantation alternated between the two groups in order to not introduce a systematic difference in total count number across groups (inspired by the considerations in Weiss et al.[24]). For each combination of effect sizes (abundance and prevalence scaling), 100 repetitions were simulated.

In each repetition, 10% of features were randomly selected for signal implantation. The set of features eligible to be selected, however, could be varied (see **SFig. 2**): *all* - all taxa were equally likely to be selected to carry a signal, *low* - only low abundance features (the 75th percentile across all samples not exceeding 0), *high* - only high abundance features (the median abundance across all samples higher than 0), *abundance* - the probability of a taxon to be selected is proportional to the mean abundance across all samples, and *inverse_abundance* - the probability of a taxon to be selected is inversely proportional to the mean abundance. Since the effect sizes from other schemes yielded unrealistic effect sizes, the downstream analyses were only carried out for the *all* and *low* implantations.

Reality assessment for simulated data

To determine how well a simulated metagenomic dataset approximated real data, several metrics are calculated by SIMBA. For each repetition of each simulation, sample sparsity and feature variance were recorded together with differences in prevalence and the generalized fold change[5] between groups. Additionally, the separation between original and simulated samples in principal coordinate space was evaluated using PERMANOVA as implemented in the vegan package[50]. As a complementary approach, a machine learning model was trained to classify real and simulated samples using the SIAMCAT R package[11] and the AUROC of the cross-validated model was recorded.

Included DA testing methods and normalization procedures

To evaluate the performance of various DA testing methods, the R implementation of each method was incorporated into SIMBA using the recommended preprocessing, if applicable. The following methods were included in the benchmark (usually available through an R package of the same name): the Wilcoxon test and linear models (available within the base R distribution), limma[51], edgeR[52], DESeq2[53], metagenomeSeq[21], ZIBSeq[54], corncob[55], ANCOM[56], and ANCOM-BC[57]. For metagenomeSeq, two different models can be fitted

within the same R package, which are included here as *mgs* (using the *fitZig* function) and *mgs2* (using the *fitFeatureModel* function), analogously to Weiss et al.[24]. For ANCOM, no dedicated R package is available from the original publication and the standard implementation is prohibitively slow, thus the implementation available through Lin et al.[58] was used (see **SFig. 7**).

Most of these methods work on the raw count data and therefore no normalization was needed. For the Wilcoxon test, the LM, and limma, different sets of normalization methods were explored, namely *pass* (no normalization), *clr* (centered log ratio transform), *rclr* (robust centered log ratio transform), *TSS* (total sum scaling), *TSS.log* (total sum scaling, followed by log10 transformation of the data), and *TSS.arcsin* (total sum scaling, followed by the arcsine square root transformation).

Benchmarking of DA testing methods at different sample sizes

To simulate different cohort (sample) sizes, SIMBA randomly selected N samples out of the two groups equally for each combination of effect size and each repetition. These samples were saved via their indices such that each method was applied to the exact same data. Seven different sample sizes were explored (12, 24, 50, 100, 200, 400, and 800) and 50 sets of test indices were created for each. For the evaluation of a single DA method, a total of 980,000 unique configurations were generated and used as input (7 abundance shifts x 4 prevalence shifts x 100 simulation repeats x 7 sample sizes x 50 repeats).

Each method was applied to each bacterial taxon in succession using the previously indexed samples. The I values across all taxa were recorded and adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure[32]. Since ANCOM does not return *P* values, its primary outputs (*W* values) were converted to be comparable to *P* values by transforming them to range between 0 and 1. The recommended decision threshold for significance in ANCOM is equal to 0.7 x number of tested taxa. Therefore, the *W* values above this decision threshold were transformed into 'significant' *P* values (lower than 0.05), whereas all other *W* values were transformed to range between 0.05 and 1 in the *P* value space. The ranking of the *W* values was conserved in this transformation.

To evaluate the performance of each method, SIMBA checked the *P* values from each testing scenario for how well bacterial taxa with differential abundance were detected. An AUROC was calculated with the *P* values as a predictor and the false discovery rate (FDR) was recorded with 0.05 serving as the decision threshold.

