nature medicine

ARTICLES

https://doi.org/10.1038/s41591-019-0406-6

# Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer

Jakob Wirbel[1,31], Paul Theodor Pyl[2,3,31], Ece Kartal[1,4], Konrad Zych[1], Alireza Kashani[2], Alessio Milanese[1], Jonas S. Fleck[1], Anita Y. Voigt[1,5], Albert Palleja[2], Ruby Ponnudurai[1], Shinichi Sunagawa[1,6], Luis Pedro Coelho[1,30], Petra Schrotz-King[7], Emily Vogtmann[8], Nina Habermann[9], Emma Niméus[3,10], Andrew M. Thomas[11,12], Paolo Manghi[11], Sara Gandini[13], Davide Serrano[13], Sayaka Mizutani[14,15], Hirotsugu Shiroma[14], Satoshi Shiba[16], Tatsuhiro Shibata[16,17], Shinichi Yachida[16,18], Takuji Yamada[14,19], Levi Waldron[20,21], Alessio Naccarati[22,23], Nicola Segata[11], Rashmi Sinha[8], Cornelia M. Ulrich[24], Hermann Brenner[7,25,26], Manimozhiyan Arumugam[2,27,32*], Peer Bork[1,4,28,29,32*] and Georg Zeller[1,32*]

**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) $< 1 \times 10^{-5}$). 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.**

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

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medicine, University of Copenhagen, Copenhagen, Denmark. [3]Division of Surgery, Oncology and Pathology, Department of Clinical Sciences Lund, Faculty of Medicine, Lund University, Lund, Sweden. [4]Molecular Medicine Partnership Unit, Heidelberg, Germany. [5]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [6]Department of Biology, ETH Zürich, Zürich, Switzerland. [7]Division of Preventive Oncology, National Center for Tumor Diseases and German Cancer Research Center, Heidelberg, Germany. [8]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [9]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [10]Division of Surgery, Department of Clinical Sciences Lund, Faculty of Medicine, Skane University Hospital, Lund, Sweden. [11]Department CIBIO, University of Trento, Trento, Italy. [12]Biochemistry Department, Chemistry Institute, University of São Paulo, São Paulo, Brazil. [13]IEO, European Institute of Oncology IRCCS, Milan, Italy. [14]School of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan. [15]Research Fellow of Japan Society for the Promotion of Science, Tokyo, Japan. [16]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. [17]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [18]Department of Cancer Genome Informatics, Graduate School of Medicine/Faculty of Medicine, Osaka University, Osaka, Japan. [19]PRESTO, Japan Science and Technology Agency, Saitama, Japan. [20]Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA. [21]Institute for Implementation Science in Population Health, City University of New York, New York, NY, USA. [22]Italian Institute for Genomic Medicine, Turin, Italy. [23]Department of Molecular Biology of Cancer, Institute of Experimental Medicine, Prague, Czech Republic. [24]Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, UT, USA. [25]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany. [26]German Cancer Consortium, German Cancer Research Center, Heidelberg, Germany. [27]Faculty of Healthy Sciences, University of Southern Denmark, Odense, Denmark. [28]Max Delbrück Centre for Molecular Medicine, Berlin, Germany. [29]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany. [30]Present address: Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China. [31]These authors contributed equally: Jakob Wirbel, Paul Theodor Pyl. [32]These authors jointly supervised this work: Manimozhiyan Arumugam, Peer Bork, Georg Zeller.
*e-mail: arumugam@sund.ku.dk; bork@embl.de; zeller@embl.de

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 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 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 1 | Fecal metagenomic studies of CRC included in this meta-analysis**

| Country code | Reference | No. of cases | No. of controls |
|---|---|---|---|
| France | Zeller et al.[8] | 53 | 61 |
| Austria | Feng et al.[9] | 46 | 63 |
| China | Yu et al.[11] | 74 | 54 |
| United States | Vogtmann et al.[10] | 52 | 52 |
| Germany | The current study | 60 | 60 |
| **External validation cohorts** | | | |
| Italy 1 | Thomas et al.[27] | 29 | 24 |
| Italy 2 | Thomas et al.[27] | 32 | 28 |
| Japan | Courtesy of T. Yamada et al. | 40 | 40 |

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 Zeller et al.[8].

**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[24] (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[25] 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 $< 1 \times 10^{-5}$; Fig. 1a) for further analysis. The latter included members of several genera previously associated

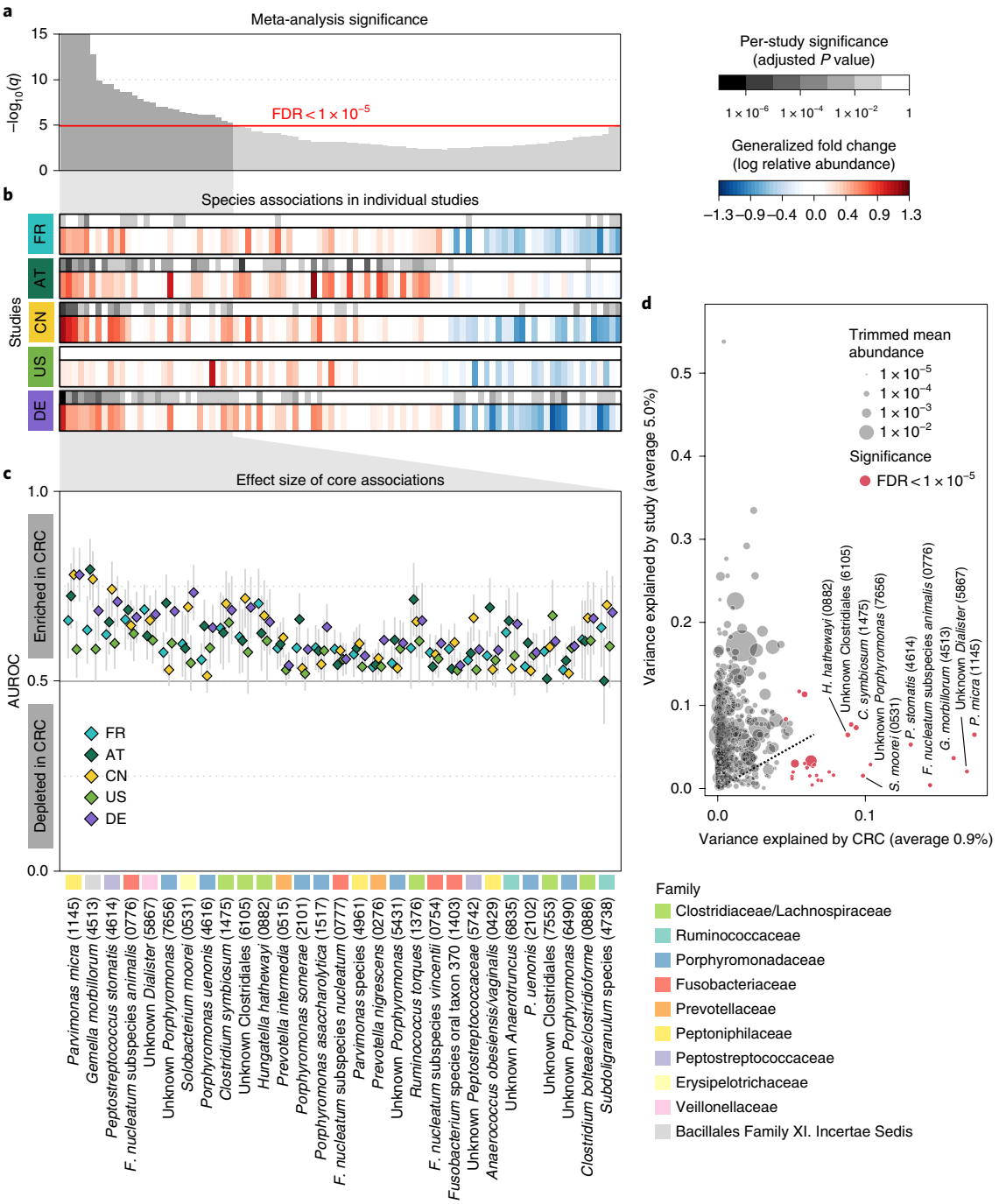



**Fig. 1 | Despite study differences, meta-analysis identifies a core set of gut microbes strongly associated with CRC. a**, The meta-analysis significance of gut microbial species derived from blocked Wilcoxon tests ($n = 574$ independent observations) is given by the bar height (FDR = 0.005). **b**, 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 = $1 \times 10^{-5}$), 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.

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; Milanese et al.[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).

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.

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[26], 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 eliminated 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 $0.07 \pm 0.12$ 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 Thomas et al.[27]), 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 $\geq 0.83$ (Fig. 3b). This corresponds to a LOSO accuracy increase of $0.076 \pm 0.03$ compared to study-to-study transfer. These results suggest that one can expect a CRC detection accuracy $\geq 0.8$ (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.

