nature medicine

ARTICLES

https://doi.org/10.1038/s41591-019-0405-7

Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation

Andrew Maltez Thomas[1,2,3,32], Paolo Manghi[1,32], Francesco Asnicar[1], Edoardo Pasolli[1], Federica Armanini[1], Moreno Zolfo[1], Francesco Beghini[1], Serena Manara[1], Nicolai Karcher[1], Chiara Pozzi[4], Sara Gandini[4], Davide Serrano[4], Sonia Tarallo[5], Antonio Francavilla[5], Gaetano Gallo[6,7], Mario Trompetto[7], Giulio Ferrero[8], Sayaka Mizutani[9,10], Hirotsugu Shiroma[9], Satoshi Shiba[11], Tatsuhiro Shibata[11,12], Shinichi Yachida[11,13], Takuji Yamada[9,14], Jakob Wirbel[15], Petra Schrotz-King[16], Cornelia M. Ulrich[17], Hermann Brenner[16,18,19], Manimozhiyan Arumugam[20,21], Peer Bork[15,22,23,24], Georg Zeller[15], Francesca Cordero[8], Emmanuel Dias-Neto[3,25], João Carlos Setubal[2,26], Adrian Tett[1], Barbara Pardini[5,27], Maria Rescigno[28], Levi Waldron[29,30,33], Alessio Naccarati[5,31,33] and Nicola Segata[1,33*]

Several studies have investigated links between the gut microbiome and colorectal cancer (CRC), but questions remain about the replicability of biomarkers across cohorts and populations. We performed a meta-analysis of five publicly available datasets and two new cohorts and validated the findings on two additional cohorts, considering in total 969 fecal metagenomes. Unlike microbiome shifts associated with gastrointestinal syndromes, the gut microbiome in CRC showed reproducibly higher richness than controls ($P < 0.01$), partially due to expansions of species typically derived from the oral cavity. Meta-analysis of the microbiome functional potential identified gluconeogenesis and the putrefaction and fermentation pathways as being associated with CRC, whereas the stachyose and starch degradation pathways were associated with controls. Predictive microbiome signatures for CRC trained on multiple datasets showed consistently high accuracy in datasets not considered for model training and independent validation cohorts (average area under the curve, 0.84). Pooled analysis of raw metagenomes showed that the choline trimethylamine-lyase gene was overabundant in CRC ($P = 0.001$), identifying a relationship between microbiome choline metabolism and CRC. The combined analysis of heterogeneous CRC cohorts thus identified reproducible microbiome biomarkers and accurate disease-predictive models that can form the basis for clinical prognostic tests and hypothesis-driven mechanistic studies.

Colorectal cancer is the second most common non-sex-specific cancer and is responsible for more deaths than any other cancer after lung cancer[1]. Because of demographic trends toward an aging population, the annual global incidence rate is expected to increase by nearly 80% to 2.2 million cases over the next two decades[2]. Sporadic CRCs, as opposed to hereditary CRCs, account for approximately 70–87% of cases[3] and genetics can explain only a small proportion of disease incidence[4]. The missing strong link

[1]Department CIBIO, University of Trento, Trento, Italy. [2]Biochemistry Department, Chemistry Institute, University of São Paulo, São Paulo, Brazil. [3]Medical Genomics Laboratory, CIPE/A.C. Camargo Cancer Center, São Paulo, Brazil. [4]European Institute of Oncology, Milan, Italy. [5]Italian Institute for Genomic Medicine, Turin, Italy. [6]Department of Surgical and Medical Sciences, University of Catanzaro, Catanzaro, Italy. [7]Department of Colorectal Surgery, Clinica S. Rita, Vercelli, Italy. [8]Department of Computer Science, University of Turin, Turin, Italy. [9]School of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan. [10]Research Fellow of Japan Society for the Promotion of Science, Tokyo, Japan. [11]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. [12]Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [13]Department of Cancer Genome Informatics, Osaka University, Osaka, Japan. [14]PRESTO, Japan Science and Technology Agency, Saitama, Japan. [15]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [16]Division of Preventive Oncology, National Center for Tumor Diseases and German Cancer Research Center, Heidelberg, Germany. [17]Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, UT, USA. [18]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany. [19]German Cancer Consortium, German Cancer Research Center, Heidelberg, Germany. [20]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [21]Faculty of Healthy Sciences, University of Southern Denmark, Odense, Denmark. [22]Molecular Medicine Partnership Unit, Heidelberg, Germany. [23]Max Delbrück Centre for Molecular Medicine, Berlin, Germany. [24]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany. [25]Laboratory of Neurosciences, Institute of Psychiatry, University of São Paulo, São Paulo, Brazil. [26]Biocomplexity Institute of Virginia Tech, Blacksburg, VA, USA. [27]Department of Medical Sciences, University of Turin, Turin, Italy. [28]Mucosal Immunology and Microbiota Unit, Humanitas Research Hospital, Milan, Italy. [29]Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA. [30]Institute for Implementation Science in Population Health, City University of New York, New York, NY, USA. [31]Department of Molecular Biology of Cancer, Institute of Experimental Medicine, Prague, Czech Republic. [32]These authors contributed equally: Andrew Maltez Thomas, Paolo Manghi. [33]These authors jointly supervised this work: Levi Waldron, Alessio Naccarati, Nicola Segata. *e-mail: nicola.segata@unitn.it

of CRC with genetics points to the potential role of other variables, including lifestyle and environmental factors, as disease co-determinants. Reported risk factors associated with CRC include age, tobacco and alcohol consumption, lack of physical activity, increased body weight and diet[5,6]. However, many non-genetic risk factors are common to several cancer types and these remain largely unsettled for CRC[7,8].

The human gut microbiome—defined as the microbial communities that populate our intestinal tract—is emerging as a relevant factor in human diseases[9,10]. Supported by some evidence of carcinogenic mechanisms induced by bacterial organisms[11–13], the gut microbiome has also been hypothesized to play a crucial role in the development of CRC. Studies using 16S ribosomal RNA (rRNA) gene amplicon sequencing have led to the discovery of the association of *Fusobacterium nucleatum* with CRC[14], which was subsequently shown to be causal in animal models of CRC carcinogenesis and progression[15,16]. Compared to 16S rRNA gene studies, a smaller number of metagenomic sequencing studies have linked other microbial species and potential functional activities of the gut microbiome to CRC[17–19,20]. However, the reproducibility and predictive accuracy of these high-resolution microbial signatures across cohorts and study design choices remain unclear. The potential use of the gut microbiome as a diagnostic tool for CRC has been proposed[17–19,21,22], but not yet validated, across multiple independent study populations.

There is thus a need to establish and validate links between the human gut microbiome and CRC carcinogenesis across populations, cohorts and microbiome tools. Some multi-cohort works have been performed based on 16S rRNA gene studies[23], but this technique has important technical limitations[24]. The recent availability of whole-metagenome shotgun datasets of CRC cohorts[17–19,21,22] enables a combined multi-population exploration of the CRC-associated microbiome with strain-level resolution[25,26] and meta-analytic predictive approaches[10,27], but the only meta-analysis study performed to date on CRC is affected by overfitting issues[28]. It is thus crucial to perform large-scale, cross-cohort studies to provide an unbiased and well-powered assessment of the link between CRC and the gut microbiome.

In this study, we have sequenced 140 samples from two different cohorts, performed an integrated analysis combining all current metagenomic CRC datasets available and assessed prediction accuracies of the gut microbiome for CRC detection across populations, datasets and conditions.

Results

A meta-analysis of metagenomic datasets to identify links between the gut microbiome and CRC. To identify reproducible relationships between the gut microbiome and CRC, we performed shotgun metagenomic sequencing[29] of the stool microbiome of 140 patients with CRC and controls recruited in two cohorts, and analyzed these in the context of 624 additional samples from five publicly available and geographically diverse metagenomic studies. We validated the results on two datasets of 60 CRC and 65 controls[30] and 40 CRC and 40 controls (see Methods), respectively. In total, we considered 413 samples from patients with CRC, 143 from subjects with adenoma and 413 control samples. Participants from all studies underwent colonoscopy to diagnose CRC or adenoma, or to confirm the absence of disease, with samples collected before diagnosis or the beginning of treatment (Supplementary Table 1 and Table 1). All datasets were sequenced at high depth except for the study in ref. [31] (Extended Data Fig. 1a and Methods).

Meta-analysis shows higher species richness in CRC-associated samples. We first tested whether microbial richness and diversity differed between CRC samples and controls, given contrasting current evidence[32–34]. In all but one study the median species richness was higher in CRC samples compared to controls, and the increase was significant in four of the six deeply sequenced datasets ($P < 0.05$; Extended Data Fig. 1b,c). Meta-analysis of standardized mean differences by random effects model for the number of microbial species confirmed the higher number of species in CRC compared to controls (meta-analysis coefficient estimate (μ) = 0.5, 95% confidence interval (0.16, 0.85), $P = 0.004$), although with significant heterogeneity across datasets (percentage of total variation due to heterogeneity (I^2) = 74.8%, $P = 0.0007$, Q-test). This difference was not meaningfully affected when controlling for potential confounding by age, body mass index (BMI) or sex (Extended Data Fig. 1d,e). Conversely, we observed no difference in diversity between carcinomas and controls (Extended Data Fig. 2a,b). We thus provide strong evidence that the CRC-associated microbiome has a quantitative species distribution that is consistent with healthy controls, but is significantly enriched in the total number of detected microbes.

We further tested whether the CRC-associated microbiome possesses more oral cavity-associated species than controls, as previously hypothesized[23,35]. Considering the 161 species we identified from multiple existing datasets[36,37] as being typical colonizers of the oral cavity (see Methods), we found increased oral species richness in CRC samples for all but one of the six deeply sequenced datasets compared to controls, and the increase was significant in meta-analysis ($\mu = 0.16$, 95% confidence interval (−0.03, 0.35), $P = 0.02$; Extended Data Fig. 2g). Similarly, the total abundance of oral species in the stool microbiome was also significantly higher in patients with CRC compared to controls (meta-analysis $\mu = 0.23$, 95% confidence interval (0.07, 0.39), $P = 0.003$). Altogether, greater species richness and abundance may be a sign of an altered gut microbiome in CRC, and it is indicative of an influx of bacterial species originating from the oral cavity.

A panel of microbial biomarkers for CRC is reproducible across cohorts. Individual biomarker discovery efforts can be sensitive to both technical artifacts and the heterogeneity of factors implicated in microbial shifts in healthy populations, including biogeography, diet and host genetics[26,38]. This is confirmed by the two newly sequenced datasets that have only partially overlapping taxonomic and functional potential biomarkers (Extended Data Fig. 3). Even so, several CRC biomarker species were identified by univariate statistics[39] independently in the majority of the datasets (Fig. 1a): *F. nucleatum*, *Solobacterium moorei*, *Porphyromonas asaccharolytica*, *Parvimonas micra*, *Peptostreptococcus stomatis* and *Parvimonas* spp. Other species were identified in fewer datasets or were dataset-specific (Fig. 1a and Supplementary Table 2). *F. nucleatum*, whose connection with CRC has been extensively reported[14,17–19], had significantly increased abundance in patients with CRC in all datasets with adequate sequencing depth, when considering single markers for this species (Extended Data Fig. 4a). Some of the cross-cohort CRC biomarker species have previously been reported[14,23,35] and many of these are commonly found in the oral cavity (eight out of the 39 total biomarkers found in at least two datasets), consistent with the increased presence of oral taxa in CRC samples mentioned above.

We then pooled evidence of differential abundance across datasets by random effects meta-analysis. Among the 26 differentially abundant species at false discovery rate (FDR) < 0.005, those with the highest effect size were again *F. nucleatum*, *S. moorei*, *P. asaccharolytica*, *P. micra* and *P. stomatis*. The meta-analysis additionally identified *Clostridium symbiosum*, which has been tested as a marker for early CRC detection[40] (Fig. 1b). Other differentially abundant species at FDR < 0.05 have not previously been reported in CRC microbiome studies, including *Streptococcus tigurinus* and *Streptococcus dysgalactiae*, and three different *Campylobacter* species. We also confirmed *Gemella morbillorum* and *Streptococcus gallolyticus* as being relevant biomarkers, as previously suggested in

Table 1 | Size and characteristics of the large-scale CRC metagenomic datasets included in this study

Dataset	Groups (*n*)	Age (average ± s.d.)	BMI (average ± s.d.)	Sex F (%)/M (%)	Country	No. of reads (×10⁹)
ZellerG_2014 (ref. [19])	Control (61)	60.6 ± 11.4	24.7 ± 3.2	54.1/45.9		
	Adenoma (42)	63 ± 9.1	25.9 ± 4.1	28.5/71.5	France	9.4
	CRC (53)	66.8 ± 10.9	25.5 ± 5.2	45.2/54.8		
YuJ_2015 (ref. [17])[a]	Control (54)	61.8 ± 5.7	23.5 ± 3	38.9/61.1	China	7.2
	CRC (74)	66 ± 10.6	24 ± 3.2	35.1/64.9		
FengQ_2015 (ref. [18])	Control (61)[b]	67 ± 6.5	27.6 ± 3.8	41/59		
	Adenoma (47)	66.5 ± 7.9	28 ± 4.7	51.1/48.9	Austria	8.3
	CRC (46)	67 ± 10.9	26.5 ± 3.5	39.1/60.9		
VogtmannE_2016 (ref. [20])	Control (52)	61.2 ± 11	25.3 ± 4.2	28.8/71.2	USA	6.9
	CRC (52)	61.8 ± 13.6	24.9 ± 4.2	28.8/71.2		
HanniganGD_2018 (ref. [31])	Control (28)				USA (54)	
	Adenoma (27)	NA	NA	NA	Canada (28)	0.5
	CRC (27)					
Cohort1 (This study)	Control (24)	67.9 ± 7.1	25.3 ± 3.5	45.8/54.1		
	Adenoma (27)	62.8 ± 8.6	25.3 ± 4.1	40.7/59.3	Italy	8.2
	CRC (29)	71.4 ± 8.2	25.7 ± 4.1	20.7/79.3		
Cohort2 (This study)	Control (28)	57.8 ± 8.3	24.6 ± 3.8	42.9/57.1	Italy	5.1
	CRC (32)	58.4 ± 8.4	26.8 ± 4.3	28.1/71.9		
Total	**Control (308)**					
	Adenoma (143)					**45.6**
	CRC (313)					

[a]There is an updated version of this paper that was published in 2017, but the paper was first published in 2015; 2015 has been included in the dataset name, as it better reflects when the data was produced. [b]Numbers differ from those of the original sample ($n = 61$ rather than 63) reported in the article due to metadata and/or sequence-processing issues. NA, not available. The last three rows in bold refer to the combined stats of all cohorts above and were included to highlight the number of samples for each condition and the total number of reads; combined stats could not be computed because of missing data for one of the cohorts.

smaller cohorts[18,41]. In contrast, only 12 species were associated with the control population in the meta-analysis and only four were significantly enriched for the same populations in at least three datasets. Control-associated species with the highest effect sizes were *Gordonibacter pamelae* and *Bifidobacterium catenulatum* (Fig. 1b, Supplementary Table 2 and Extended Data Fig. 4c), which are generally considered beneficial microbes and have been used as probiotic supplements[42]. Adjustment for potential confounding by host characteristics did not meaningfully affect crude estimates in the meta-analysis (Fig. 1d and Extended Data Fig. 4b). The substantially higher number of species enriched in CRC as opposed to controls (49 versus 12), even when focusing only on species with putative oral origin (15 versus 2; Extended Data Fig. 5a), points to the existence of a reproducible taxonomic signature of the CRC-associated microbiome.