Confounder implantation by mixing data from different studies

To simulate a setting with a realistic confounding present, the taxonomic profiles from two different studies were combined (Zeevi and TwinsUK studies). Study effects are known to affect a large number of bacterial taxa and a systematic difference between the studies was indeed apparent in a PCoA (see **Fig. 3a**). In selecting samples for the positive and negative groups in the different simulation repetitions, the probability of a sample to be selected for the positive group was then contingent on the study affiliation. When this probability was biased towards one of the studies, a systematic shift between the groups could be introduced into most taxa. At a bias of 0.5, there is no difference between the probability for a sample of one study versus the other study to be selected for the positive group, and at a bias of 1, only the samples from one of the studies will be selected for the positive group. Implantation of differentially abundant features was carried out as described above. We

created three different benchmarking simulations with a bias of 0.5 (no study confounding), 0.7 (moderate study confounding), and 0.9 (strong study confounding).

Confounder-aware DA testing

For the confounded benchmarking, only tests with a reasonable performance in the not-confounded setting were run through SIMBA, namely the Wilcoxon test, the LM, limma, and ANCOM-BC. All tests could also be adjusted by the confounder covariate, usually by including the covariate into the test formula. For the Wilcoxon test, confounder-aware testing was performed using the blocked Wilcoxon test implemented in the coin package[59] and for the LM, the confounder covariate was included as a random effect in the formula. The significance of the original study variable was then tested by fitting the model using the *lmer* function within the *lmerTest* package[41]. The evaluation procedure was otherwise unchanged compared to the benchmarking without confounding.

Effect size assessment in real case-control datasets for colorectal cancer and inflammatory bowel disease

To compare simulated data to real case-control microbiome studies, we collected datasets for two diseases with a well-described microbiome signal. For colorectal cancer (CRC), we included the data from five studies[5,60–63] across three continents, which were the basis for an earlier meta-analysis that identified consistent and predictive microbial biomarkers for CRC[5]. For Crohn's disease (CD), we similarly included five case-control studies[3,64–67] that had been analyzed previously[11]. For CD, the data were restricted to the first measurement for each individual, whenever applicable. The data from all studies were taxonomically profiled via mOTUs2 (version 2.5, ref[49]) and features were filtered for at least 5% prevalence in at least three of the studies. Differences in prevalence across groups and the generalized fold change were calculated for each microbial feature as previously described[5] and the significance of enrichment was calculated using the blocked Wilcoxon test from the coin package in R[59].

Author contributions

G.Z. and S.K.F. conceived the study and supervised the work. J.W. and M.E. implemented the software and performed the statistical analyses with guidance from G.Z. and S.K.F.. J.W., M.E., and G.Z. designed the figures with input from S.K.F.. J.W., M.E., and G.Z. wrote the manuscript with contributions from S.K.F.. All authors discussed and approved the final manuscript.

Code availability

The software package to simulate metagenomic data (SIMBA) is available on Gitlab: https://git.embl.de/jawirbel/SIMBA. Similarly, the repository containing the scripts to run a benchmark (BAMBI) is also available on Gitlab: https://git.embl.de/jawirbel/BAMBI.