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

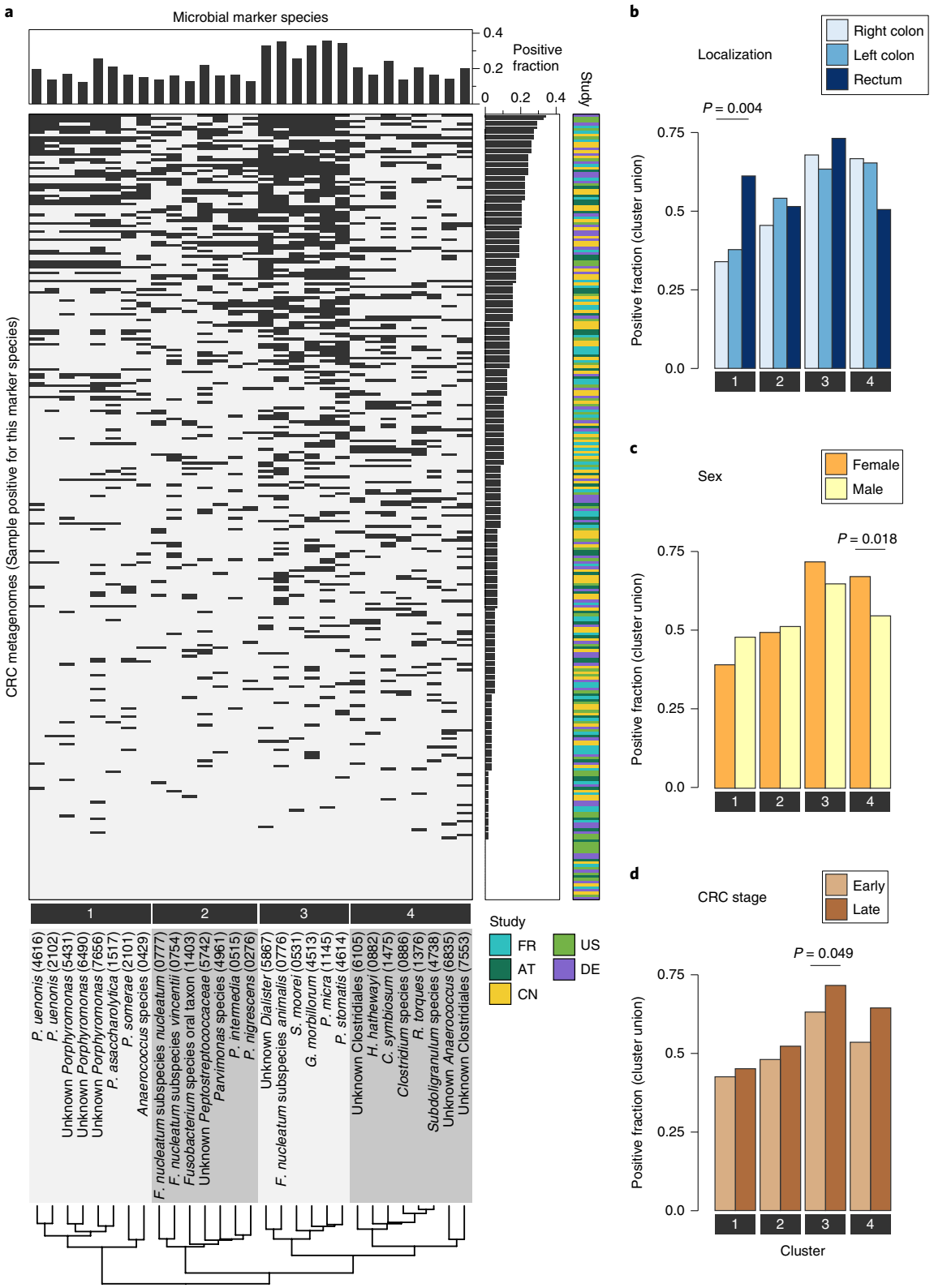



**Fig. 2 | Co-occurrence analysis of CRC-associated gut microbial species reveals four clusters preferentially linked to specific patient subgroups. a**, For all CRC patients ($n = 285$ independent samples), the heatmap shows whether the respective sample is positive for each of the core set of microbial 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.

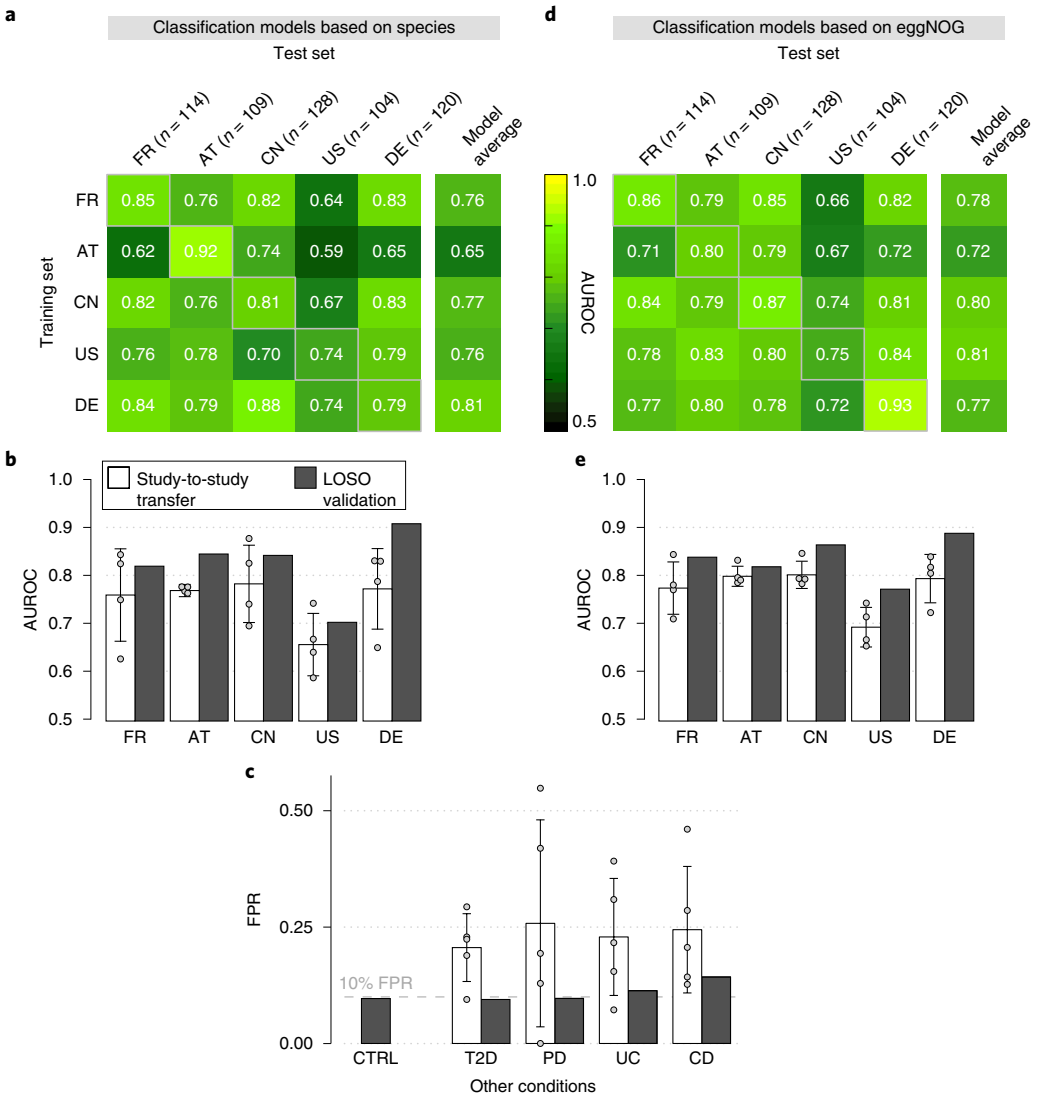



**Fig. 3 | Both taxonomic and functional metagenomic classification models generalize across studies, in particular when trained on data from multiple studies. a**–**e**, CRC classification accuracy resulting from cross-validation within each study (gray boxes along the diagonal) and study-to-study model 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.

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 $\geq 0.96$ 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 Vieira et al.[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].

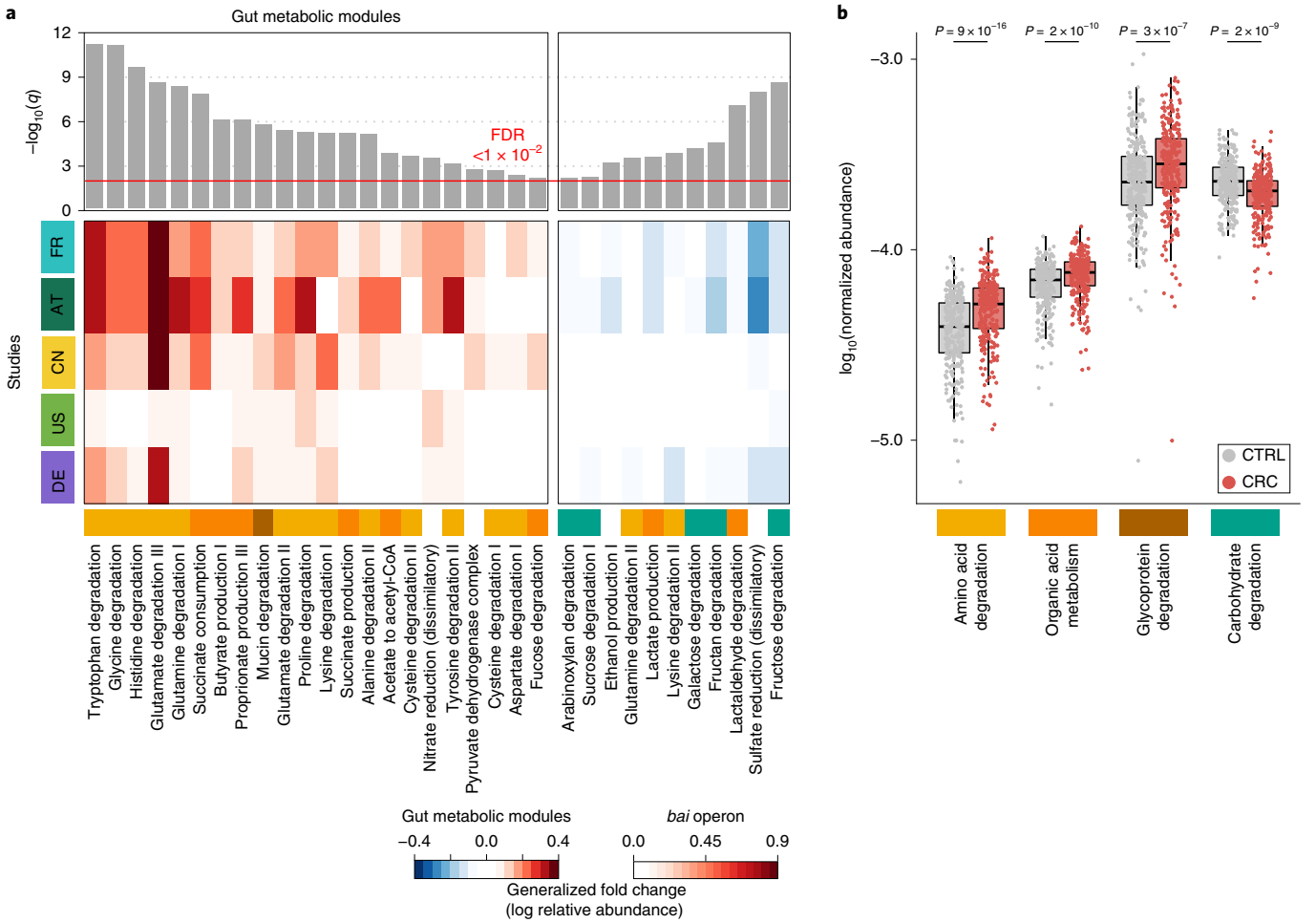
The large metagenomic data set analyzed in this study allowed us to quantify the prevalence of the gut microbial virulence and toxicity mechanisms thought to play a role in colorectal carcinogenesis. Prominent examples include the *Fusobacterium nucleatum* adhesion protein A (encoded by the *fadA* gene), 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 ($P = 5.3 \times 10^{-10}$ and $4.1 \times 10^{-4}$, 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 ($P = 1.6 \times 10^{-9}$) 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[47] 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. 4 | 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. **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.