Functional potential of the microbiome was also significantly associated with CRC samples when compared to healthy controls. We found overall increased richness of UniRef gene families[43] in CRC samples in two datasets, with percentages of unmapped reads ranging between 20 and 40 (Extended Data Fig. 5e). We found 33,840 of the 2,479,274 single gene families detected at least once to be associated with CRC and 30,475 associated with controls at FDR < 0.05 (9,154 and 7,115 differential gene families at FDR < 0.005). We further observed 136 out of 590 metagenomically reconstructed microbial functional pathways to be CRC-associated, and only 37 associated with controls (Supplementary Table 3). Among the most differentially abundant pathways (Fig. 1c) that are at worst only minimally affected by potential confounding factors (Fig. 1e), we found starch, stachyose and galactose degradation to be associated with controls. These associations may indicate how potentially diet-associated changes in the functional repertoire of the microbiome can influence host conditions. The CRC-associated microbiome showed an association with both gluconeogenesis and the capacity for uptake and metabolism of amino acids via putrefaction and fermentation pathways (Supplementary Tables 3 and 4). These included pathways responsible for the conversion of different amino acids to tumor-promoting compounds[19,44], such as polyamines (for example, L-arginine and L-ornithine degradation to putrescine) and ammonia (L-histidine and L-arginine degradation, and L-lysine and L-alanine fermentation to acetate, butyrate and propionate). These pathways (Fig. 1c) and the set of species described above (Fig. 1a,b) thus constitute a collection of microbiome biomarkers that are reproducible across cohorts.

Predicting CRC from single metagenomic datasets in independent cohorts leads to reduced accuracy. To test the hypothesis that the stool microbiome could be used as a reproducible CRC prescreening tool, we performed intra-cohort, cross-cohort and combined-cohort prediction validation on the overall set of 621 CRC and control samples using a random forest classifier (Table 1). In intra-cohort cross-validation using species-level taxonomic relative abundances, we observed performances ranging in area under the receiver operating characteristic curve (AUC) score from 0.92 to 0.58, with an average in the deeply sequenced datasets of 0.81 (Fig. 2a). When using the functional potential of the gut microbiome by means of pathway abundances we observed decreased single-dataset cross-validation accuracies, with the exception of our Cohort1 (maximum 0.82 AUC, average 0.71 AUC; Extended Data Fig. 6a).

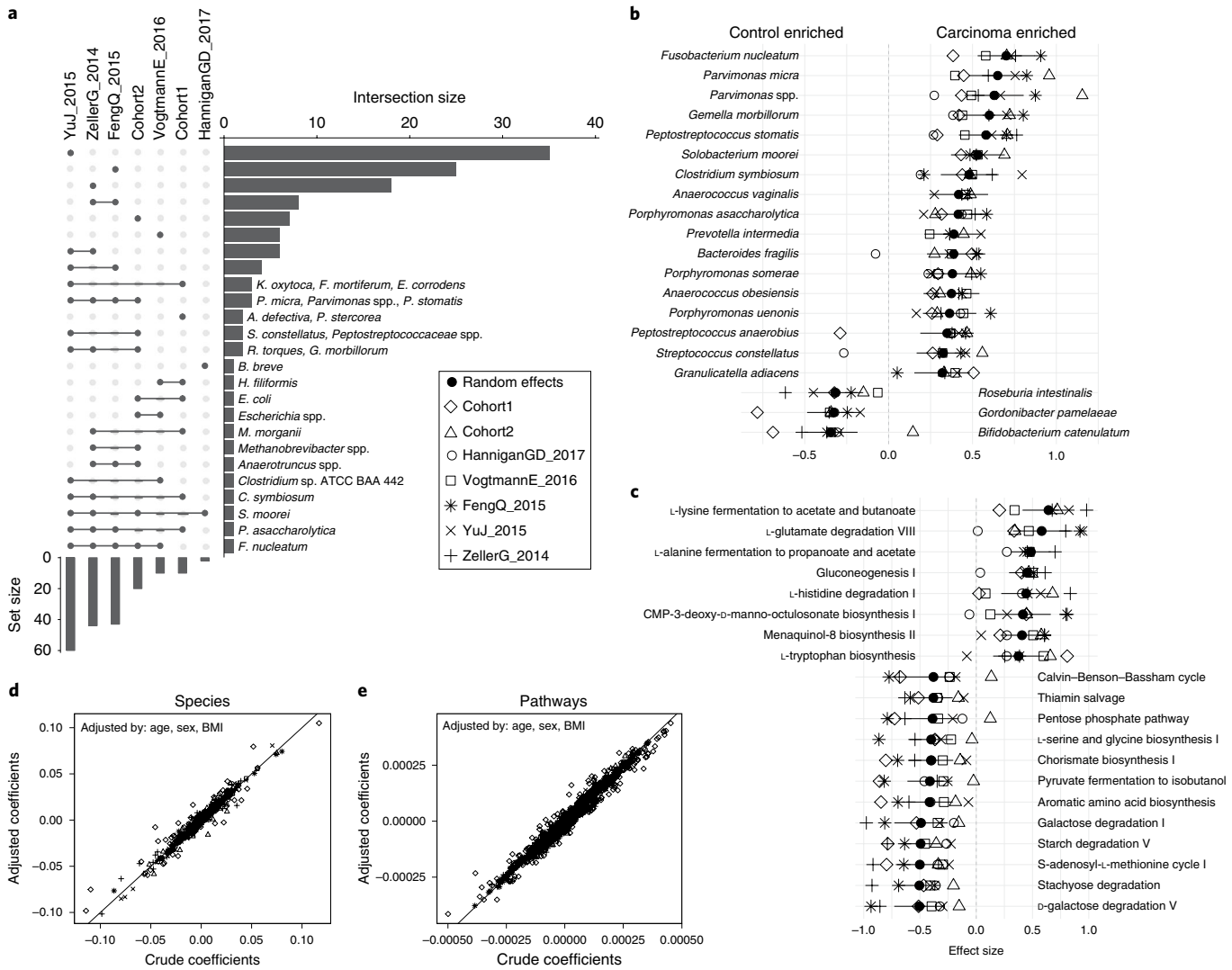


Fig. 1 | Reproducible taxonomic and functional microbial biomarkers across datasets when comparing carcinoma to healthy controls (no adenoma samples considered). a, UpSet plot showing the number of taxonomic biomarkers identified using LEfSE on MetaPhlAn2 species profiles shared by combinations of datasets (see Supplementary Table 3 for all single significant associations). **b**,**c**, Pooled effect sizes for the 20 significant features with the largest effect size, calculated using a meta-analysis of standardized mean differences and a random effects model on MetaPhlAn2 species abundances (**b**) and HUMANn2 pathway abundances (**c**). Bold lines represent the 95% confidence interval for the random effects model coefficient estimate. **d**, Scatter plot of crude and age-, sex- and BMI-adjusted coefficients obtained from linear models using MetaPhlAn2 species abundances. **e**, Scatter plot of crude and age-, sex- and BMI-adjusted coefficients obtained from linear models using HUMANn2 pathway abundances.

The profiling of the more fine-grained UniRef90 gene-family abundances improved predictions, with AUC scores reaching 0.8 for Cohort2 and an average of 0.77 in the deeply sequenced datasets (Fig. 2b). These results show that, while cross-validation AUCs can be high for prediction of CRC in certain datasets, these are highly variable and dataset dependent.

We then tested whether and how much the microbial signatures of CRC remained predictive across distinct datasets and cohorts. To this end, we trained the classifier on each single 'training' dataset and applied the model to each distinct 'testing' dataset. For most datasets this led to decreased AUC values when compared to single cross-validation AUCs, and AUCs showed high variability across cohorts (minimum 0.5 and maximum 0.86 AUC for cross-dataset). These results were consistent when using either pathway or gene-family abundances as predictors (Extended Data Fig. 6a and Fig. 2b). Overall, we highlight a poor transportability of the microbiome signature from one dataset to the other, and experimental choices[45] and cohort or population characteristics[26] may explain the reduced cross-study predictability when considering single datasets to train the model (Extended Data Fig. 6c,d).

Pooling of training cohorts substantially improves prediction across datasets. To overcome the limitations of training on single datasets (Supplementary Table 5), we performed a leave-one-dataset-out (LODO) analysis[46] in which classifiers were trained on six datasets combined, and validated on the left-out dataset, for each dataset in turn. For taxonomic profiles, this approach improved both AUC values and inter-dataset consistency, producing AUCs $\geq$0.80 (average = 0.84, s.d. = 0.03) for all six deeply sequenced datasets (Fig. 2a). Predictors based on clade-specific markers also produced high, albeit more variable, AUC values, outperforming taxonomic profiles in some datasets (Extended Data Fig. 6b). Gene families achieved slightly reduced performances, whereas pathway abundances produced substantially less accurate predictions (Fig. 2b and Extended Data Fig. 6a). The technical and host population diversity embedded in these training

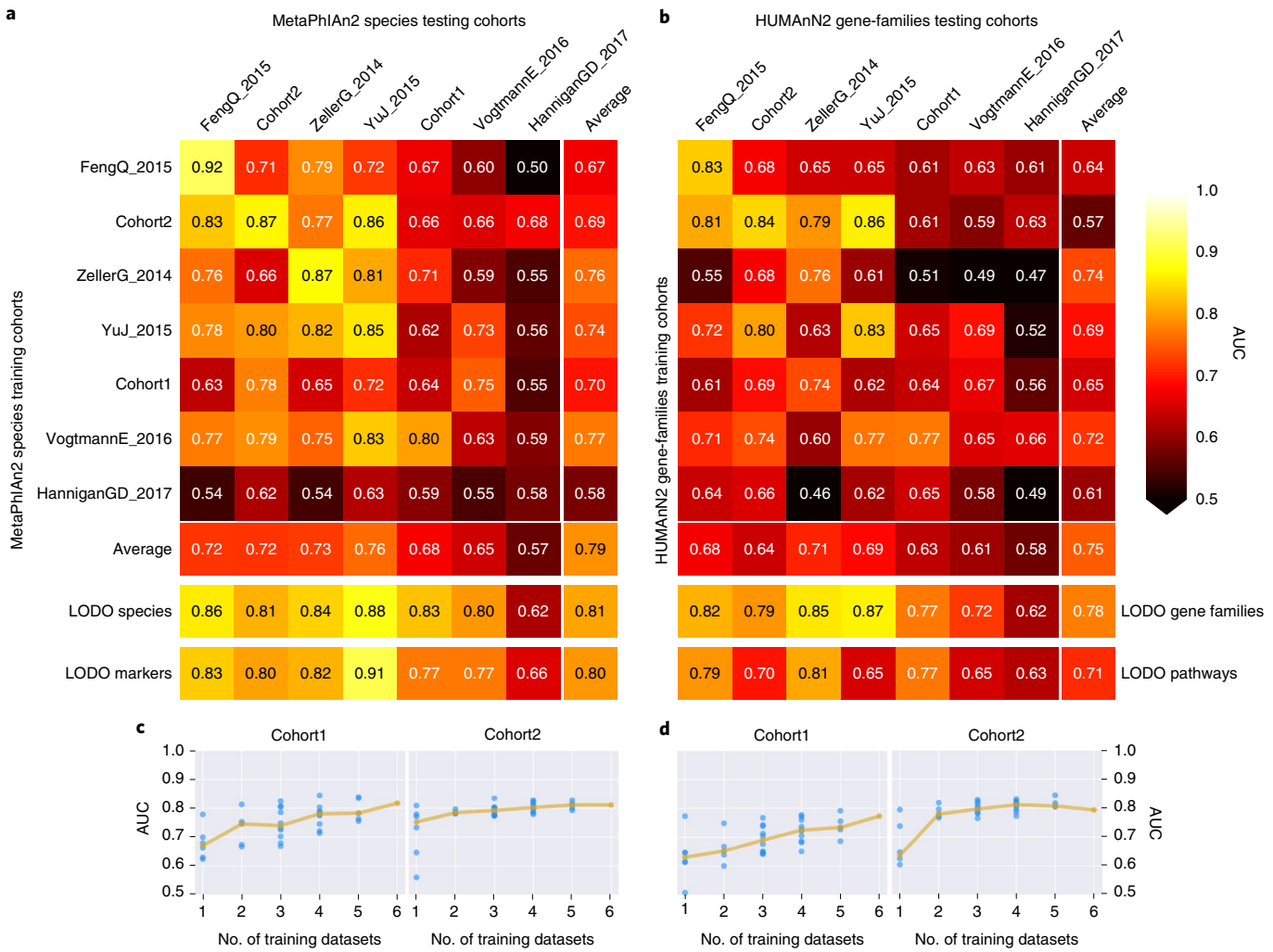


Fig. 2 | Assessment of prediction performances of the gut microbiome for CRC detection within and across cohorts. a, Cross-prediction matrix reporting prediction performances as AUC values obtained using a random forest model on species-level relative abundances (see Methods). Values on the diagonal refer to 20 times repeated tenfold stratified cross-validations. Off-diagonal values refer to the AUC values obtained by training the classifier on the dataset of the corresponding row and applying it to the dataset of the corresponding column. The LODO rows refer to the performances obtained by training the model on the species-level abundances and MetaPhlAn2 markers, using all but the dataset of the corresponding column and applying it to the dataset of the corresponding column. See Extended Data Fig. 6 for the marker cross-study validation matrix. **b**, Cross-prediction matrix of AUC values obtained using HUMANn2 UniRef90 gene-family abundances and HUMANn2 pathway relative abundances. See Extended Data Fig. 6 for the pathway cross-study validation matrix. **c**, Prediction performances for the two Italian cohorts at increasing numbers of external datasets considered for training the model. The dark yellow line interpolates the median AUC at each number of training datasets considered. See Extended Data Fig. 7 for the plots referred to in the other testing datasets. **d**, Prediction performances at increasing number of datasets considered in the training, using HUMANn2 UniProt90 gene-family abundances. See Extended Data Fig. 7 for the plots referred to in the other testing datasets.

meta-cohorts may be crucial in improving the generalizability of classifiers, as we found this LODO approach to be substantially and consistently more informative than a single-dataset cross-validation, and independent investigations found similarly high LODO performances using different metagenomic profilers and machine learning tools[30].