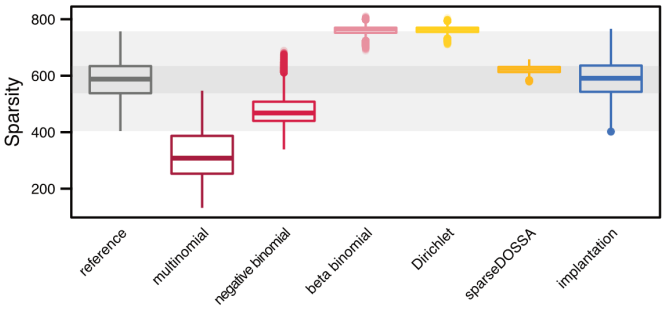
References

1. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16**, 73 (2015).
2. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
3. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293–305 (2019).
4. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
5. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
6. Li, J. *et al.* Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome* **5**, 14 (2017).
7. Wu, H. *et al.* Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.* **23**, 850–858 (2017).
8. Rubinstein, M. R. *et al.* Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe* **14**, 195–206 (2013).
9. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
10. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).
11. Wirbel, J. *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 93 (2021).
12. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
13. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
14. Vieira-Silva, S. *et al.* Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**, 310–315 (2020).
15. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
16. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
17. Bartolomaeus, T. U. P. *et al.* Quantifying technical confounders in microbiome studies. *Cardiovasc. Res.* **117**, 863–875 (2021).
18. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, 2224 (2017).
19. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
20. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, (2012).
21. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
22. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).
23. Thorsen, J. *et al.* Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* **4**, 62 (2016).
24. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).
25. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20**, 210–221 (2019).
26. Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol.* **21**, 191 (2020).
27. Kleine Bardenhorst, S., Berger, T., Klawonn, F. & Vital, M. Data Analysis Strategies for Microbiome Studies in Human Populations—a Systematic Review of Current Practice. *Msystems* (2021).
28. Nearing, J. T., Douglas, G. M., Hayes, M. G. & MacDonald, J. Microbiome differential abundance methods produce disturbingly different results across 38 datasets. *bioRxiv* (2021).
29. Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R. & Boulesteix, A.-L. On the optimistic performance

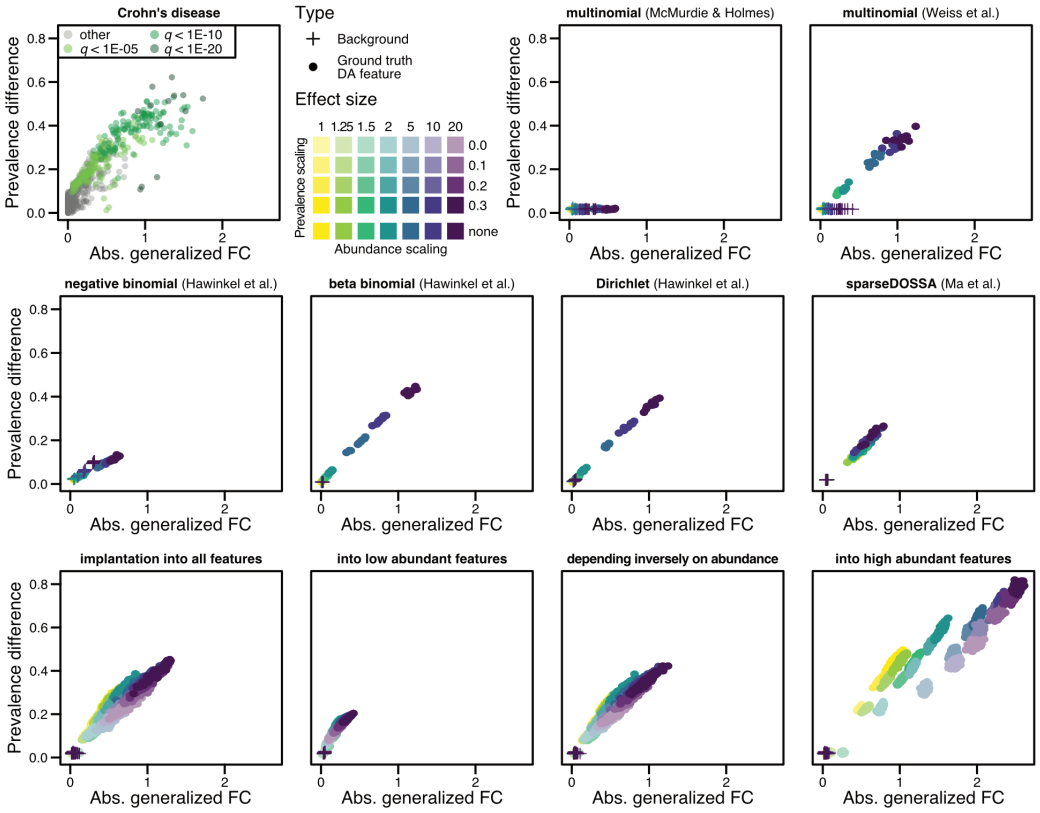
evaluation of newly introduced bioinformatic methods. *Genome Biol.* **22**, 152 (2021).
30. Ma, S., Ren, B., Mallick, H., Moon, Y. S. & Schwager, E. A Statistical Model for Describing and Simulating Microbial Community Profiles. *bioRxiv* (2021).
31. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
32. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
33. Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **172**, 1198–1215 (2018).
34. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
35. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nature Biotechnology* **35**, 1077–1086 (2017).
36. Xie, H. *et al.* Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Syst* **3**, 572–584.e3 (2016).
37. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
38. Forslund, S. K. *et al.* Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature* (2021).
39. Gibbons, S. M., Duvallet, C. & Alm, E. J. Correcting for batch effects in case-control microbiome studies. *PLoS Comput. Biol.* **14**, e1006102 (2018).
40. Stram, D. O. Meta-analysis of published data using a linear mixed-effects model. *Biometrics* **52**, 536–544 (1996).
41. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, Articles* **82**, 1–26 (2017).
42. Mallick, H., Rahnavard, A., McIver, L. J., Ma, S. & Zhang, Y. Multivariable association discovery in population-scale meta-omics studies. *Biorxiv* (2021).
43. Tierney, B. T., Tan, Y., Kostic, A. D. & Patel, C. J. Gene-level metagenomic architectures across diseases yield high-resolution microbiome diagnostic indicators. *Nat. Commun.* **12**, 2907 (2021).
44. Westfall, J. & Yarkoni, T. Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLoS One* **11**, e0152719 (2016).
45. Prill, R. J., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K. & Stolovitzky, G. Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.* **4**, mr7 (2011).
46. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
47. Boulesteix, A.-L., Wilson, R. & Hapfelmeier, A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Med. Res. Methodol.* **17**, 138 (2017).
48. Lang, M., Bischl, B. & Surmann, D. batchtools: Tools for R to work on batch systems. *J. Open Source Softw.* **2**, 135 (2017).
49. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
50. Oksanen, J. *et al.* Package 'vegan'. *Community ecology package, version* **2**, 1–295 (2013).
51. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
52. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
54. Peng, X., Li, G. & Liu, Z. Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. *J. Comput. Biol.* **23**, 102–110 (2016).
55. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Stat.* **14**, 94–115 (2020).
56. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
57. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 3514 (2020).
58. Lin, F. H. HuangLin/ANCOM: third release of ANCOM. *Zenodo10* **5281**, (2019).
59. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. A Lego System for Conditional Inference. *Am. Stat.* **60**, 257–263 (2006).
60. Vogtmann, E. *et al.* Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome

Shotgun Sequencing. *PLoS One* **11**, e0155362 (2016).
61. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
62. 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).
63. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
64. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
65. He, Q. *et al.* Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* **6**, 1–11 (2017).
66. Lewis, J. D. *et al.* Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host Microbe* **18**, 489–500 (2015).
67. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

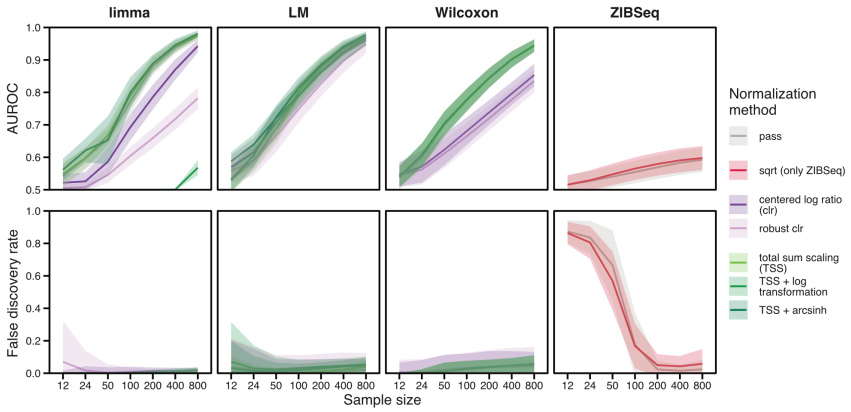
Supplementary Figures



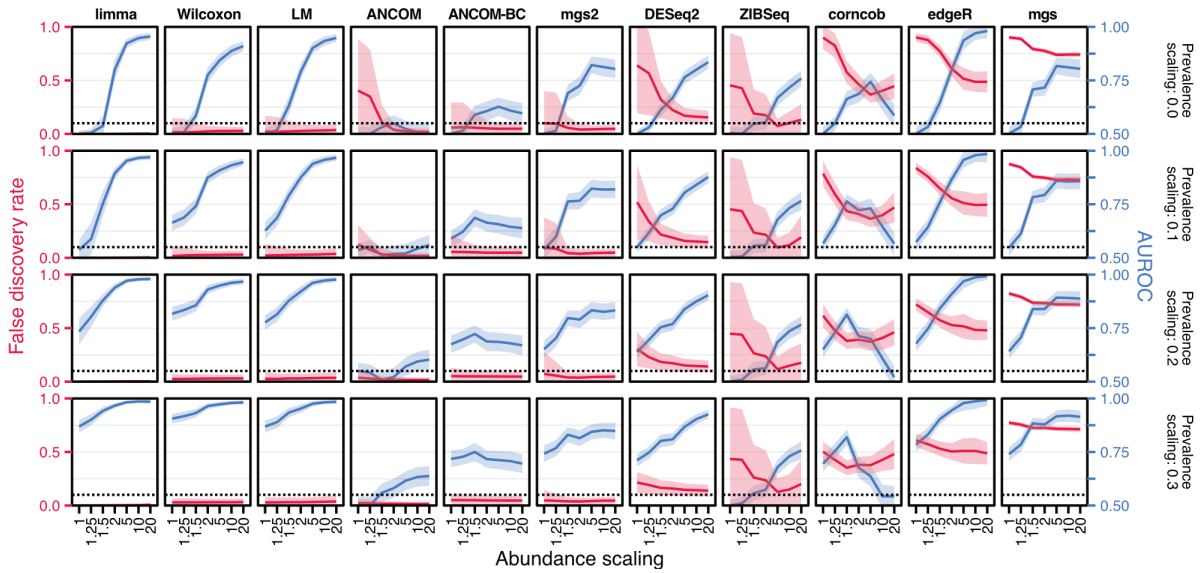
SFig. 1: Sample sparsity is preserved in signal implantation, but not parametric simulations
Sample sparsity (measured as the number of zero entries per sample) was recorded for all simulated samples across all repetitions and effect sizes. Reference indicates the real sample sparsity observed in the baseline dataset from Zeevi (see **Methods**), with the shaded grey area denoting the interquartile range (darker area) and 1.5*interquartile range (lighter area) from the reference (analogous to the boxplot definition). For multinomial simulations, the implementation from Weiss et al. was used.