## 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 Thomas et al.[27]). 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 Zeller et al.[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).



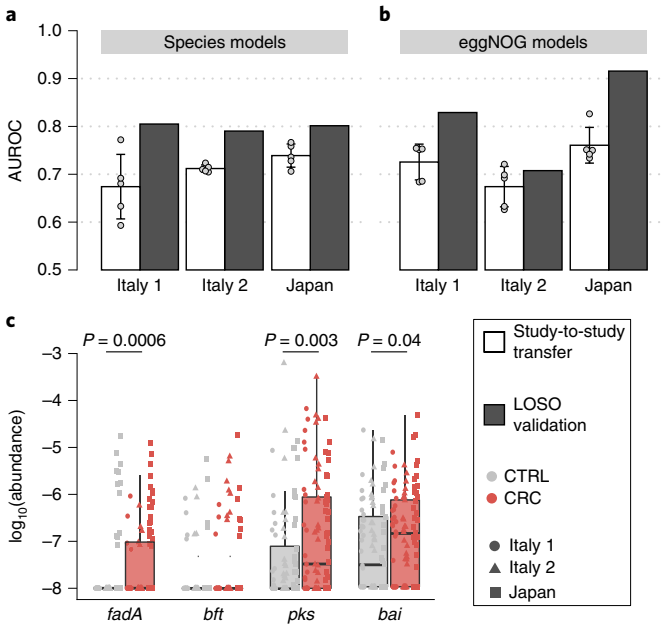



**Fig. 5 | Meta-analysis results are validated in three independent study populations. a**,**b**, CRC classification accuracy for independent data sets, two from Italy and one from Japan (see Table 1 and Supplementary Table 2), is indicated by the bar height for single-study (white) and LOSO (gray) models 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.

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 Thomas et al.[27] 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 Thomas et al.[27], 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.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0406-6.

Received: 30 July 2018; Accepted: 20 February 2019;
Published online: 1 April 2019

## 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, M. 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. Hothorn, T. et al. A Lego system for conditional inference. *Am. Stat.* **60**, 257–263 (2006).
25. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
26. 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).
27. 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.* https://doi.org/10.1038/s41591-019-0405-7 (2019).
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–D293 (2016).
29. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (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. et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol.* **1**, 16088 (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. 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.* **9**, 288 (2015).
38. *Diet, Nutrition, Physical Activity and Cancer: a Global Perspective. A Summary of the Third Expert Report* (World Cancer Research Fund, 2018).
39. Dutilh, B. E., Backus, L., van Hijum, S. A. & 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 7α-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. D. 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**, S190–S201 (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. et al. Integrative analysis of exogenous, endogenous, tumour and immune factors for precision medicine. *Gut* **67**, 1168–1180 (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**, e02248-18 (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).


## Acknowledgements

We are thankful to members of the Zeller, Bork, and Arumugam groups for inspiring discussions. Additionally, we thank Y. P. Yuan and the EMBL Information Technology Core Facility for support with high-performance computing, and the EMBL Genomics Core Facility for their sequencing support. We are also grateful for the advice provided by B. Klaus, EMBL Centre for Statistical Data Analysis. We acknowledge funding from EMBL, the German Cancer Research Center, the Huntsman Cancer Foundation, the Intramural Research Program of the National Cancer Institute, ETH Zürich, and the following external sources: the European Research Council (CancerBiome grant no. ERC-2010-AdG_20100317 to P.B., Microbios grant no. ERC-AdG-669830 to P.B., and Meta-PG grant no. ERC-2016-STG-716575 to N.S.); the Novo Nordisk Foundation (grant no. NNF10CC1016515 to M.A.); the Danish Diabetes Academy supported by the Novo Nordisk Foundation and TARGET Research Initiative (Danish Strategic Research Council grant no. 0603-00484B to M.A.); the Matthias-Lackas Foundation (to C.M.U.); the National Cancer Institute (grant nos. R01 CA189184, R01 CA207371, U01 CA206110, and P30 CA042014 to C.M.U.); the Federal Ministry of Education and Research (BMBF; the de.NBI network no. 031A537B to P.B. and the ERA-NET TRANSCAN project no. 01KT1503 to C.M.U.); the Helmut Horten Foundation (to S.Sunagawa); and the Fundação de Amparo à Pesquisa do Estado de São Paulo (grant no. 16/23527-2 to A.M.T.). For the Italy validation cohorts, funding was provided by the Lega Italiana per La Lotta contro i Tumori. For the Japan validation cohort, funding was provided to T.Y. and S.Y. by the National Cancer Center Research and Development Fund (grant nos. 25-A-4,28-A-4, and 29-A-6); Practical Research Project for Rare/Intractable Diseases from the Japan Agency for Medical Research and Development (grant no. JP18ek0109187); Japan Science and Technology Agency-PRESTO (grant no. JPMJPR1507); Japan Society for the Promotion of Science KAKENHI (grant nos. 16J10135, 142558, and 221S0002); Joint Research Project of the Institute of Medical Science, University of Tokyo; and the Takeda Science and Suzuken Memorial Foundations.

## 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.

## Competing interest

P.B., G.Z., A.Y.V., and S.Sunagawa are named inventors on a patent (EP2955232A1: Method for diagnosing adenomas and/or colorectal cancer (CRC) based on analyzing the gut microbiome).

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41591-019-0406-6.

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41591-019-0406-6.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to M.A., P.B. or G.Z.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

## 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 Yu et al.[11], PRJEB12449 for Vogtmann et al.[10], ERP008729 for Feng et al.[9], and ERP005534 for Zeller et al.[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 Zeller et al.[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 Thomas et al.[27].

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 $1 \times 10^{-3}$ 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 (ref. [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 (ref. [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 were otherwise stated. To account for potential confounders and heterogeneity between data sets, we employed a stratified version of the Wilcoxon test[24]. 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[24]. 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[24].

**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 Pasolli et al.[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), $\log_{10}$-transformed (after adding a pseudo-count of $1\times10^{-5}$ to avoid non-finite values resulting from $\log_{10}(0)$[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 $1\times10^{-6}$ as the maximum abundance cutoff and $1\times10^{-9}$ 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 overoptimistically biased performance estimates (see Hastie et al.[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 where 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 $\log_{10}$-transformed abundances (pseudo-count of $10\times10^{-9}$, 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 $0.001^2$ to $0.2^2$. 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 Qin et al.[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 $1\times10^{-10}$ 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: $1 \times 10^{-7}$) 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.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

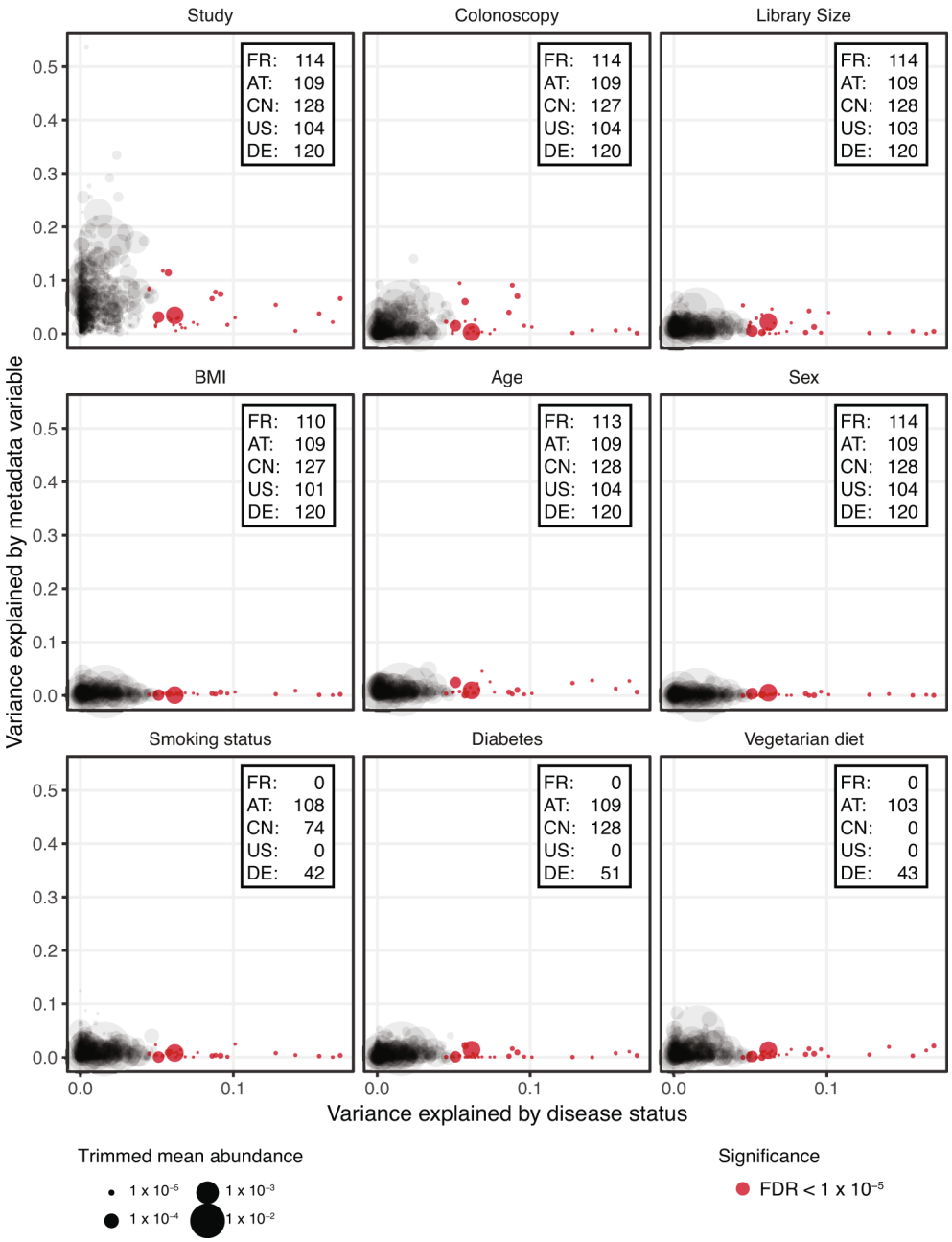
The raw sequencing data for the samples in the German study that have not been published before (see Methods) are available from the European Nucleotide Archive under study no. PRJEB27928. The metadata for these samples are available as Supplementary Table 6.