The model trained on taxonomic or functional features was also shown to capture the above whole-microbiome biomarkers because the direct inclusion of alpha-diversity metrics, oral species abundance and a measure of metagenome mappability did not provide substantial improvements (average = 0.83, s.d. = 0.03 for the deeply sequenced datasets when the additional features were combined with the taxonomic model). However, based on the performance and variability of the predictive models across datasets, we recommend using species-level microbial abundance as the main feature set for CRC status prediction in a LODO setting.

To assess the relation between population diversity in the training meta-cohort and prediction performance, we considered increasingly larger subsets of the available training cohorts. AUC values sharply increased when moving from one to two training datasets (10–13% median AUC improvement depending on the features considered in the model), with less marked improvements at further dataset additions (Fig. 2c,d and Extended Data Fig. 7). Large and heterogeneous combined training sets thus generate improved accuracy for identifying CRC cases in independent metagenomic datasets.

Accurate predictive models using a minimal microbial signature. The predictive CRC-associated microbiome signatures identified above considered all observed species and gene functions, and would thus be impractical for clinical application without whole-microbiome profiling. We thus sought to identify a minimal set of

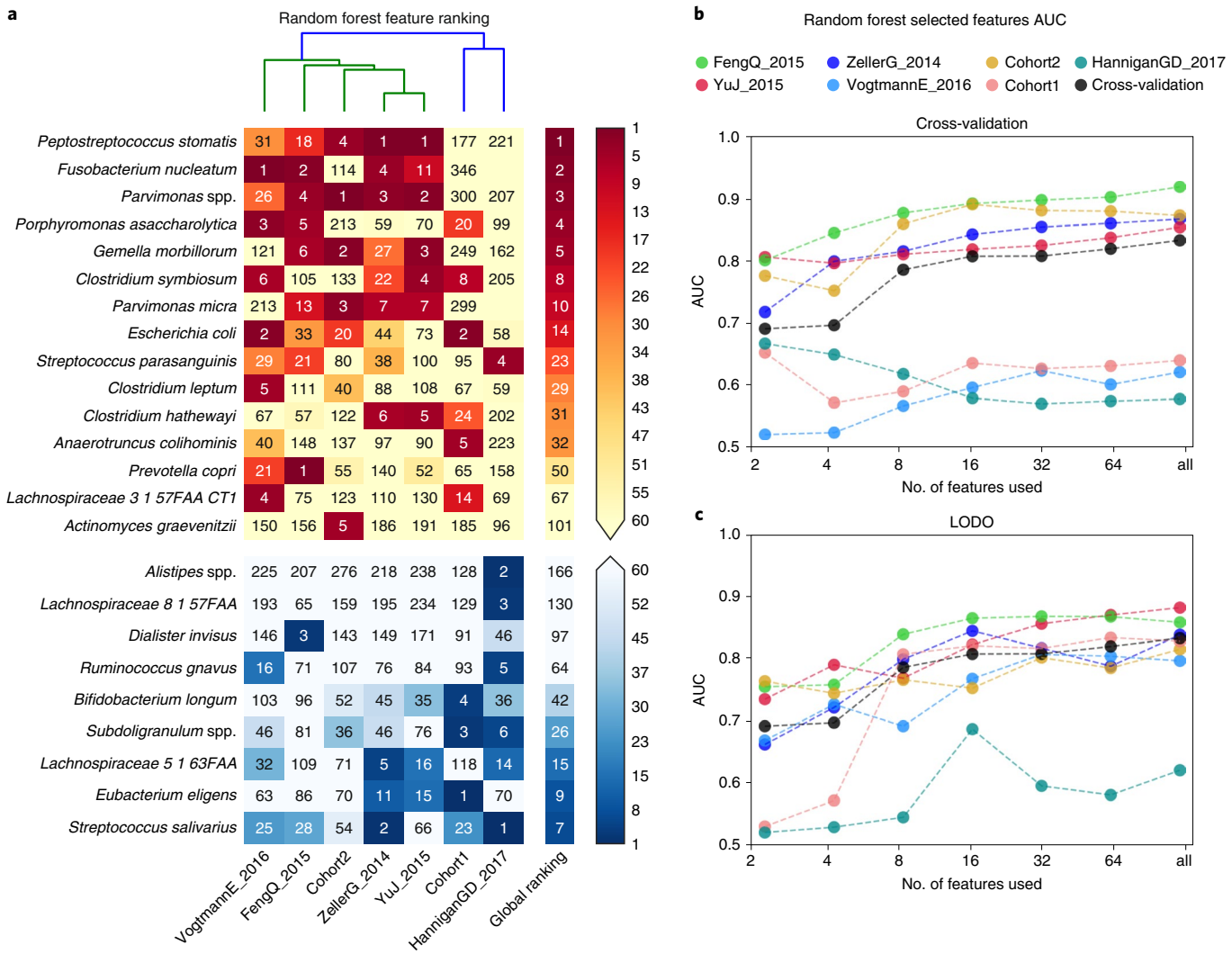


Fig. 3 | Ranking relevance of each species in the predictive models for each dataset and identification of a minimal microbial signature for CRC detection. a, The importance of each species for the cross-validation prediction performance in each dataset estimated using the internal random forest scores. Only species appearing in the five top-ranking features in at least one dataset are reported. **b**,**c**, Prediction performances at increasing number of microbial species obtained by retraining the random forest classifier on the top-ranking features identified with a first random forest model training in a cross-validation (**b**) and LODO setting (**c**). The rankings are obtained excluding the testing datasets to avoid overfitting.

highly predictive microbial features by exploiting the internal feature ranking of the random forest classifier[10]. We found that *P. stomatis* was the species with the highest average rank. As expected, other CRC-associated species including *F. nucleatum*, *Parvimonas* spp., *P. asaccharolytica*, *G. morbillorum*, *C. symbiosum* and *P. micra* were also crucial to prediction accuracy (Fig. 3a), with the seven top-ranked species for CRC detection amongst those with the largest effect sizes in the meta-analysis. Very few control-associated species were ranked high in the learning models, further highlighting that successful discrimination is achieved by CRC-specific rather than control-specific microbial features.

To evaluate how many microbial species or gene families are necessary to achieve prediction scores comparable to those obtained using the full set of features, we computed AUC values at increasing numbers of features. Feature ranking was performed internally to each training fold to avoid overfitting. By applying this approach to all datasets (Fig. 3b,c), we found that using as few as 16 species achieved cross-validation AUC >0.8 for the majority of the datasets, with little improvement from using all remaining species (2% average improvement in AUC value). We also found that using only 64 gene families achieved prediction values >0.8 for the same datasets, and that using all 8,192 gene families improved AUC only slightly (2% improvement; Extended Data Fig. 8). Therefore, these results suggest that a stool-based diagnostic test using genetic markers targeting a limited number of microbial species or genes would serve as a promising clinical tool.

Microbiome signatures for adenomas are only partially predictive. We assessed the ability to discriminate adenomas from controls or carcinomas, using 27 newly sequenced adenoma-associated samples and 116 adenoma-associated samples from available studies (Table 1). Adenomas could be distinguished from patients with CRC with lower accuracy than controls (average AUC = 0.69 versus 0.79; Extended Data Fig. 6e,f), and there are only eight species that differentiate patients with adenomas from patients with carcinomas in the meta-analysis (FDR < 0.05). Seven of these eight biomarkers are in common with the comparison between patients with carcinomas and healthy individuals, and the LODO approach did not improve discrimination of adenomas from CRC (average AUC = 0.68). Moreover, we found that no dataset could accurately

predict adenomas from control samples (maximum AUC = 0.58), even when using a LODO approach (average AUC = 0.54). In the meta-analysis, none of the species were significantly different when contrasting samples from patients with adenomas and healthy controls. These results reinforce previous findings[18,19] that the adenoma-associated stool microbiome closely resembles that of the healthy gut.

Increased abundance of choline trimethylamine-lyase-encoding genes in CRC. Microbiome-derived metabolites, and specifically polyamines, have been implicated in carcinogenesis in both animal models and humans[44]. We chose to focus on trimethylamine (TMA), an amine produced by bacteria from choline and carnitine, because it has been shown to play a role in complex diseases such as atherosclerosis and primary sclerosing cholangitis[9,47]. Since dietary components have been linked with CRC risk[5,6], we hypothesized that the TMA-producing potential of the human gut microbiome could also be associated with CRC[48]. To test this hypothesis, we considered the genes belonging to the main TMA-synthesis pathways to reconstruct and quantify the presence of such genes in the CRC-associated metagenomes. The main genes associated with TMA synthesis are those encoding choline TMA-lyase (*cutC*), L-carnitine dioxygenase (*yeaW*) and the L-carnitine/gamma-butyrobetaine antiporter (*caiT*), and we identified these in 923, 5,185 and 5,709 available bacterial genomes, respectively.

Screening the seven CRC-associated metagenomic datasets, we found that only one of these had a significant increase of *caiT* in CRC samples compared to controls, whereas no significant differences were detected for *yeaW* (Extended Data Fig. 9a). However, we found increased abundance of *cutC* in CRC samples compared to controls in all seven datasets ($P < 0.05$ by Wilcoxon rank-sum test on reads per kilobase million (RPKM) abundances for five datasets; Fig. 4a). Meta-analysis indicated an overall strong association with no evidence of heterogeneity ($P = 0.001$, $\mu = 0.27$, 95% confidence interval (0.1, 0.42), $I^2 = 4.2\%$, Q-test = 0.65; Fig. 4b). We also analyzed the abundance of the gene encoding the choline TMA-lyase-activating enzyme (*cutD*), finding a significant increase in CRC (meta-analysis $P = 0.001$, $\mu = 0.32$, 95% confidence interval (0.16, 0.47), $I^2 = 0\%$, Q-test = 0.96; Extended Data Fig. 9b,c). These results indicate that TMA production might happen preferentially via choline degradation, and not via carnitine, and could substantially affect the amounts of TMA and trimethylamine oxide (TMAO) present in an individual[49]. Intermediate levels of *cutC* in adenomas (Fig. 4a) are further suggestive of a TMA action along the adenoma–carcinoma axis. We validated the increased *cutC* gene abundance in CRC by quantitative PCR (qPCR)[50] on a subset of samples from Cohort1 with sufficient DNA left after sequencing, and confirmed the metagenomic findings (one-tailed Wilcoxon signed-rank test $P = 0.024$; Fig. 4d). Further quantification of *cutC* transcript abundance from the co-extracted RNA in the same dataset also pointed to an overexpression of this gene in CRC ($P = 0.035$; Fig. 4e).

We further explored the role of *cutC* in the gut microbiome by reconstructing sample-specific sequence variants using a reference-aided targeted assembly approach (see Methods). We found a large sequence divergence for the gene encoding this enzyme that is known to occur in single copies in the genomes[50], and we identified four main sequence variants that are associated with the taxonomic structure (Fig. 4c and Extended Data Figs. 9d,e and 10a,b). Interestingly, the most prevalent (46.5%) *cutC* sequence type (>95% identity over the full lenght of the gene) belonged to an unknown species that was only recently assembled from metagenomics[51] and assigned to species-level genome bin ID 3957. This candidate species comprises 56 metagenomically assembled species[51] and is placed within the Lachnospiraceae family, but the missing genus assignment confirms that several microbes remain undercharacterized in the human microbiome. This *cutC* variant was associated with non-CRC samples (odds ratio 0.38, 95% confidence interval (0.25, 0.57), $P = 0.0001$, Fisher test), whereas *cutC* sequence types mostly belonging to *Hungatella hathewayi* and *Clostridium asparagiforme* (Firmicutes) were significantly CRC-associated (odds ratio 2.14, 95% confidence interval (1.29, 3.56), $P = 0.004$, Fisher test), as were sequence types belonging to *Klebsiella oxytoca* and *Escherichia coli* (odds ratio 1.85, 95% confidence interval (1.13, 3.00), $P = 0.02$, Fisher test; Fig. 4b). Altogether, these findings highlight that sequence variants of *cutC* can be strongly associated with disease, potentially because of corresponding differences in the efficacy of choline degradation and TMA production.