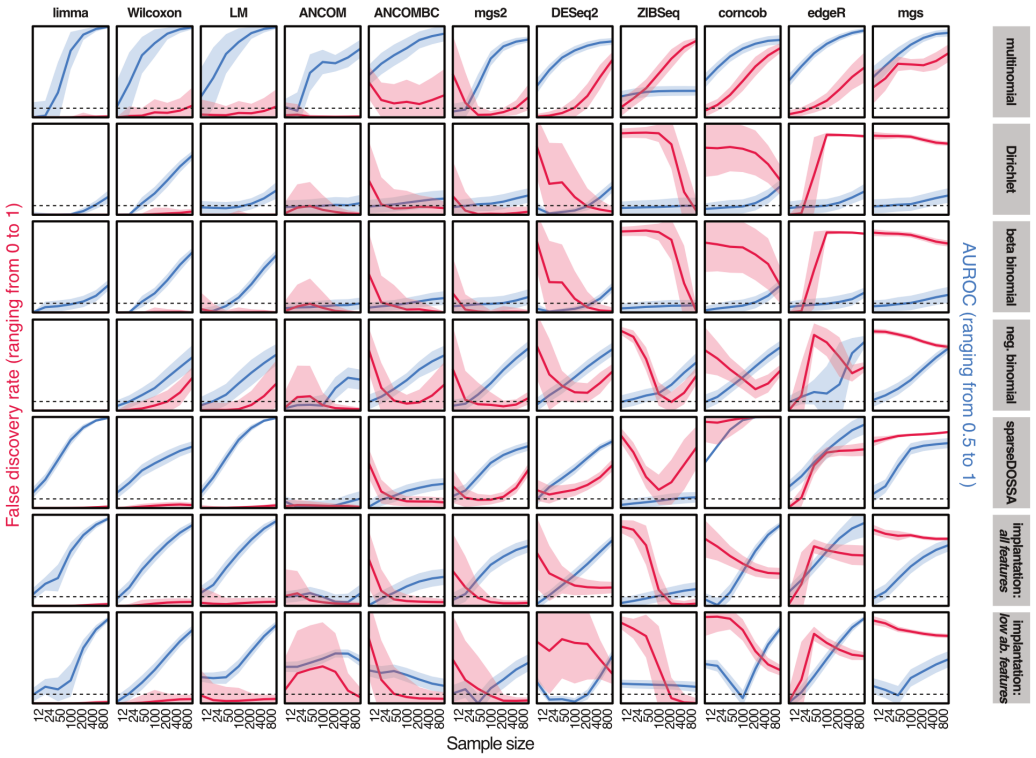
SFig. 2: Implanted effect sizes vary across different simulation scheme and eligible feature sets
The absolute generalized fold change (gFC) and the absolute prevalence difference between groups was calculated for all features across all repetitions in every simulation scheme. For each repetition, the mean gFC and prevalence shift values were calculated for both background and ground truth differentially abundant (DA) features. As a reference point, the real gFC and prevalence shift values observed across all features in the Crohn's disease meta-analysis (see **Methods**) are shown in the top left panel. In the bottom row, mean gFC and prevalence shift values are shown for different signal implantation simulations that vary in which feature set was eligible for implantation. When DA features were implanted into high abundant features, the resulting (mean) effect sizes were too high and unrealistic.



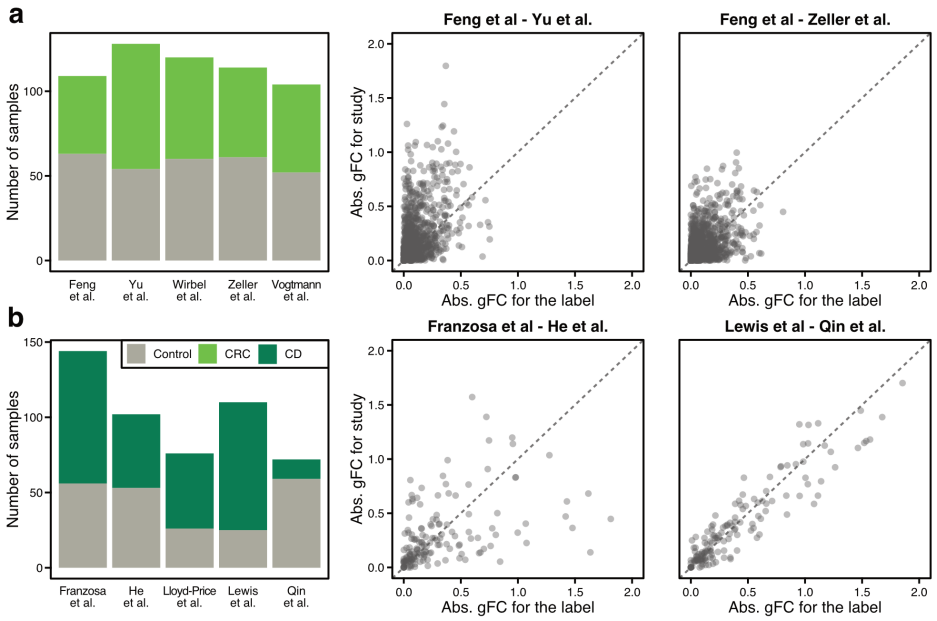
SFig. 3: The choice of normalization method has in general little influence on the resulting performance
AUROC for the detection of ground truth DA features and FDR are shown for a single effect size in the signal implantation setting (same setting as **Fig. 2**, abundance scaling factor of 2, no prevalence scaling, all features eligible for implantation) with varying normalization methods (see **Methods**). The shaded area indicates the standard deviation across repetitions. For ZIBSeq, only no normalization (*pass*) and the *sqrt* method are implemented. In general, the choice of normalization method has little effect on the resulting performance, with two exceptions: Both the *clr* or *rclr* method lower the AUROC for detection of the ground truth DA features in combination with the Wilcoxon test and limma. Additionally, the *TSS.log* method results in extremely low AUROC values for limma only.