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 Yu et al.[11]; PRJEB12449 for Vogtmann et al.[10]; ERP008729 for Feng et al.[9]; and ERP005534 for Zeller et al.[8]. The independent validation cohorts can be found in the Sequence Read Archive under the identifier no. SRP136711 for Thomas et al.[27] 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.

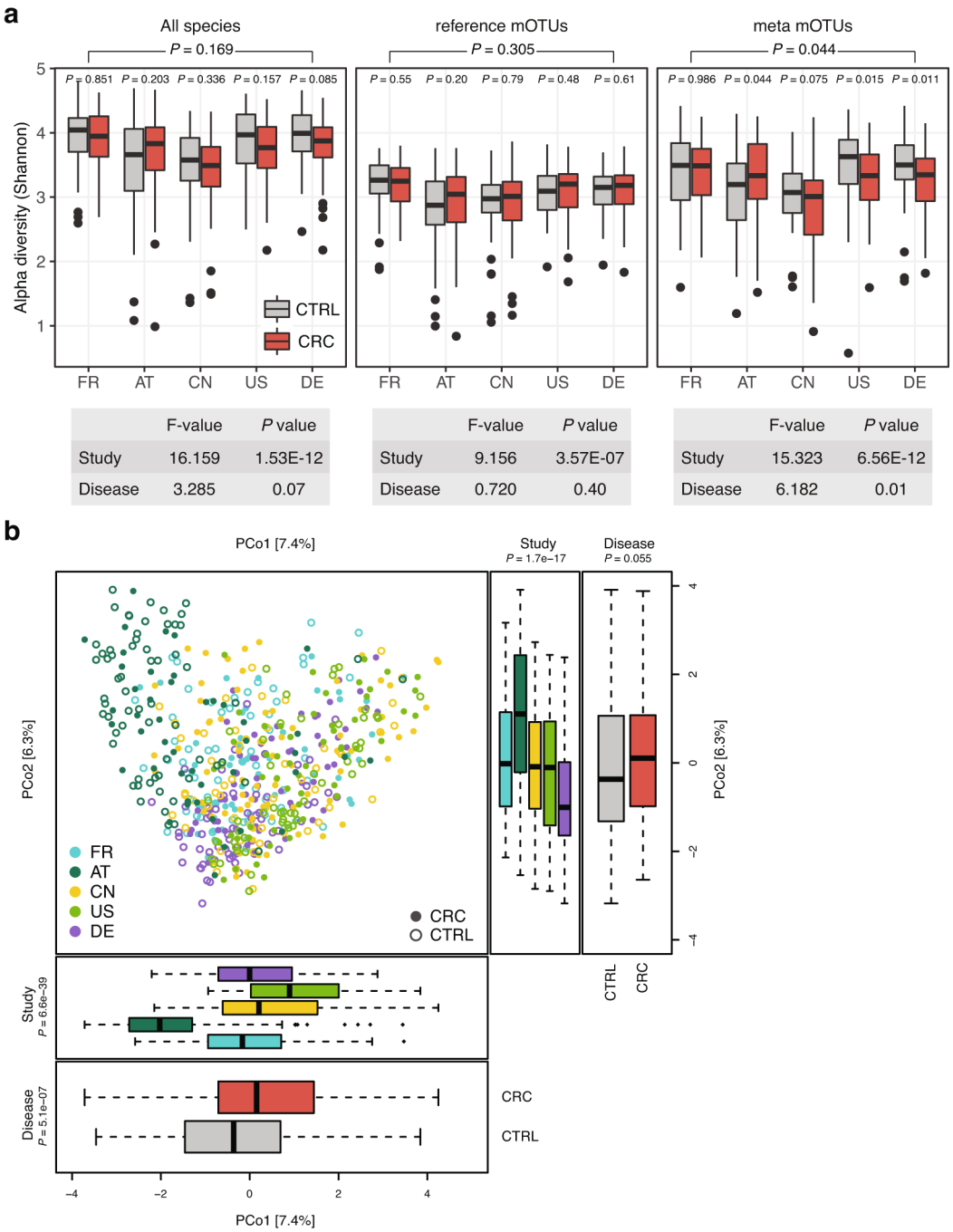
The code and all analysis results can be found under https://github.com/zellerlab/crc_meta.

## References

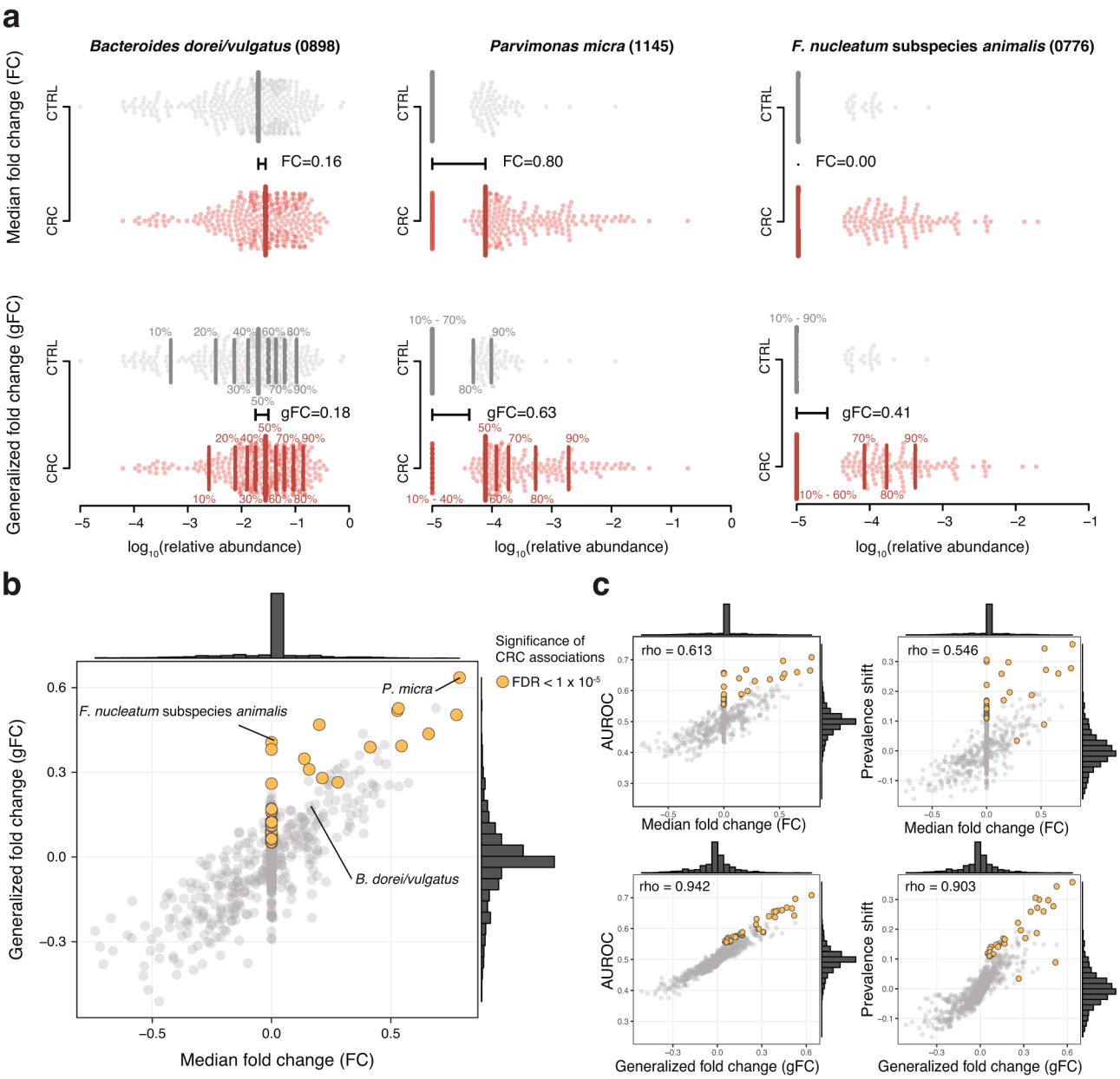

61. Böhm, J. et al. Discovery of novel plasma proteins as biomarkers for the development of incisional hernias after midline incision in patients with colorectal cancer: The ColoCare study. *Surgery* **161**, 808–817 (2017).
62. Liesenfeld, D. B. et al. Metabolomics and transcriptomics identify pathway 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**, 1460–1467.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. et al. 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. B* **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. B* **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. vegan: Community Ecology Package (The Comprehensive R Archive Network, 2018).
73. Costea, P. I., Zeller, G., Sunagawa, S. & Bork, P. A fair comparison. *Nat. Methods* **11**, 359 (2014).
74. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 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. USA* **108**, 4516–4522 (2011).