Additional independent validation of predictive models. To further validate our meta-analysis results, we considered two additional independent metagenomic cohorts from Germany[30] (Validation Cohort1) and Japan (Validation Cohort2) comprising a total of 100 patients with CRC and 105 controls (see Methods). The metagenomic predictive model was confirmed to be highly accurate on these new cohorts (Fig. 5a), with an AUC of 0.90 and 0.81 for the German and Japanese cohorts, when using the species-level taxonomic abundance model. Species newly associated with the CRC microbiome, such as *S. tigurinus* and *S. dysgalactiae,* were confirmed as having higher prevalence in CRC than in controls in the two validation datasets (blocked Wilcoxon test[52] $P = 0.049$ and $P = 0.011$ for *S. tigurinus* and *S. dysgalactiae*, respectively). Enrichment in the CRC-associated microbiome of these two species was confirmed also by the analysis of additional metagenomic datasets of inflammatory bowel disease[53] and type 2 diabetes[54,55], in which the prevalence of *S. tigurinus* was always below 10% in both cases and controls, whereas *S. dysgalactiae* was never detected in these additional datasets. We also confirmed species richness to be significantly higher in CRC ($P = 0.0005$ for both validation datasets after rarefaction at the tenth percentile; Fig. 5b) as well as richness of oral microbial species in the rarefied samples (blocked Wilcoxon test[52] $P = 0.003$), and the abundance of *cutC* in CRC ($P < 1 \times 10^{-6}$).

CRC specificity of microbiome predictive models. We performed additional experiments to validate the discriminative power of the above microbial signatures specifically for CRC and not for other potentially microbiome-linked disease conditions. To this end, we first considered 13 additional fecal samples sequenced from patients who underwent colonoscopy in our Cohort1 and that were originally discarded because the final diagnosis pointed at diseases other than adenoma or carcinoma, such as ulcerative colitis, Crohn's disease, uncategorized colitis and diverticular diseases. These were distinguishable from CRC samples based on the taxonomic model (0.78 cross-validation AUC, 0.80 AUC using only 16 species), and only slightly decreased the AUC of the model trained on all the other datasets when these were added to the non-disease (that is, healthy) category (from 0.83 to 0.79 in AUC). We then expanded this analysis to diseases for which at least two distinct large metagenomic datasets are available in the public domain, and this includes ulcerative colitis and Crohn's disease[53,56] as well as non-gastrointestinal diseases such as type 2 diabetes[54,55]. For this purpose we added samples randomly drawn from each of the case and control conditions of these additional disease cohorts to the control class of the new validation cohort, and recorded the variations in AUC when attempting to predict CRC (see Methods). By comparing the AUCs obtained when adding non-CRC external cases and when adding the corresponding external controls, we found for both validation cohorts a small decrease in prediction accuracy for both ulcerative colitis (3 and 4% for Validation Cohort1 and Validation Cohort2, respectively; Fig. 5c) and Crohn's disease (5 and 9% for Validation Cohort1 and Validation Cohort2, respectively; Fig. 5c), pointing to a limited effect on the CRC model of samples from these two diseases. For type 2 diabetes we observed an increase in the predictive

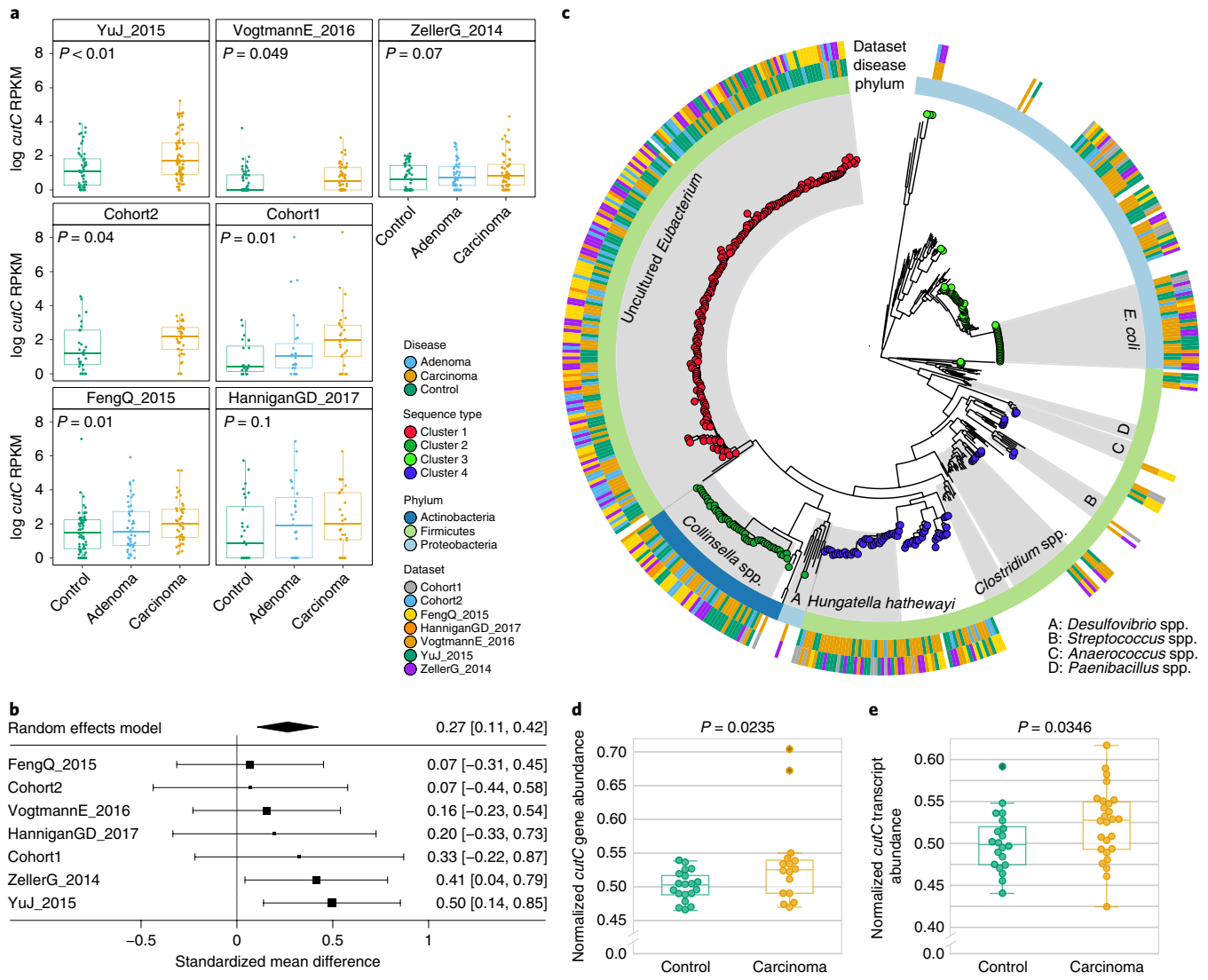


Fig. 4 | Choline TMA-lyase gene *cutC* and its genetic variants are strong biomarkers for CRC-associated stool samples. a, Distribution of RKPM abundances obtained using ShortBRED for the choline TMA-lyase enzyme gene *cutC*. *P* values were computed by two-tailed Wilcoxon signed-rank tests comparing values between controls and carcinomas for each dataset. **b**, Forest plot reporting effect sizes calculated using a meta-analysis of standardized mean differences and a random effects model on *cutC* RPKM abundances between carcinomas and controls. **c**, A phylogenetic tree of sample-specific *cutC* sequence variants identified four main sequence variants. Tips with no circles represent *cutC* sequence variants from genomes absent from the datasets. Taxonomy was assigned based on mapping against existing *cutC* sequences (criteria of >80% breadth of coverage, >97% identity and minimum 2,000-nt alignment length). **d**,**e**, qPCR validation of *cutC* gene abundance (**d**) and *cutC* transcript abundance (normalized by total 16S rRNA gene/transcript abundance) (**e**) on a subset of DNA samples from Cohort1. qPCR validation *P* values were obtained by one-tailed Wilcoxon signed-rank test. The lower and upper hinges of boxplots presented in the Figures correspond to the 25th and 75th percentiles, respectively. The midline is the median. The upper and lower whiskers extend from the hinges to the largest (or smallest) value no further than ×1.5 interquartile range from the hinge, defined as the distance between the 25th and 75th percentiles. Data beyond the end of the whiskers are plotted individually.

power in one dataset[54] and a decrease in the other[55] in both validation datasets, and the CRC model always remained highly predictive (AUC ≥ 0.80). Altogether, these results point to the existence of a clear microbiome signature of CRC that is distinct from other relevant diseases with a gastrointestinal component.

Relationship to currently available non-invasive clinical screening tests. To assess the potential of microbiome-based prediction models in comparison and in combination with currently used non-invasive clinical screening tests, we considered the fecal occult blood test (FOBT) and the Wif-1 methylation test available for 110 samples of the ZellerG_2014 cohort[19]. The LODO microbiome model tested on this dataset proved to be slightly superior to the FOBT at multiple combinations of specificity and sensitivity levels (Fig. 5d), and on a par with the Wif-1 methylation test. Considering the LODO model predictions and the FOBT together in the same test improves the sensitivity/specificity trade-off at high specificity levels when the integration is based on having at least one predictor positive, and at relatively lower specificity levels when requiring both predictors to be positive (Fig. 5d). Integrating the microbiome model with the Wif-1 methylation test resulted in similar performances, and the use of the reduced microbiome model with only 16 species generally improved the results (Fig. 5d). We thus provide evidence for the potential clinical value of microbiome predictive models, especially when considered together with other available non-invasive clinical tests.

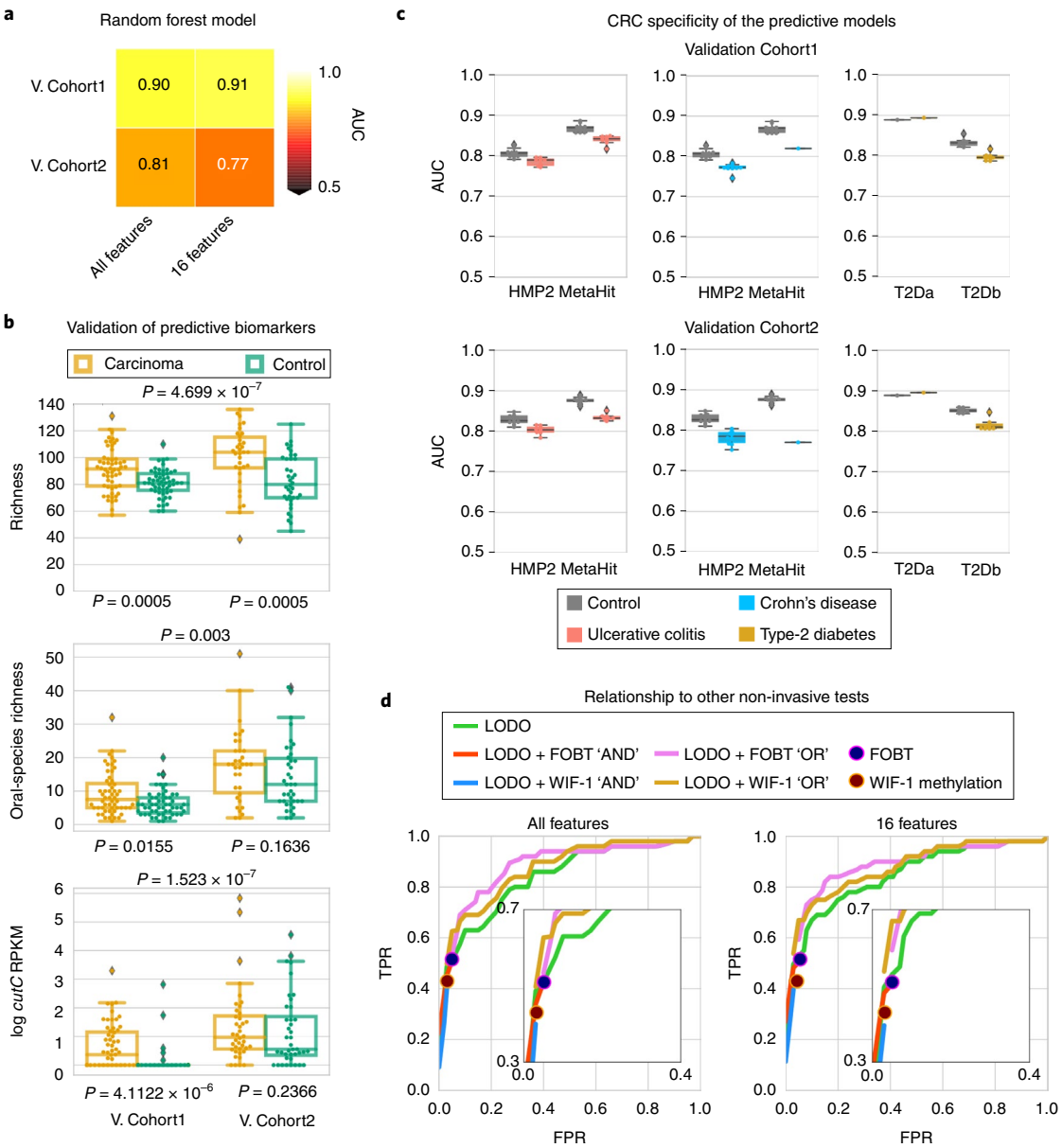


Fig. 5 | Clinical potential and validation of the predictive biomarkers. a, Prediction performance of the taxonomic models trained on the seven datasets of Table 1 and applied to the new validation cohorts (using all features and the top 16 features) confirmed the strong reproducibility of metagenomic models for CRC across cohorts when sufficiently large training cohorts are available. Random forest internal ranking of the 16 most predictive species was obtained on each training fold separately to avoid overfitting. **b**, Species richness, rarefied oral species richness and *cutC* gene abundance (RPKM) are confirmed as strong biomarkers of CRC in the validation datasets[30]. *P* values underlying the panels refer to one-tailed Wilcoxon signed-rank test; *P* values overlying the panels refer to one-sided permutation-based Wilcoxon–Mann–Whitney tests, blocked for cohort. **c**, Prediction performances as AUC values on the validation cohorts when adding an external set of case and control samples from metagenomic cohorts of diseases other than CRC (Crohn's disease, ulcerative colitis, type 2 diabetes (T2D)). The studies that were included are HMP[56], MetaHit[53], T2Da[54] and T2Db[55]. **d**, Assessment of the potential of microbiome-based prediction models in comparison, and in combination, with current non-invasive clinical screening tests. Models integrating our LODO machine learning approach with the FOBT or the Wif-1 methylation test are termed OR and AND, depending on whether only one or both is required to be positive for the combined test to be positive. FPR, false-positive rate; TPR, true-positive rate; V., validation. The lower and upper hinges of boxplots presented in the Figures correspond to the 25th and 75th percentiles, respectively. The midline is the median. The upper and lower whiskers extend from the hinges to the largest (or smallest) value no further than ×1.5 interquartile range from the hinge, defined as the distance between the 25th and 75th percentiles. Data beyond the end of the whiskers are plotted individually.