SFig. 4: Performance of differential abundance testing methods under various effect sizes
AUROC for the detection of ground truth DA features and FDR are shown across all included methods for varying effect sizes of the same signal implantation benchmark (all features eligible for implantation). The shaded area indicates the standard deviation across repetitions. All values were recorded for a sample size of 100. With higher effect sizes (both prevalence shift and abundance scaling), the AUROC for the detection of ground truth DA features generally increases. In some methods, such as edgeR, the AUROC nears a value of 1 at extreme effect sizes, even though the FDR remains high, which indicates that the *P* values from edgeR can distinguish between background and ground truth DA features but are poorly calibrated (see also Main text).

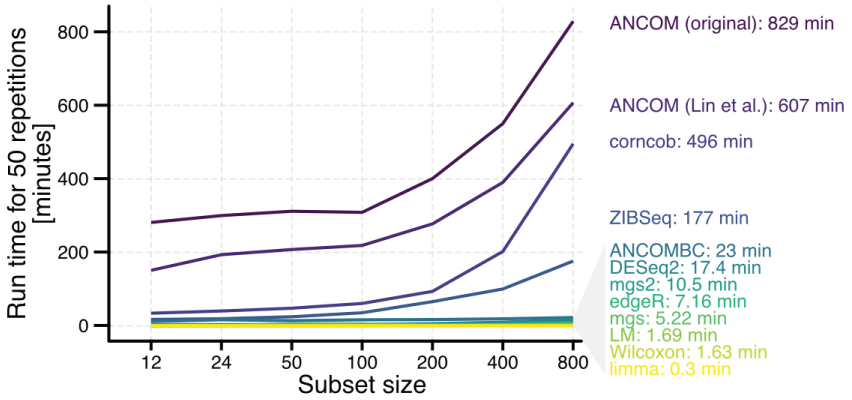


SFig. 5: Performance of differential abundance testing methods across simulation frameworks
AUROC for the detection of ground truth DA features and FDR are shown for a single effect size (same setting as **Fig. 2**, abundance scaling factor of 2, no prevalence scaling, all features eligible for implantation) from different simulation methods. The shaded area indicates the standard deviation across repetitions. For multinomial simulations, the implementation from Weiss et al. was used.



SFig. 6: Examples of study confounding observed in real data mirror effects from the confounded signal implantation benchmark
For colorectal cancer (CRC, panel **a**) and Crohn's disease (CD, panel **b**), the number of samples in each group (control and respective case) is shown across studies as a bar plot on the left. For CRC, study design led to generally balanced comparisons across studies, limiting the risk for strong study confounding. For CD, however, some comparisons are very unbalanced and can therefore introduce strong study confounding (for example, in

the comparison between Lewis et al. and Qin et al.). On the right, generalized fold change associated with the label and with study differences are shown for selected comparisons between two studies analogously to **Fig. 3b**. For the comparisons in CRC, the disease effect is generally not as strong as in CD (also see **Fig. 1**). For some study comparisons in CD, study and disease effects are almost perfectly aligned (Lewis et al. and Qin et al.), mirroring the strong study confounding situation in the signal implantation benchmark (see **Fig. 3**).



SFig. 7: Comparison of runtime across differential abundance testing methods
Runtime was recorded on the same machine for 50 repetition of different subset sizes from a single repetition in the same signal implantation benchmark (abundance scaling of 2, prevalence shift of 0.1, all features eligible for implantation). Methods are annotated with the time needed to run the subset size of 800 samples. The original ANCOM implementation was obtained from the website of the first author of the ANCOM manuscript at https://sites.google.com/site/siddharthamandal1985/research.