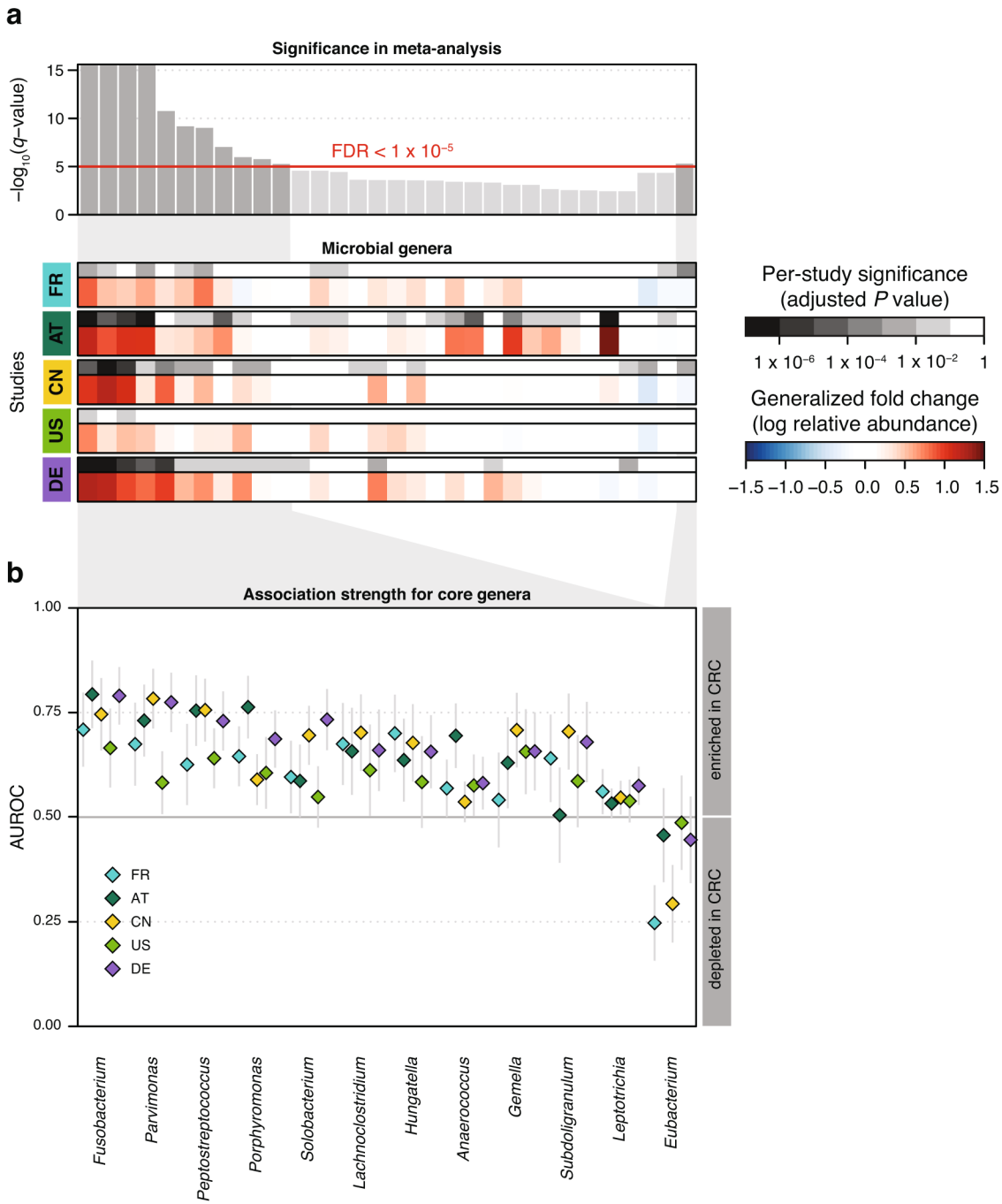

**Extended Data Fig. 1 | Potential confounding of individual microbial species associations by patient demographics and technical factors.** Variance explained by disease status (CRC versus CTRL) is plotted against variance explained by different putative confounding factors for individual microbial species. Each species is represented by a dot proportional in size to its abundance (see legend and Methods); core microbial markers identified in the meta-analysis are highlighted in red. For the confounder analysis, factors with continuous values were discretized into quartiles and the BMI was split into lean/overweight/obese according to conventional cutoffs. The variance explained by disease status was computed for all data; accordingly, the *x* values are the same in all panels and also in Fig. 1d. The variance explained by different confounding factors was computed using all samples for which data were available (indicated by the insets).



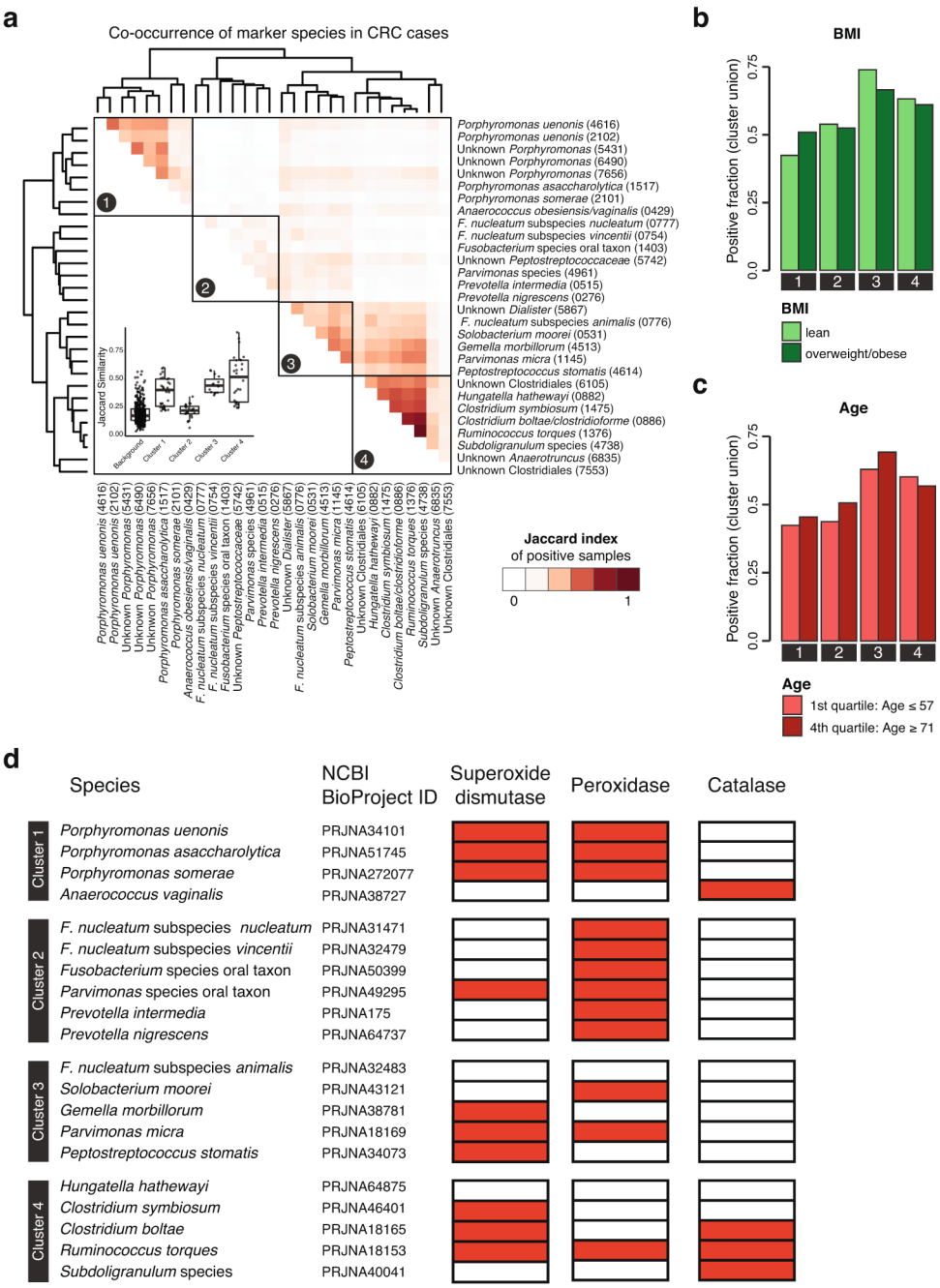

**Extended Data Fig. 2 | Study heterogeneity shows a strong influence on alpha and beta diversity. a**, Alpha diversity as measured with the Shannon index was computed for all gut microbial species ($n = 849$), reference mOTUs ($n = 246$), and meta-mOTUs ($n = 603$) separately. *P* values were computed using a two-sided Wilcoxon test, while the overall *P* value (on top) was calculated using a two-sided blocked Wilcoxon test ($n = 575$ independent observations; see Methods). The ANOVA F-statistics below the panel was calculated using the R function 'aov'. **b**, Principal coordinate analysis of samples from all five included studies based on Bray–Curtis distance; the study is color-coded and disease status (CRC versus CTRL) is indicated by filled/unfilled circles. The boxplots on the side and below show samples projected onto the first two principal coordinates broken down by study and disease status, respectively. *P* values were computed using a two-sided Wilcoxon test for disease status and a Kruskal–Wallis test for study ($n = 575$ independent observations). For all boxplots, boxes denote the IQR with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR. Country codes are as in Fig. 1b.