Discussion

In the present study, we comprehensively assessed the CRC-associated gut microbiome and its ability to distinguish newly diagnosed patients with CRC from tumor-free controls. Our study was performed across multiple datasets and populations, through a combined analysis of fecal CRC metagenomes from four previously unpublished cohorts and five publicly available datasets. Whereas direct specific host–microbe interactions have been shown to cause certain malignancies in both in vitro and in vivo animal models[11–13,57], and genotoxic determinants such as colibactin tend to be over-represented in the analyzed datasets[30], indirect metabolite-mediated mechanisms may be more important in the development

of carcinomas although causality relations need to be tested. In our analysis, we indeed found a reproducible panel of microbiome species (Fig. 1), whole-microbiome characteristics and strain-level biomarkers (Fig. 4) beyond the validated mechanisms of specific variants of *E. coli*[11,57] and *Bacteroides fragilis*[57]. We found that the gut microbiome in CRC has greater richness than controls, partially due to the presence of oral cavity-associated species rarely found in the healthy gut, challenging the widespread assumption that decreased alpha-diversity is generally associated with intestinal dysbiosis[58,59].

The identification of reproducible microbial biomarkers for CRC may enable the design of non-invasive diagnostic tools. We developed machine learning models able to distinguish patients with carcinomas from controls with an average performance of 0.84 AUC when validated on datasets excluded from the training of the model (Fig. 2a). Importantly, these performances are quite independent of specific methodological choices given that complementary investigations[30] using different metagenomic profilers and machine learning approaches achieved very similar results. Further increase in prediction performance can be achieved using larger datasets ($n > 1{,}000$) rather than different methodologies (Figs. 2c,d and Extended Data Fig. 7), and the combination of a microbiome model with other clinical tests and patient risk factors could substantially improve this diagnostic accuracy (Fig. 5d). Current clinical pre-colonoscopy screening tests (for example, FOBT, Wif-1) remain cheaper, but microbiome-based CRC prediction models enable a very high diagnostic potential that increases with the number of microbes or microbial genes used, with single biomarkers being much inferior to multi-feature diagnostic models. However, near maximal accuracy was achieved with as few as 15–25 microbes (Fig. 3b,c) or a few hundred genes (Extended Data Fig. 8), potentially enabling inexpensive clinical microbiological tests to be performed on stool samples. Prospective studies of these biomarkers are needed to establish whether they can identify individuals at elevated risk of CRC and provide the possibility of disease prevention.

The diversity and subject-specificity of the human gut microbiome is not yet fully uncovered, with many microbial genes having unknown function and with strain-level diversity that is missed by many current analysis pipelines[51]. Large-scale shotgun metagenomics can begin to overcome these limitations, as shown here by the identification of a link between CRC and the microbial pathway producing TMA from choline[49]. The gene encoding for the key enzyme for this pathway, cutC choline TMA-lyase, is both more abundant and expressed in the gut microbiome of patients with carcinomas, with specific variants of *cutC* characterizing controls, adenomas and carcinomas (Fig. 4). TMA-producing choline lyases have been found to be associated with atherosclerosis[9], and higher plasma TMA oxide and choline levels have been reported to be correlated with CRC risk[60,61]. We highlight the importance of strain-level gene resolution in understanding any potential carcinogenic role of *cutC*. CRC-associated variants mostly originated from *H. hathewayi*, *C. asparagiforme*, *K. oxytoca* and *E. coli*, whereas no significant enrichment was detected for a *cutC* variant carried by an unexplored, recently discovered candidate species in the family Lachnospiraceae[51]. Thus, genetic variants in key microbial genes involved in choline-induced TMA production by the gut microbiome are a plausible and potential mechanism for colorectal carcinogenesis. Other partially diet-dependent microbiome factors can contribute to the promotion of carcinogenesis, and we found in our parallel work that genes for secondary bile acid conversion are consistently enriched in CRC-associated microbiomes[30]. Further work is needed to establish the changes in protein structure and function associated with the genetic variants of the diet-related microbial genes found here to be enriched in the CRC microbiome.

Analysis of cancer cohorts that are heterogeneous for geography, ethnicity and lifestyle presents a distinct opportunity for studying the cancer-associated microbiome. By combining multiple small cohorts of potentially low generalizability, it is possible to obtain better representation of the spectrum of cancer cases and controls. With appropriate methodology, artifactual findings due to batch effects present in any individual dataset can be avoided. The use of large, diverse training sets enables the creation of more accurate diagnostic models, and the availability of independent validation datasets enables more realistic estimation of that accuracy. Future shotgun metagenomic studies of the intestinal mucosa-associated microbiome, which are currently infeasible due to excessive human DNA contamination[29], will be important to further refine the list of CRC-associated gut microbes. Nevertheless, this study identifies highly reproducible microbial CRC biomarkers and points to the potential for non-invasive microbial diagnostic tests to supplement existing screening.

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/s41591-019-0405-7.

Received: 29 May 2018; Accepted: 20 February 2019;
Published online: 1 April 2019

References

1. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. Siegel, R., Desantis, C. & Jemal, A. Colorectal cancer statistics, 2014. *CA Cancer J. Clin.* **64**, 104–117 (2014).
3. Frank, C., Sundquist, J., Yu, H., Hemminki, A. & Hemminki, K. Concordant and discordant familial cancer: familial risks, proportions and population impact. *Int. J. Cancer* **140**, 1510–1516 (2017).
4. Foulkes, W. D. Inherited susceptibility to common cancers. *N. Engl. J. Med.* **359**, 2143–2153 (2008).
5. Johnson, C. M. et al. Meta-analyses of colorectal cancer risk factors. *Cancer Causes Control* **24**, 1207–1222 (2013).
6. Huxley, R. R. et al. The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence. *Int. J. Cancer* **125**, 171–180 (2009).
7. Schmidt, T. S. B., Raes, J. & Bork, P. The human gut microbiome: from association to modulation. *Cell* **172**, 1198–1215 (2018).
8. Thomas, R. M. & Jobin, C. The microbiome and cancer: is the 'oncobiome' mirage real? *Trends Cancer Res.* **1**, 24–35 (2015).
9. Jie, Z. et al. The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.* **8**, 845 (2017).
10. 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).
11. Cougnoux, A. et al. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* **63**, 1932–1942 (2014).
12. 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).
13. Chung, L. et al. *Bacteroides fragilis* toxin coordinates a pro-carcinogenic inflammatory cascade via targeting of colonic epithelial cells. *Cell Host Microbe* **23**, 203–214.e5 (2018).
14. Kostic, A. D. et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
15. Kostic, A. D. et al. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).
16. 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).
17. 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).
18. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
19. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
20. Vogtmann, E. et al. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS One* **11**, e0155362 (2016).

21. Baxter, N. T., Ruffin, M. T., Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 37 (2016).
22. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T. 4th & Schloss, P. D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res.* **7**, 1112–1121 (2014).
23. Drewes, J. L. et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microbiomes* **3**, 34 (2017).
24. Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The madness of microbiome: attempting to find consensus 'best practice' for 16S microbiome studies. *Appl. Environ. Microbiol.* **84**, e02627–17 (2018).
25. Segata, N. On the road to strain-resolved comparative metagenomics. *mSystems* **3**, e00190–17 (2018).
26. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
27. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023 (2017).
28. Dai, Z. et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* **6**, 70 (2018).
29. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
30. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* https://doi.org/10.1038/s41591-019-0406-6 (2019).
31. 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).
32. Thomas, A. M. et al. Tissue-associated bacterial alterations in rectal carcinoma patients revealed by 16S rRNA community profiling. *Front. Cell. Infect. Microbiol.* **6**, 179 (2016).
33. Gao, Z., Guo, B., Gao, R., Zhu, Q. & Qin, H. Microbiota disbiosis is associated with colorectal cancer. *Front. Microbiol.* **6**, 20 (2015).
34. Ahn, J. et al. Human gut microbiome and risk for colorectal cancer. *J. Natl Cancer Inst.* **105**, 1907–1911 (2013).
35. Flemer, B. et al. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454–1463 (2017).
36. Brito, I. L. et al. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).
37. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
38. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407 (2016).
39. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
40. Xie, Y.-H. et al. Fecal *Clostridium symbiosum* for noninvasive detection of early and advanced colorectal cancer: test and validation studies. *EBioMedicine* **25**, 32–40 (2017).
41. Boleij, A., van Gelder, M. M. H. J., Swinkels, D. W. & Tjalsma, H. Clinical Importance of *Streptococcus gallolyticus* infection among colorectal cancer patients: systematic review and meta-analysis. *Clin. Infect. Dis.* **53**, 870–878 (2011).
42. Fijan, S. Microorganisms with claimed probiotic properties: an overview of recent literature. *Int. J. Environ. Res. Public Health* **11**, 4745–4767 (2014).
43. Apweiler, R. et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119 (2004).
44. Gerner, E. W. & Meyskens, F. L. Jr Polyamines and cancer: old molecules, new understanding. *Nat. Rev. Cancer* **4**, 781–792 (2004).
45. Costea, P. I. et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
46. Riester, M. et al. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl Cancer Inst.* **106**, dju048 (2014).
47. Kummen, M. et al. Elevated trimethylamine-N-oxide (TMAO) is associated with poor prognosis in primary sclerosing cholangitis patients with normal liver function. *United European Gastroenterol. J.* **5**, 532–541 (2017).
48. Oellgaard, J., Winther, S. A., Hansen, T. S., Rossing, P. & von Scholten, B. J. Trimethylamine N-oxide (TMAO) as a new potential therapeutic target for insulin resistance and cancer. *Curr. Pharm. Des.* **23**, 3699–3712 (2017).
49. Kalnins, G. et al. Structure and function of CutC choline lyase from human microbiota bacterium *Klebsiella pneumoniae*. *J. Biol. Chem.* **290**, 21732–21740 (2015).
50. Rath, S., Heidrich, B., Pieper, D. H. & Vital, M. Uncovering the trimethylamine-producing bacteria of the human gut microbiota. *Microbiome* **5**, 54 (2017).
51. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 1–14 (2019).
52. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. A lego system for conditional inference. *Am. Stat.* **60**, 257–263 (2006).
53. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
54. Karlsson, F. H. et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
55. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
56. Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276–289 (2014).
57. Dejea, C. M. et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592–597 (2018).
58. Manichanh, C. et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205–211 (2006).
59. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
60. Bae, S. et al. Plasma choline metabolites and colorectal cancer risk in the Women's Health Initiative Observational Study. *Cancer Res.* **74**, 7442–7452 (2014).
61. Xu, R., Wang, Q. & Li, L. A genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine N-oxide (TMAO), a gut microbial metabolite of dietary meat and fat. *BMC Genomics* **16**(Suppl 7), S4 (2015).

Acknowledgements

We thank the members of the Segata, Naccarati and Waldron groups for insightful discussions, all the volunteers enrolled in the study, the NGS facility at the University of Trento for performing the metagenomic sequencing, and the HPC facility at the University of Trento for supporting the computational experiments. This work was primarily supported by Lega Italiana per La Lotta contro i Tumori to N.S., F.C. and A.N., and by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant No. 16/23527-2) to A.M.T. This work was also partially supported by the Conselho Nacional de Pesquisa e Desenvolvimento (CNPq, Brazil) to J.C.S. and E.D.-N.; by FAPESP (grant No. 14/26897-0); by Associação Beneficente Alzira Denise Hertzog Silva (ABADHS, Brazil) and PRONON/SIPAR to E.D.-N.; by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) (Finance Code No. 001 to J.C.S.; by the Italian Institute for Genomic Medicine (IIGM) and Compagnia di San Paolo Torino to A.N, A.F., B.P. and S.T.; by Fondazione Umberto Veronesi 'Post-doctoral Fellowship Years 2014, 2015, 2016, 2017 and 2018' to B.P. and S.T.; by the Grant Agency of the Czech Republic (grant No. 17-16857S) to A.N.; by Fondazione Umberto Veronesi (grant No. FUV-14-SG-GANDINI) to S.G.; by the European Union H2020 Marie Curie grant (No. 707345) to E.P.; by the European Research Council (ERC-STG project MetaPG) to N.S.; by MIUR 'Futuro in Ricerca' (grant No. RBFR13EWWI_001) to N.S.; by the People Programme (Marie Curie Actions) of the European Union FP7 and H2020 to N.S.; and by the National Cancer Institute (grant No. U24CA180996) and National Institute of Allergy and Infectious Diseases (grant No. 1R21AI121784-01) of the National Institutes of Health to L.W. B.P. is the recipient of a Fulbright Research Scholarship (year 2018). We acknowledge funding from EMBL, DKFZ, 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 (No. ERC-AdG-669830) to P.B.; 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 ll to C.M.U.); the BMBF (the de.NBI network, grant No. 031A537B) to P.B.; the ERA-NET TRANSCAN project (No. 01KT1503) to C.M.U.; and the Helmut Horten Foundation (to S.S.). For Validation Cohort2, funding was provided by grants from 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 (AMED, grant No. JP18ek0109187); JST (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; the University of Tokyo; the Takeda Science Foundation; and the Suzuken Memorial Foundation.