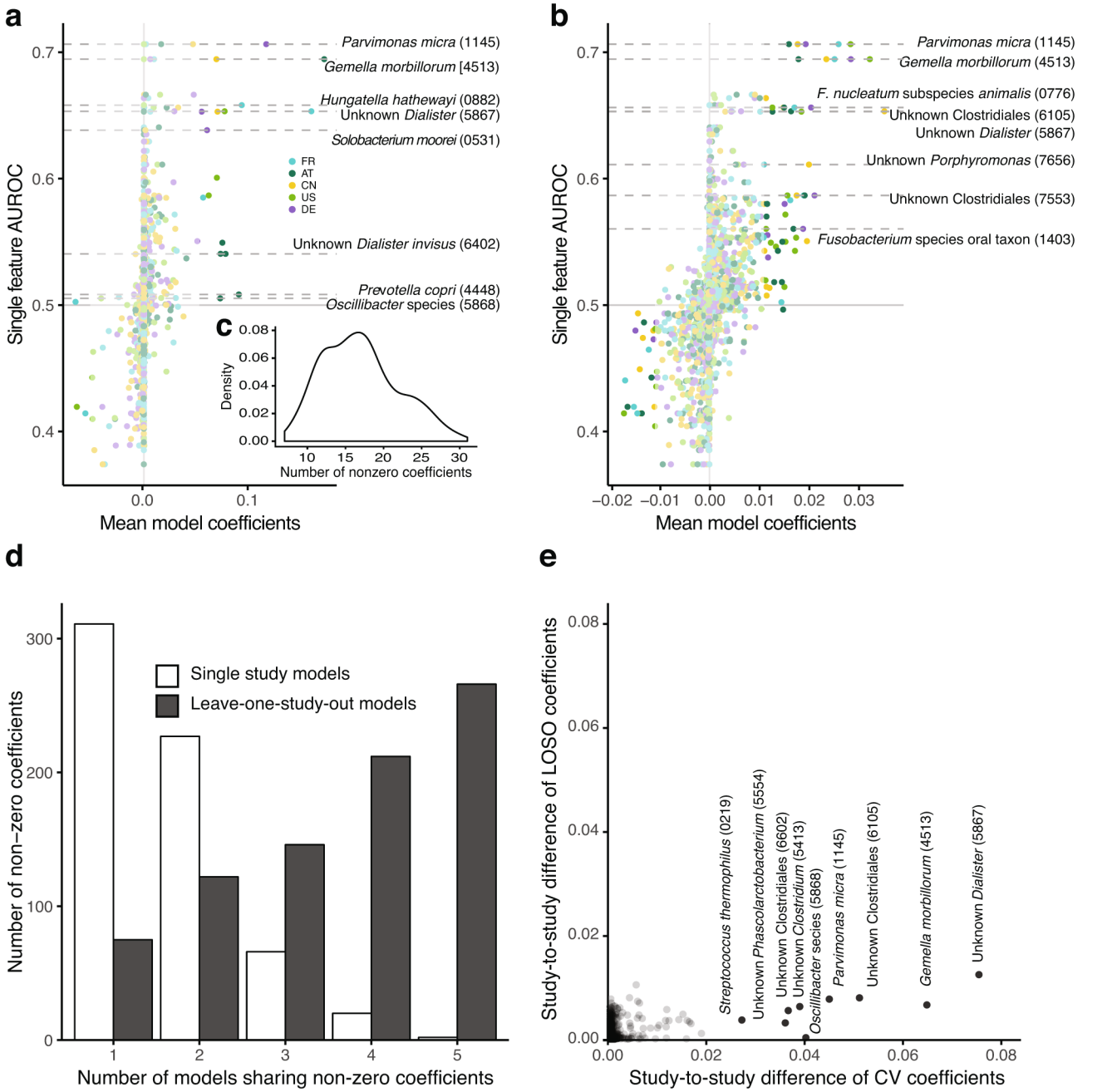

**Extended Data Fig. 3 | The generalized fold change extends the established (median-based) fold change to provide higher resolution in sparse microbiome data. a**, In the top row, the logarithmic relative abundances for *Bacteroides dorei/vulgatus*, *Parvimonas micra*, and *F. nucleatum* subspecies *animalis*—examples for one high-prevalence and two low-prevalence species—are shown as swarm plots for the CTRL and CRC groups. The thick vertical lines indicate the medians in the different groups and the black horizontal line shows the difference between the two medians, which corresponds to the classical (median-based) fold change. Since *F. nucleatum* subspecies *animalis* is not detectable in more than 50% of cancer cases, there is no difference between the CTRL and CRC median; thus, the fold change is 0. The lower row shows the same data, but instead of only the median (or 50th percentile), 9 quantiles ranging from 10 to 90% are shown by the thinner vertical lines. The generalized fold change is indicated by the horizontal black line again, computed as the mean of the differences between the corresponding quantiles in both groups. In the case of the sparse data (for example, *F. nucleatum*), the differences in the 70, 80, and 90% quantiles cause the generalized fold change to be higher than 0. **b**, The median fold change is plotted against the newly developed generalized fold change for all microbial species. (The core set of microbial CRC marker species is highlighted in orange.) Marginal histograms visualize the distribution for both fold change and generalized fold change. **c**, Scatter plots showing the relationship between fold change and generalized fold change and the area under the ROC curve (AUROC) or the shift in prevalence between CRC and CTRL, with Spearman's rank correlations (rho) added in the top left corners; the generalized fold change provides higher resolution (wider distribution around 0) and better correlation with the non-parametric AUROC effect size measure as well as prevalence shift, which captures the difference in prevalence of a species in CRC metagenomes relative to CTRL metagenomes.



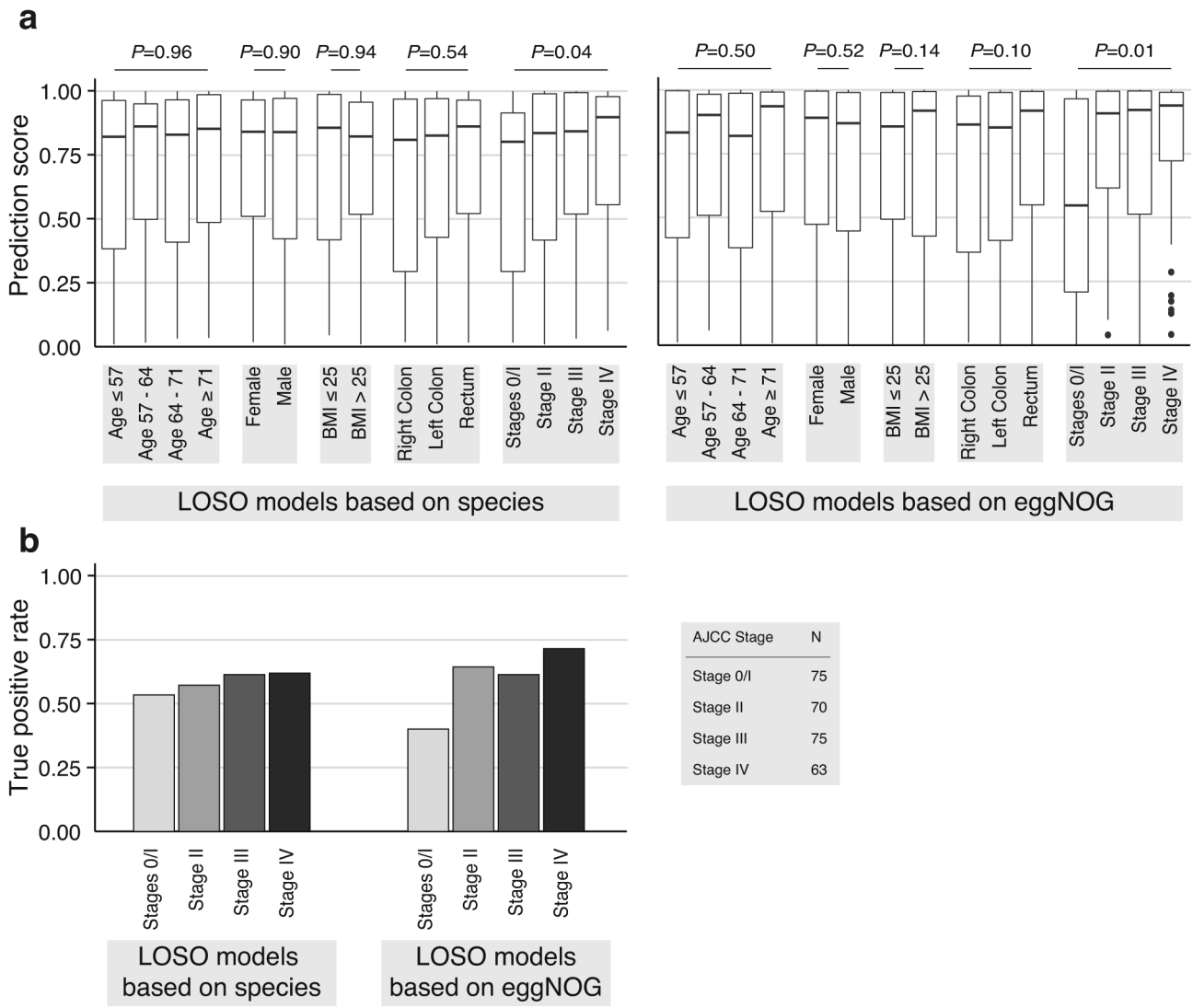

**Extended Data Fig. 4 | Microbial genera identified in the meta-analysis to be associated with CRC. a**, The meta-analysis significance of microbial genera, computed using a univariate, two-sided Wilcoxon test blocked for 'study' and 'colonoscopy' ($n = 574$ independent observations), is given by bar height (FDR = 0.005). Underneath, significance (FDR-corrected $P$ value computed using a two-sided Wilcoxon test) and generalized fold changes (see Methods) within individual studies are displayed as heatmaps in gray and color, respectively (see keys). Genera are ordered by meta-analysis significance and direction of change. **b**, For highly significant genera (meta-analysis FDR = $1 \times 10^{-5}$), association strength is quantified by the area under the ROC curve across individual studies (color-coded diamonds); 95% confidence intervals are depicted by gray lines. Country codes are as in Fig. 1b.



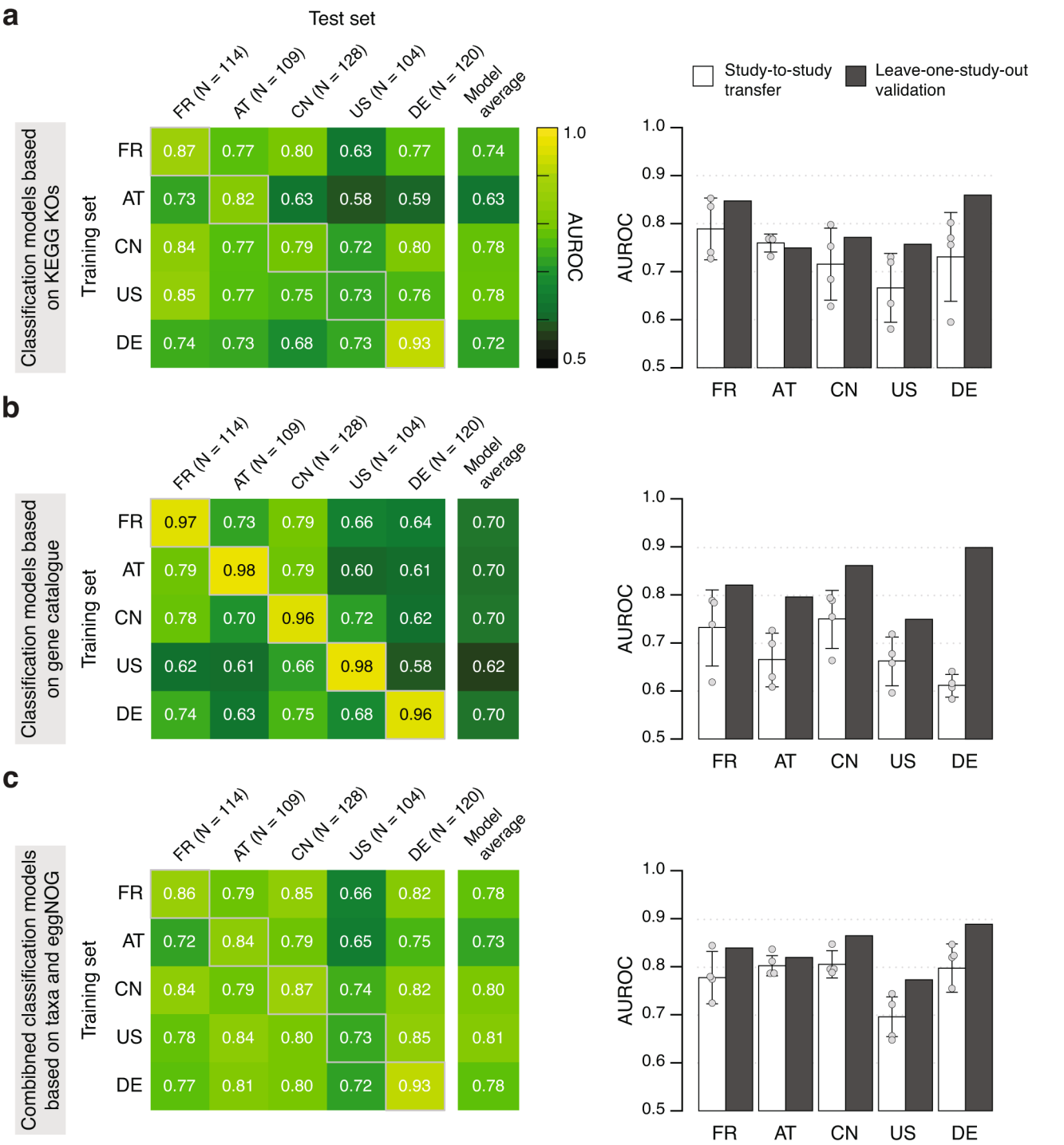

**Extended Data Fig. 5 | The core set of CRC-enriched microbial species can be stratified into four clusters based on co-occurrence in CRC metagenomes.** **a**, The heatmap shows the Jaccard index (computed by comparing marker-positive samples; see Methods) for the core set of microbial marker species, computed on CRC cases only. Clustering was performed using the Ward algorithm as implemented in the R function 'hclust'. The inset shows the distribution of Jaccard similarities within each cluster and for the background (all similarities between species not in the same cluster). **b**,**c**, Barplots show the fraction of CRC samples that are positive for a marker species cluster (defined as the union of positive marker species) broken down by patient subgroups based on BMI (**b**) and age (**c**) (see Fig. 2b–d for other patient subgroups). The significance of the associations between CRC subgroups and marker species clusters was tested using the Cochran–Mantel–Haenszel test blocked for 'study' and 'colonoscopy'. (No significant associations were detected.) **d**, For the core set of microbial species with a genomic reference, the presence (red) or absence (white) of superoxide dismutase, peroxidase, and catalase are shown as heatmaps. Presence of the enzyme was determined by checking the protein annotations for the reference projects (see NCBI BioProject ID) in http://progenomes.embl.de/.



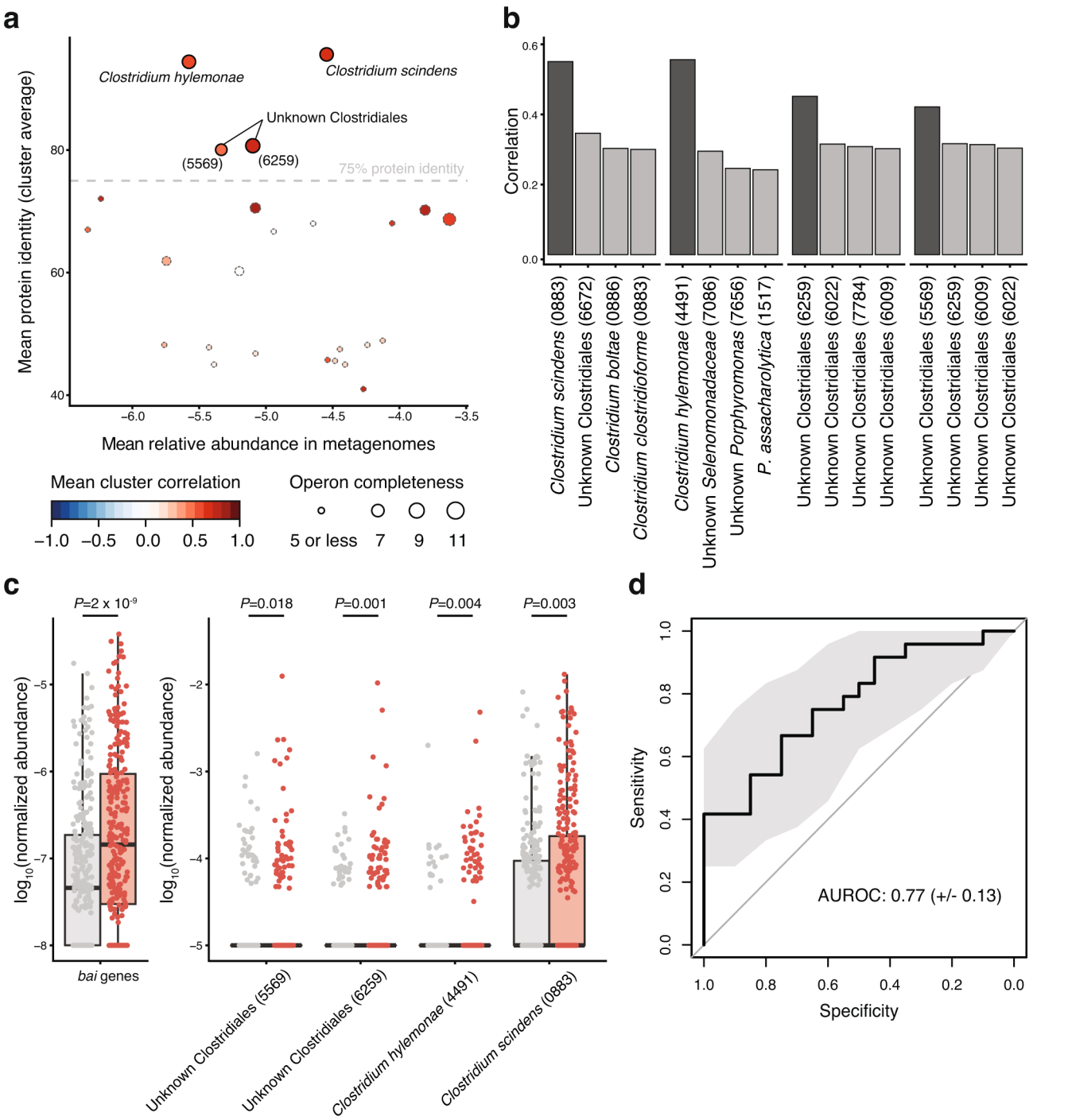

**Extended Data Fig. 6 | Coefficients of LOSO LASSO logistic regression models compared to models trained on individual studies. a**, The mean coefficients (feature weights) from LASSO cross-validation models trained on single studies (color-coded) are plotted against the single-feature AUROC for each species feature. The horizontal lines highlight the microbial species that are—for at least one study—selected in more than 50% of the models in cross-validation and account for more than 10% of the absolute model weight in at least 10% of the cross-validation models. **b**, Similarly, **b** shows the same for the models trained in the LOSO setting (see Methods). The colors indicate which study has been left out of the training set (and is used for validation). The weights of the LOSO models are spread across more species; thus, generally, lower species are highlighted by the horizontal lines if their weights explain more than 2.5% of the absolute model in at least 10% of cross-validation models and they have been selected in more than 50% of models in cross-validation. **c**, The inset shows the distribution of the number of non-zero coefficients across all cross-validation models. **d**, The bar height indicates the number of non-zero coefficients that are shared between the mean models for each study or left-out study, respectively. **e**, The study-to-study difference (computed as the median of all pairwise differences between model weights for a single species across the mean models) for cross-validation single-study models are plotted against the same measure for the LOSO models. Species with a study-to-study difference of more than 0.02 in the cross-validation models are highlighted and annotated, showing much larger variability between models trained on single studies compared to LOSO models. Country codes as in Fig. 1b.