Author contributions

N.S., A.M.T., L.W. and A.N conceived the study. N.S. supervised the study. C.P., S.G., D.S., S.T., A.F., G.G., M.T., B.P, M.R. and A.N. organized the clinical study, recruited patients and collected samples. F. Armanini generated metagenomic data. A.M.T., P.M., F. Asnicar, E.P., M.Z., F.B., N.K. and G.F. collected and analyzed the metagenomic data. A.M.T., P.M., F. Asnicar, E.P., M.Z., G.F., J.W., G.Z. and L.W. performed machine learning and statistical analyses. F. Armanini, S.T., S. Manara, A.T., B.P. and A.N. performed validation experiments. S. Mizutani., H.S., S. Shiba, T.S., S.Y., T.Y., J.W., P.S.-K, C.M.U.,

H.B., M.A., P.B. and G.Z. provided additional validation data. A.M.T., P.M., L.W. and N.S. designed and produced the figures. A.M.T., P.M. and N.S. wrote the manuscript with contributions from S. Manara, F.C., E.D.-N., J.C.S., M.R., L.W. and A.N. All authors discussed and approved the manuscript.

Competing interests

P.B., G.Z., A.Y.V. and S.S. are named inventors on a patent (EP2955232A1: Method for diagnosing colorectal cancer based on analyzing the gut microbiome). All other authors declare no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0405-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0405-7.

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

Correspondence and requests for materials should be addressed to N.S.

Methods

Italian cohorts of patients with CRC, adenomas and controls. The two clinical studies performed here were approved by the relevant ethics committees (Cohort1: Ethics Committee of Azienda Ospedaliera 'SS. Antonio e Biagio e C. Arrigo' of Alessandria, Italy, protocol No. Colorectal_miRNA_CEC2014; and Cohort2: Ethics Committee of European Institute of Oncology of Milan, Italy, protocol No. R107/14-IEO 118). Informed consent was obtained from all participants.

For Cohort1, samples were collected from patients at the Clinica S. Rita in Vercelli, Italy. Patients with hereditary CRC syndromes, with a previous history of CRC and with uncompleted or poorly cleaned colonoscopy were excluded from the study. Patients were recruited at initial diagnosis and had not received any treatment before fecal sample collection. Subjects reporting the use of antibiotics during the 6 months before the sample collection were excluded from the study. On the basis of colonoscopy results, recruited subjects were classified into three categories: (1) healthy subjects: individuals with colonoscopy negative for tumor, adenomas and other diseases; (2) patients with adenoma: individuals with colorectal adenoma(s); and (3) patients with CRC: individuals with newly diagnosed CRC. A total of 93 subjects were initially recruited, and the 80 that passed quality control (see below) were divided into 29 patients with CRC, 27 adenomas and 24 controls. An additional 13 subjects affected by inflammatory gastrointestinal tract diseases (ulcerative and Crohn's colitis, diverticular diseases) were recruited and fecal samples were subsequently used as a part of the final validation. Stool was collected in stool nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp.) and returned before performing the colonoscopy. Aliquots of stool samples were stored at −80 °C until use. DNA was extracted from aliquots of fecal samples using the Qiamp DNA stool kit (Qiagen) following the manufacturer's instructions. Total RNA from feces was extracted using the Stool Total RNA Purification Kit (Norgen Biotek Corp.) following the manufacturer's instructions.

For Cohort2, a total of 60 subjects were recruited at the European Oncology Institute in Milan, Italy and were divided into 32 patients with CRC and 28 controls. Controls, matched for age (±5 years) and season when blood withdrawn (±2 years), were recruited among subjects who underwent recent colonoscopy and had negative or no other relevant gastrointestinal disorders. Subjects reporting the use of antibiotics in the 6 months before sample collection were excluded. Fecal samples were collected from healthy subjects and patients (before surgery, or any other cancer treatment) and directly frozen at −80 °C in resuspension buffer (TES buffer: 50 mM Tris-HCL, 10 mM NaCl, 10 mM EDTA, pH 7.5) and kept in liquid nitrogen until DNA extraction. DNA was extracted from fecal samples with the GNOME DNA isolation kit (MP Biochemicals).

Sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina), following the manufacturer's guidelines. Sequencing was performed on a HiSeq2500 (Illumina) at the internal sequencing facility at University of Trento, Trento, Italy.

Public metagenomic cohorts of patients with CRC, adenomas and controls. We downloaded five public fecal shotgun CRC datasets covering samples from six different countries, totaling 313 patients with CRC, 143 adenomas and 308 controls (Table 1), and now available in curatedMetagenomicData[27]. We manually curated metadata tables for the public cohorts according to the curatedMetagenomicData[27] R package grammatical rules. The metadata table includes ten fields (sampleID, subjectID, body_site, country, sequencing_platform, PMID, number_reads, number_bases, minimum_read_length, median_read_length) that are mandatory for all datasets, in addition to other fields that are dataset-specific.

Description of the two validation cohorts. We considered an additional set of samples from two independent cohorts that were not available at the time we performed the meta-analysis on the other seven datasets, and we therefore used these as validation cohorts. Validation Cohort1 consisted of 60 CRC metagenomes collected in Germany after colonoscopy and 65 sex- and age-matched healthy controls, and is described in depth in the study accompanying this work[30]. Shotgun metagenomic sequencing was performed by Illumina HiSeq 2000/2500/4000 (Illumina) platforms at the Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany. Validation Cohort2 consisted of 40 CRC samples and 40 controls from a Japanese cohort from Tokyo. DNA was extracted for Validation Cohort2 from frozen fecal samples by bead-beating using the GNOME DNA Isolation Kit (MP Biomedicals), and DNA quality was assessed with an Agilent 4200 TapeStation (Agilent Technologies). Sequencing libraries were generated with a Nextera XT DNA Sample Prep Kit (Illumina), and shotgun metagenomics of fecal samples was carried out on the HiSeq 2500 platform (Illumina) at a targeted depth of 5.0 Gb (150 base pair (bp) paired-end reads).

The samples and clinical information used from both validation cohorts in this study were obtained under conditions of informed consent and with approval of the institutional review boards of each participating institute.

Public metagenomic cohorts of patients without CRC. We used the curatedMetagenomicData[27] resource to retrieve taxonomical and functional potential profiles as well as the metadata of three public cohorts: NielsenHB_2014 (ref. [53]), comprising 21 patients with Crohn's disease, 127 patients with ulcerative colitis and 248 controls; KarlssonFH_2013 (ref. [54]), comprising 53 patients with type 2 diabetes and 43 controls; QinJ_2012 (ref. [55]), comprising 172 patients with type 2 diabetes and 174 controls; and we downloaded 1,339 metagenomes from the Human Microbiome Consortium phase-2 cohort[56], comprising 598 patients with Crohn's disease, 375 patients with ulcerative colitis and 365 controls.

Sequence preprocessing and taxonomic and functional profiling. Fecal metagenomic shotgun sequences obtained from the Italian cohorts were subjected to a preprocessing pipeline whereby sequences were quality filtered using trim_galore (parameters: --nextera --stringency 5 --length 75 --quality 20 --max_n 2 --trim-n), discarding all reads of quality less than 20 and shorter than 75 nucleotides. Filtered reads were then aligned to the human genome (hg19) and the PhiX genome for human and contaminant DNA removal using bowtie2 (ref. [62]). Thirteen samples having less than 2 Gb of host-decontaminated DNA were excluded from the study.

We used MetaPhlAn2 (ref. [63]) for quantitative profiling of the taxonomic composition of the microbial communities of all metagenomic samples, whereas HUMANn2 (ref. [64]) was used to profile pathway and gene-family abundances. The profiles generated for the six public cohorts, along with their metadata and the two newly sequenced cohorts, are available through the curatedMetagenomicData R package[27]. Oral species were defined in this work by analyzing the 463 oral samples from the Human Microbiome Project dataset[37] and the 140 saliva samples from ref. [36]. Specifically, all species with >0.1% abundance and >5% prevalence were deemed to be of oral origin. For *F. nucleatum* marker analysis, we extracted MetaPhlAn2 clade-specific markers from each sample sam file and considered a marker to be present if the coverage was greater than zero.

The random forest-based machine learning approach. Our machine learning analyses exploited four types of microbiome quantitative profiles: taxonomic species-level relative abundances and marker presence or absence patterns inferred by MetaPhlAn2 (ref. [63]) and gene-family and pathway-relative abundances estimated by HUMAnN2 (ref. [64]).

All machine learning experiments used random forest[65], as this algorithm has been shown to outperform, on average, other learning tools for microbiome data[10]. The code generating the analyses and the Figures is available at https://bitbucket.org/CibioCM/multidataset_machinelearning/src/, and is based on MetAML[10] with the random forest implementation taken from Scikit-Learn v.0.19.0 (ref. [66]). We used an ensemble of 1,000 estimator trees and Shannon entropy to evaluate the quality of a split at each node of a tree. The two hyperparameters for the minimum number of samples per leaf and for the number of features per tree were set, as indicated elsewhere[67], to 5 and 30%, respectively. For the marker presence/absence profiles we used a number of features equal to the square root of the total number of features, and this percentage was further decreased to 1.0 when using gene-family profiles as these have a substantially higher number of features (>2 million). The experiments ran on reduced sets of input features (Fig. 4 and Extended Data Fig. 8) avoided feature subsampling when fewer than 128 features were used (Extended Data Fig. 8).

Application and evaluation of the learning models. The inside-dataset prediction capability was measured through tenfold cross-validation, stratified so that each fold contained a balanced proportion of positive and negative cases. The procedure of forming the folds and assessing the models was repeated 20 times. The final result is therefore an average over 200 validation folds. In the cross-study validation, datasets were considered two by two: one is used for training the model, the other to validate.

The LODO approach consists of training the model on the pooled samples from all cohorts except the one used for model testing. This mimics the scenario in which all available samples from multiple cohorts are used to predict CRC-positive samples in a newly established cohort. As part of the meta-analysis, we iterated along all the cohorts, performing a LODO validation on each set of samples (Fig. 2).

Additional validation experiments on independent datasets and other diseases. We built a validation LODO model trained on MetaPhlAn2 taxonomic abundances from the previously described set of seven cohorts and applied it to the independent validation cohorts. To test the performance of the model when challenged with other diseases, we selected four metagenomic cohorts[53–56] covering three non-CRC diseases (ulcerative colitis, Crohn's disease and type 2 diabetes) and we used these for further experiments. For each disease in each dataset, we randomly drew 60 samples from the control class as well as 60 samples from the cases and added them to each validation dataset in turn, labeled as controls. The random selection was repeated ten times, and the validation AUC was computed on the model's prediction accordingly. The rationale is to observe the decrease in AUC when the external cases are added to the controls of the validation cohort with respect to the addition of healthy controls.

Specificity of the prediction model was also assessed by the addition of 13 inflammatory bowel disease samples to Cohort1: we used these 13 samples either as controls for Cohort1 or added them to the original controls; we performed a cross-validation and a LODO on Cohort1 (no validation cohorts in the training) using MetaPhlAn2 microbial species.

To assess the prediction ability of our random forest approach with respect to more traditional non-invasive tests such as the FOBT and the Wif-1 methylation tests, we recorded the true-positive rate (sensitivity) and false-positive rate (1 – specificity) for a subset of the ZellerG_2014 cohort according to these two tests, and 100 positive detection thresholds in the case of random forest models. We then combined the random forest approach with the two tests in turn, first assigning the positive class when both predictors are positive ('AND' model) and secondly when just one predictor is positive ('OR' model).

Statistical analysis. Univariate analyses on a per-dataset basis was performed using linear discriminant analysis effect size (LEfSe)[39] to identify features that were statistically different among groups and to estimate their effect size. Analysis of Composition of Microbes was also applied[68], but showed reduced power on our datasets (for example, it identified *F. nucleatum* as a biomarker in only one dataset) probably due to the low relative abundance of CRC biomarkers that are thus only minimally affected by the problem of compositionality. For these reasons, we chose to use LEfSe for the univariate analysis and focused on those biomarkers with the highest effect size. To overcome the limitations of univariate statistics, we performed multivariate analysis using linear models fitted to the data using the limma R package[69], and potential confounders such as age, sex and BMI were included in the models. For the meta-analysis on taxonomic and functional profiles, we converted relative abundances to arcsine-square root-transformed proportions and used the *escalc* function from the R metafor package that employs Cohen's standardized mean difference statistic to calculate random effects model estimates. We quantified study heterogeneity using the I^2 estimate (percentage of variation reflecting true heterogeneity), as well as Cochran's *Q*-test to assess statistically significant heterogeneity. *P* values obtained from the random effects models were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure, with corrected $P < 0.05$ considered statistically significant. Cluster analysis was conducted by calculating distance matrices from phylogenetic trees using the APE R package, clustering using partitioning around medoids and computing the prediction strength of clusters using the cluster R package. When validating differential species richness, oral species richness and increased abundance of the *cutC* gene, we also assessed significance through one-sided, permutation-based Wilcoxon–Mann–Whitney tests where we blocked for cohort[52], as implemented in the 'coin' R package.

Identification and quantification of the genes encoding TMA-producing enzymes. To obtain a more comprehensive database of choline TMA-lyase enzyme sequences, we downloaded amino acid sequences that matched the keywords *cutC* and *cutD* from UniProt90 (ref. [43]), mapped their identifiers to those of the European Molecular Biology Laboratory's coding sequences using UniParc and used the resulting DNA sequences to search, using BLASTn[70], all 48,902 Prokka[71] annotated genomes available in our repository[72]. Matching queries were filtered to include only alignments with >80% identity and length >1,000 nt for *cutC* and >800 nt for *cutD*, and an e-value $<1 \times 10^{-15}$. We used ShortBRED[73] to identify short seed sequences that were representative of the filtered queries using UniProt's UniRef100 database and quantified these in the metagenomes, normalizing by the number of RPKM. The pipeline was also applied to identify and quantify the L-carnitine/gamma-butyrobetaine antiporter (*caiT*) and the dioxygenase *yeaW*, responsible for producing TMA preferentially via carnitine degradation. To investigate differences in *cutC* sequence types, we clustered *cutC* sequences at 97% sequence identity using UCLUST[74] and aligned raw reads to the clustered *cutC* database using bowtie2 (ref. [62]). From the bam files we calculated the breadth and depth of each sequence and generated their corresponding consensus sequence using Samtools[75] and VCF utils[76]. We chose the representative *cutC* sequence for each sample as the one with the highest breadth or the highest depth, if there were multiple *cutC* sequences with the same breadth. We filtered representative *cutC* sequences from each sample to include only those of breadth >80%, aligned them using MAFFT[77] and built a phylogenetic tree using fastTree[78], which was refined with RAxML[79] and visualized using GraPhlAn[80].