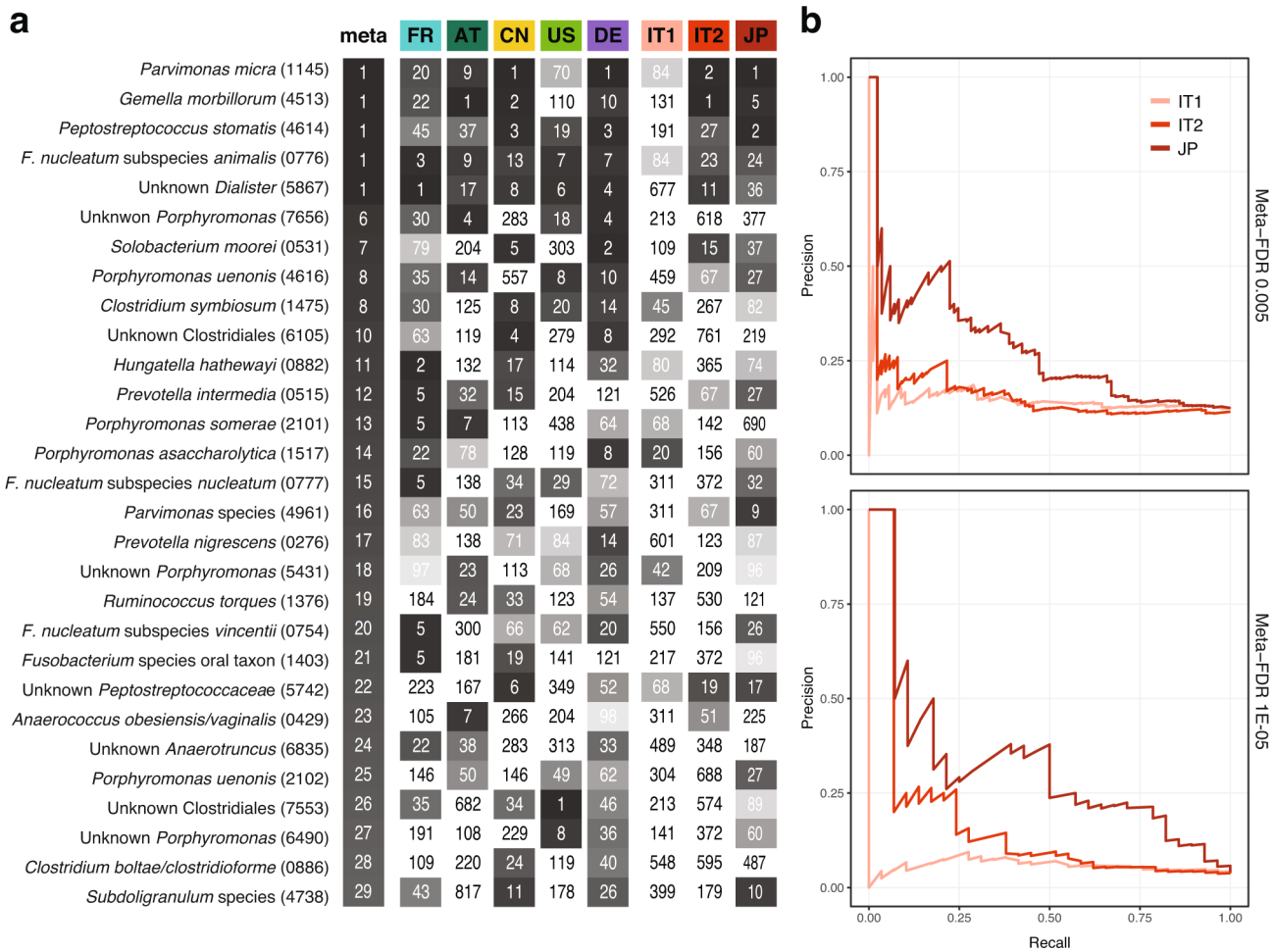

**Extended Data Fig. 7 | Analysis of LOSO models for prediction bias. a**, To examine whether species- and gene family-level classification models are confounded, that is, biased toward certain patient subgroups, the prediction scores from the LOSO models are broken down into strata for each clinical parameter (for example, female and male for sex). The prediction bias for each variable was tested by Wilcoxon (for sex and BMI) or Kruskal–Wallis (all others) tests while blocking for study as the confounder. The boxes denote the IQRs, with the median as the horizontal black line and the whiskers extending up to the most extreme point within the 1.5-fold IQR. A significant difference in prediction score was detected only for the CRC stage. This stage bias is more pronounced for gene family then for species models. **b**, To examine the CRC stage bias further, the barplots show the true positive rate corresponding to an overall 10% FPR (see also Fig. 3c) for the different CRC stages, displaying slightly higher classification sensitivity for late-stage CRC for both species and gene family models.




**Extended Data Fig. 8 | Cross-study performance of statistical models based on KEGG KO abundances, single-gene abundances from the metagenomic gene catalog (IGC), and the combination of taxonomic and eggNOG database abundance profiles. a–c**, CRC classification accuracy resulting from cross-validation within each study (gray boxed along the diagonal) and study-to-study model transfer (external validations off the diagonal) as measured by the AUROC for the classification models trained on KEGG KOs (**a**), models based on the gene catalog (**b**), and models based on the combination of taxonomic and eggNOG database abundance profiles (**c**) (see Methods for the details on the statistical modeling workflows). The last column depicts the average AUROC across external validations. The barplots on the right show that the classification accuracy on a hold-out study improves if the 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, indicated by the bar color) consistently across different types of input data. Country codes as in Fig. 1b.




**Extended Data Fig. 9 | Identification of *bai* genes in metagenomes.** Putative *bai* genes identified in the metagenomic IGC were clustered by co-abundance in metagenomes to infer genomic linkage (see Methods) to be able to infer operon completeness and species of origin. **a**, For each resulting cluster of putative bile acid-converting genes, the mean relative abundance was plotted against the mean percentage of protein identity derived from global alignment against the know bile acid-converting genes from *C. scindens* and *C. hylemonae* (see Methods). Completeness, that is, how many of the 11 different *bai* gene functions are represented in each cluster, and mean gene-to-gene correlation of relative abundance within each cluster are encoded by dot size and color, respectively (see legend). The four clusters with a mean protein identity > 75% to known *bai* operon-containing genomes were included in the subsequent analysis and labeled with the most highly correlated mOTU (see **b**). **b**, Pearson correlation between gene cluster abundances and the relative abundance of the most highly correlated species (in logarithmic space) is given by the bar height for the four gene clusters identified in **a**. The most highly correlating species is highlighted in darker gray (see labeling of gene clusters in **a**). **c**, The log-transformed abundances of all *bai* genes and the four species identified in **b** are shown as boxplots for CTRLs (gray) and CRC cases (red). Assessing the significance of the differences between CRC and CTRLs (using a Wilcoxon test blocked for 'study' and 'colonoscopy') demonstrates a much more significant CRC enrichment of the aggregated metagenomic *bai* gene abundance than of the individual clostridial species to which these belong. **d**, ROC curve for the qPCR quantification of the *baiF* gene in the genomic DNA of a subset of samples in the German study (see Methods and Fig. 4e).




**Extended Data Fig. 10 | Validation of the meta-analysis of single-species associations in three independent cohorts. a**, Heatmap showing for the core set of CRC-associated species (see Fig. 1) the rank of the respective species within the associations of each study, including the three independent validation cohorts (see Table 1), compared to the rank in the meta-analysis (meta) on the left. **b**, Precision-recall curves for the different independent validation cohorts using the meta-analysis set of associated species at FDR = 0.005 (top) and FDR = $1 \times 10^{-5}$ (bottom) as the 'true' set (see Methods) and the naïve (uncorrected) within-cohort significance as the predictor (see Supplementary Fig. 2). IT1, Italy 1; IT2, Italy 2; JP, Japan; other country codes are as in Fig. 1b.

nature research

Corresponding author(s): Georg Zeller, Peer Bork, Manimozhiyan Arumugam

Last updated by author(s): Feb 11, 2019

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Data was downloaded manually |
| Data analysis | mOTUS v2.0.0<br>SIAMCAT v1.1.0<br>R v3.5.1<br>MOCAT v2.0<br>MUSCLE v3.8.31<br>HMMER v3.1b2<br>EMBOSS v6.6.0<br>BWA v0.7.15-r1140<br>mRMR downloaded from http://home.penglab.com/proj/mRMR/ code version from 20 April 2009<br>GMMs v1.07<br>Additional code published on https://github.com/zellerlab/crc_meta |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw sequencing data for the samples in the DE study that were not published before (see Methods), are made available in the European Nucleotide Archive (ENA) under the study identifier PRJEB27928. Metadata for these samples are available as Supplementary Table S5. For the other studies included, the raw sequencing data can be found under the following ENA identifiers: PRJEB10878, PRJEB12449, ERP008729, and ERP005534 . The independent validation cohorts can be found in SRA under the identifier SRP136711 and in the DDBJ database under the ID DRA006684. The code and all analysis results can be found under https://github.com/zellerlab/crc_meta.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | No sample size calculation was performed for this meta-analysis, all publicly available data sets meeting a minimal set of inclusion criteria (see Methods and Supplementary Table S1) were included. New data sets generated (DE, JP, IT) were of similar sample size as previously published ones that described microbiome alterations in colorectal cancer. |
| Data exclusions | We used all data from cancer patients and neoplasia-free controls, but did not include any adenoma samples. Inclusion of studies is described in Table S1. |
| Replication | All attempts at replicating qPCR experiments were successful. |
| Randomization | Not applicable for this observational case-control meta-analysis. |
| Blinding | Blinding was not possible because statistical analyses depended on information about cancer status (statistical tests for differences between groups and supervised statistical modeling). |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |
| ☒ | ☐ Clinical data |

### Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | In our meta-analysis there is heterogeneity in the population characteristics recorded by the primary studies. Participant age, sex, BMI and sampling time relative to colonoscopy were recorded for all subjects with older age of CRC patients apparent in some studies (see Supplementary Table S2). |
| Recruitment | This is a meta-analysis with heterogeneous recruitment procedures performed in primary studies, not all of which are clearly |

| | |
|---|---|
| Recruitment | documented. At the level of this meta-analysis, selection bias by clinical investigators or due to differences in participant compliance with recommended CRC screening programs cannot be ruled out. |
| Ethics oversight | Patient recruitment for data that was newly generated for this study and consenting procedures were approved by Ethics Committees of the University of Heidelberg, Azienda Ospedaliera of Alessandria, the European Institute of Oncology of Milan, the National Cancer Center Japan - Research Institute, the Tokyo Institute of Technology, and EMBL. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.