Validation of *cutC* gene and transcript abundances by qPCR. Real-time qPCR was used to assess differences in *cutC* genes and transcripts between CRC samples and controls. We used a previously described protocol[50] that employs 16S rRNA abundances as an internal sample normalization. For first-strand cDNA synthesis, 400 ng of RNA templates were retrotranscribed using the High-capacity cDNA Reverse Transcription Kit with Random Primers (ThermoFisher Scientific), following the manufacturer's instructions. The *cutC* and 16S rRNA genes (and transcripts from complementary DNA (cDNA)) were amplified using degenerate primers and cycling conditions as described previously[50]. Briefly, reactions were performed in triplicate with 10 ng of template DNA or 30 ng of cDNA on the Rotor Gene Q (QIAGEN) using HOT FIREPol EvaGreen qPCR mix (SOLIS BIODYNE) at a final primer concentration of 0.5 µM (16S) or 0.75 µM (*cutC*). Cycling conditions were as follows: initial denaturation of 95 °C for 15 min, followed by 40 cycles of denaturing at 95 °C for 45 s, annealing at 57 °C (*cutC*) or 55 °C (16 S) for 45 s and an extension step of 72 °C for 45 s. Melting curves were subsequently performed for all reactions using the following program: 95 °C for 5 s, followed by 65 °C for 60 s and a final continuous reading step of seven acquisitions per second between 65 and 97 °C.

Quantification of the *cutC* gene by means of the qPCR protocol was applied to 44 samples belonging to Cohort1 for which sufficient DNA was available. Samples for which either the *cutC* or the 16S rRNA amplification failed were removed, and we retained measurements for a total of 16 CRC and 19 control samples. Relative gene fold change was calculated by applying the ΔΔcycle threshold (Ct) method[81], with ΔCt calculated as the difference between *cutC* and 16S rRNA Ct values. Significance of the *cutC* versus 16S rRNA comparison was assessed by one-tailed Wilcoxon signed-rank test. The same procedure was applied to the quantification of *cutC* and 16S rRNA transcripts from cDNA, which was computed using 26 CRC and 20 control samples for which we obtained a reliable quantification of both *cutC* and 16S rRNA.

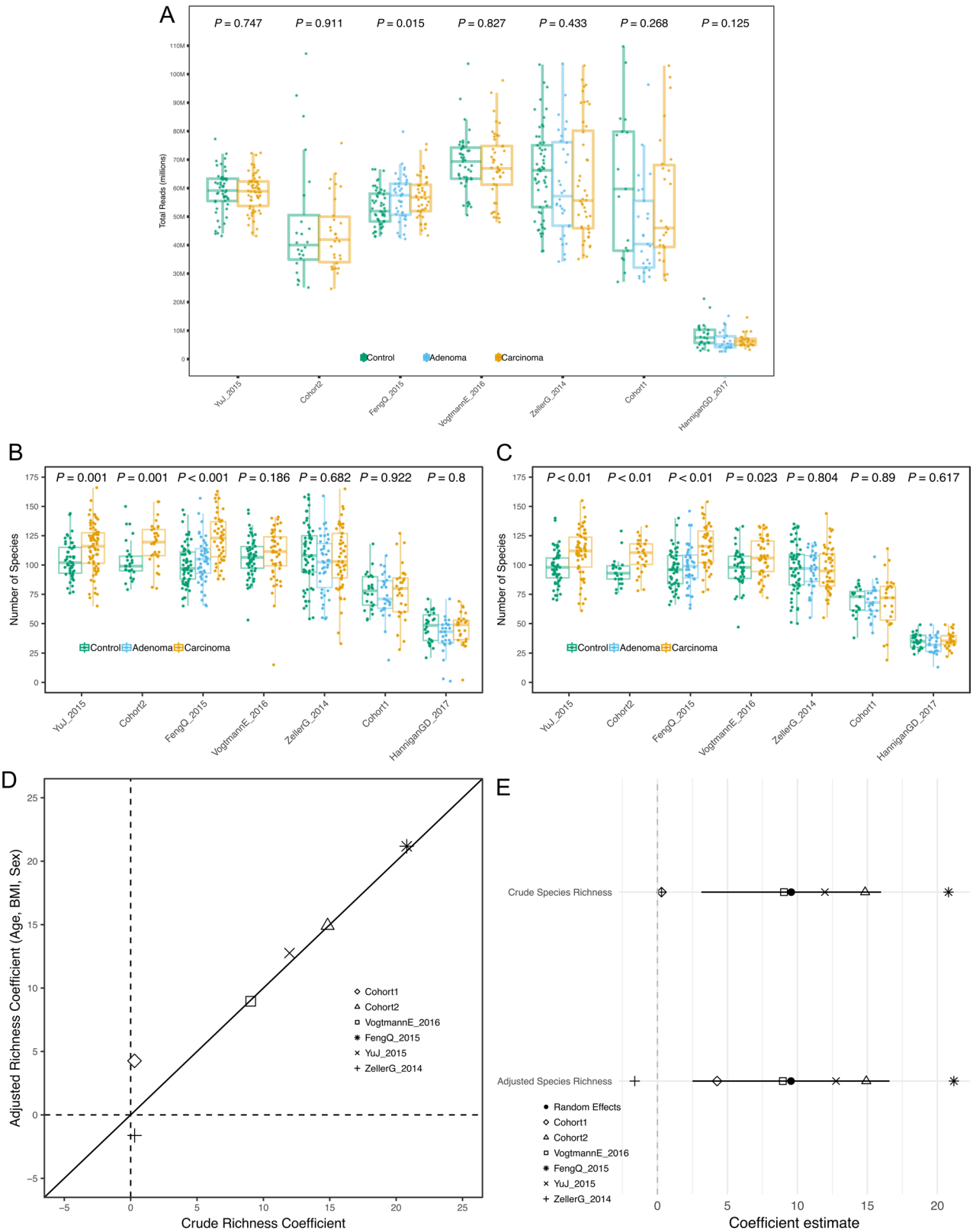
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

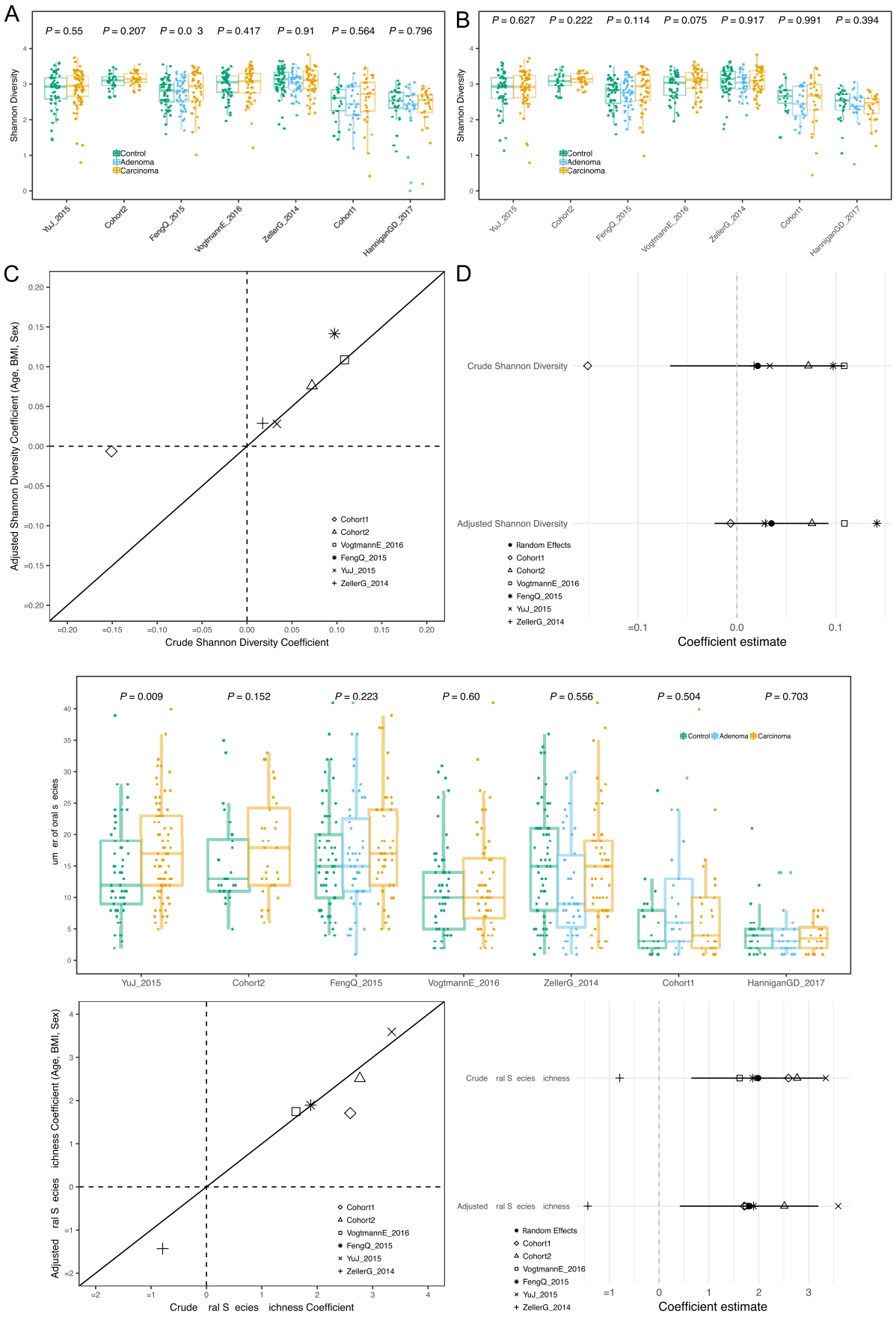
Nucleotide sequences for the two new Italian cohorts are available in the Sequence Read Archive under accession No. SRP136711. MetaPhlAn2 and HUMANn2 profiles for the new cohorts were also added to the curatedMetagenomicData R package[27] along with their corresponding metadata. Validation Cohort1 is available in the European Nucleotide Archive under the study identifier PRJEB27928; Validation Cohort2 is available in the DNA data bank of Japan databases under the accession No. DRA006684.

References

62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
63. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
64. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
65. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
66. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. Hastie, T, Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* Vol. 1 (Springer, 2009).
68. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
69. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
70. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
71. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
72. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
73. Kaminski, J. et al. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput. Biol.* **11**, e1004557 (2015).
74. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
75. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
76. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
77. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
78. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
79. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
80. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).
81. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C(T)}$ Method. *Methods* **25**, 402–408 (2001).

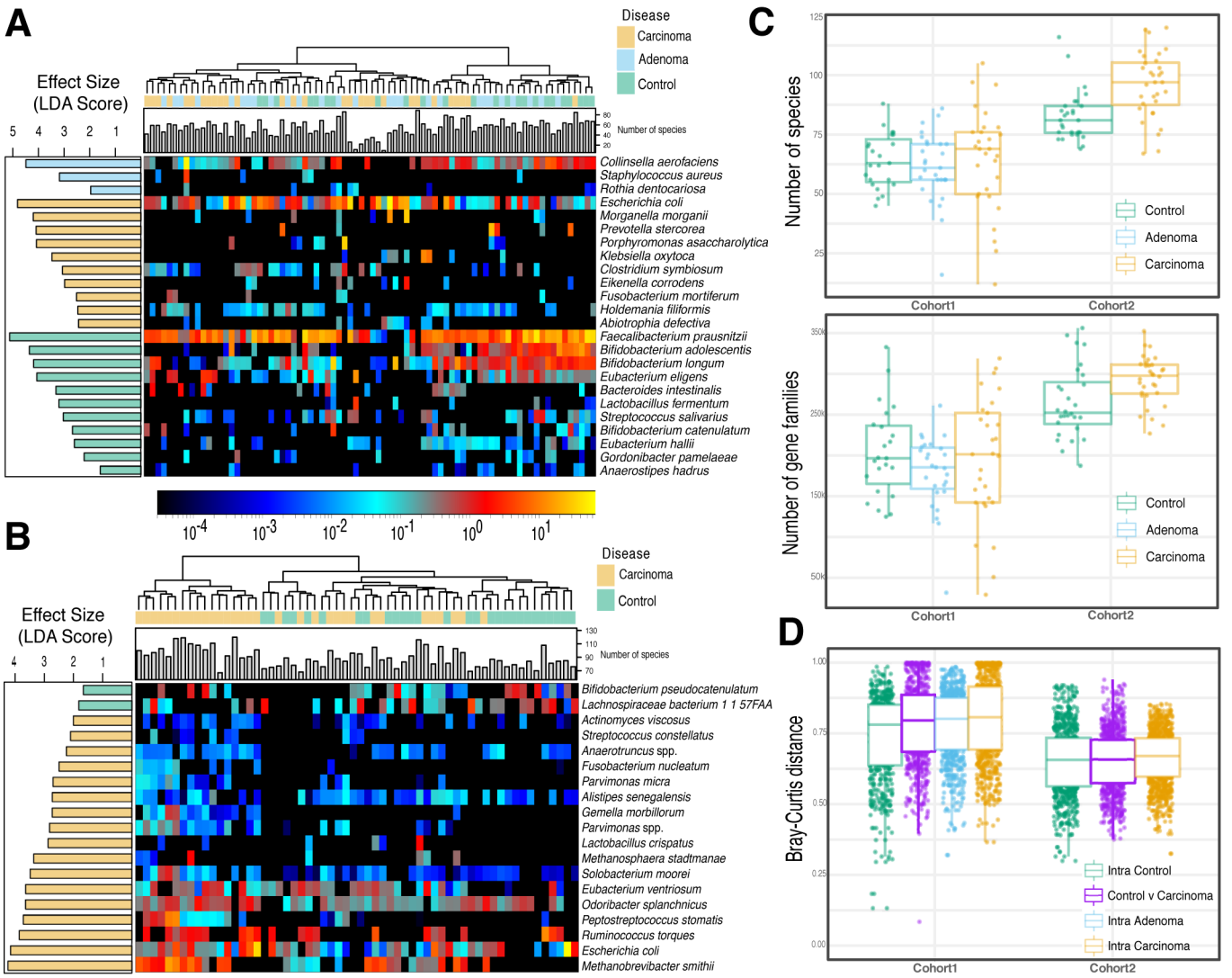


Extended Data Fig. 1 | Sequencing depths and species richness across CRC datasets a, Boxplots reporting the total number of reads in each dataset. *P* values between the carcinoma and control groups were calculated by two-tailed Wilcoxon rank-sum tests. **b**, Boxplots showing the total number of microbial species per dataset. *P* values were calculated by two-tailed Wilcoxon rank-sum tests. **c**, Boxplots showing the total number of microbial species per dataset calculated on metagenomes subsampled in each dataset to the number of reads of the tenth percentile. *P* values were calculated by two-tailed Wilcoxon rank-sum tests. **d**, Multivariate analysis of species richness using crude and age-, sex- and BMI-adjusted coefficients obtained from linear models. **e**, Meta-analysis of crude and adjusted multivariate richness coefficients using a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate.

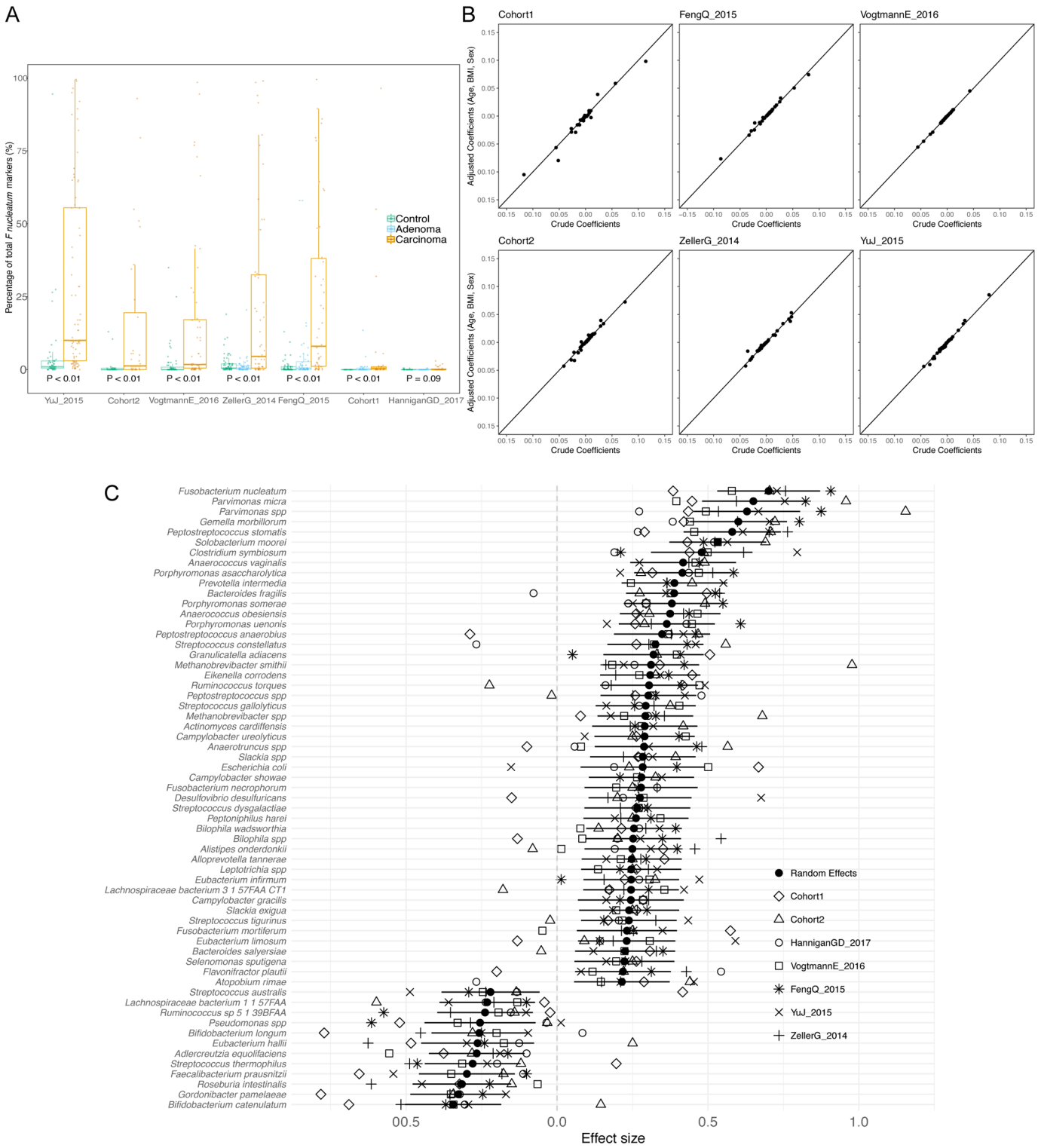


Extended Data Fig. 2 | See next page for caption

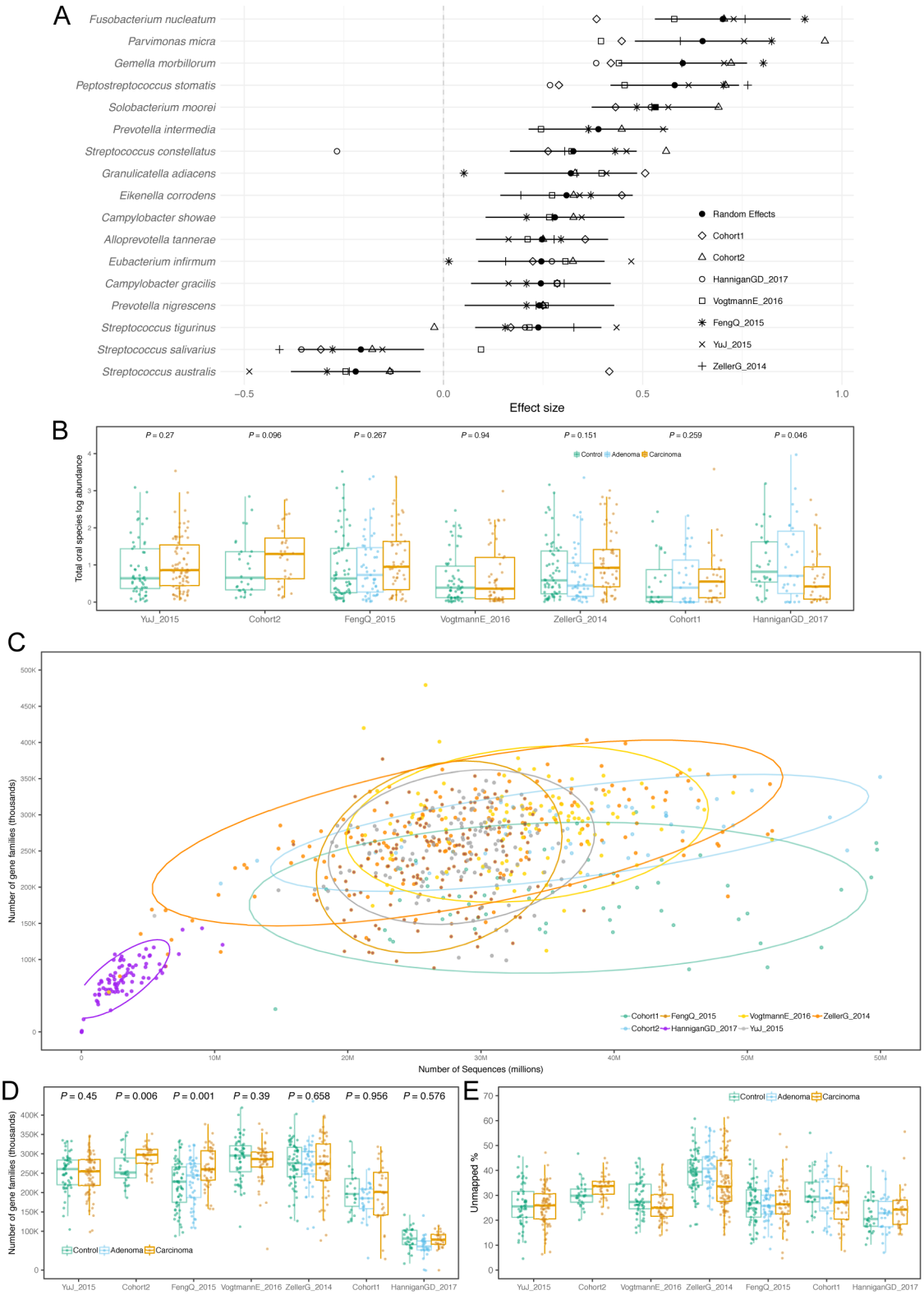
Extended Data Fig. 2 | Meta-analysis of species diversity and oral species richness in CRC datasets. a, Boxplots reporting the Shannon species diversity in each dataset. *P* values between the carcinoma and control groups were calculated by two-tailed Wilcoxon rank-sum tests. **b**, Boxplots reporting the Shannon species diversity calculated on metagenomes subsampled in each dataset to the number of reads of the tenth percentile. *P* values were calculated by two-tailed Wilcoxon rank-sum tests. **c**, Multivariate analysis of species diversity using crude and age-, sex- and BMI-adjusted coefficients obtained from linear models. **d**, Meta-analysis of crude and adjusted multivariate Shannon diversity coefficients using a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate. **e**, Boxplots reporting the total number of oral microbial species per dataset. *P* values were calculated by two-tailed Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **f**, Multivariate analysis of putative oral species richness using crude and age-, sex- and BMI-adjusted coefficients obtained from linear models. **g**, Meta-analysis of crude and adjusted multivariate putative oral species richness coefficients using a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate.



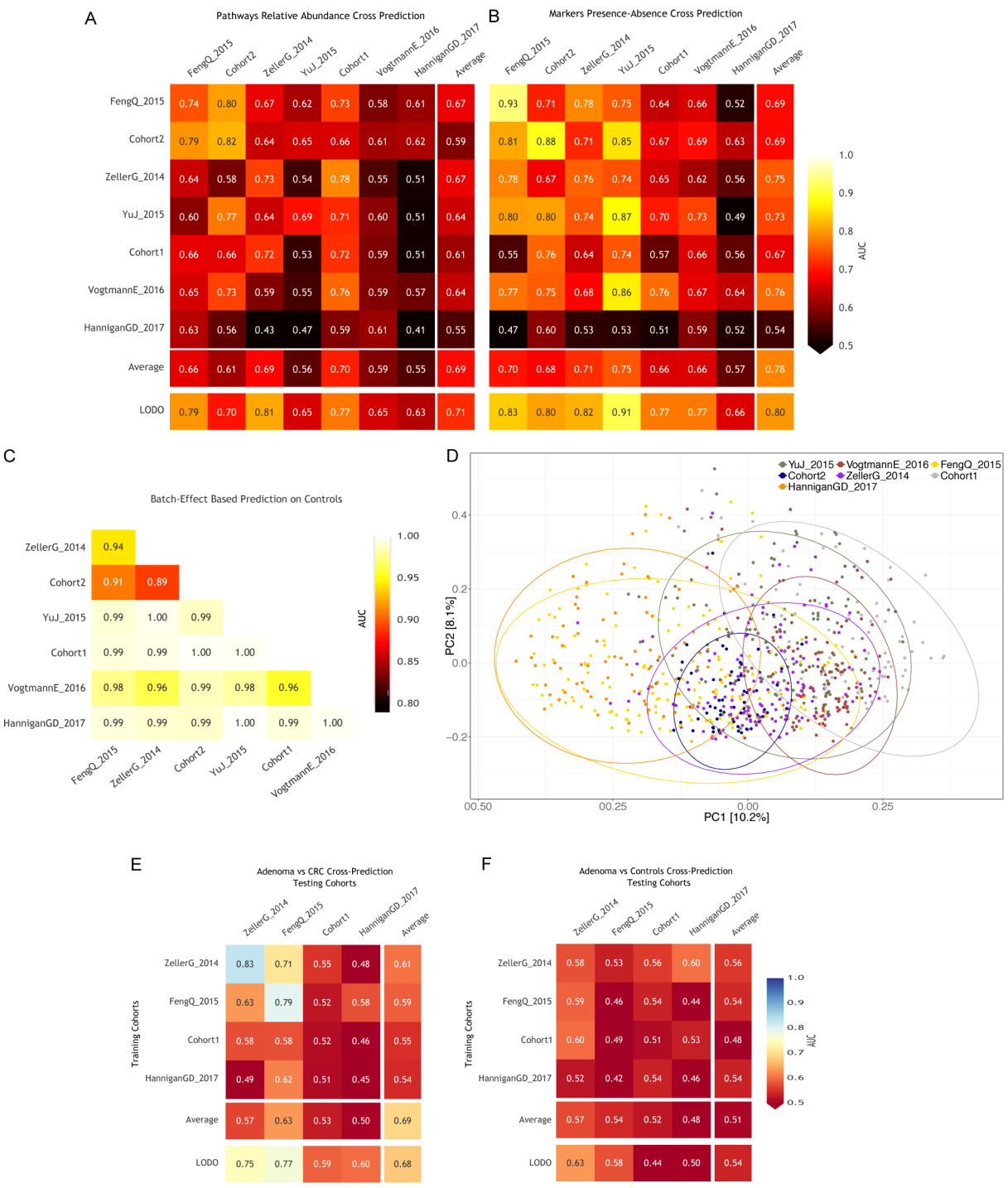
Extended Data Fig. 3 | Two metagenomic cohorts identify clear but only partially overlapping microbiome signatures associated with CRC. a,**b**, Relative abundances (log scale) and effect sizes (estimated using the linear discriminant analysis score in LEfSe) for the significantly different microbial species in CRC samples compared to control samples for Cohort1 (significance assessed by the non-parametric test in LEfSe) (**a**) and Cohort2 (**b**). **c**, Alpha-diversities measured as the total number of species and total number of UniProt90 gene families in each sample for the two cohorts. **d**, Beta-diversities estimated with the Bray–Curtis dissimilarity metric for intra- and inter-condition comparisons in the two cohorts.



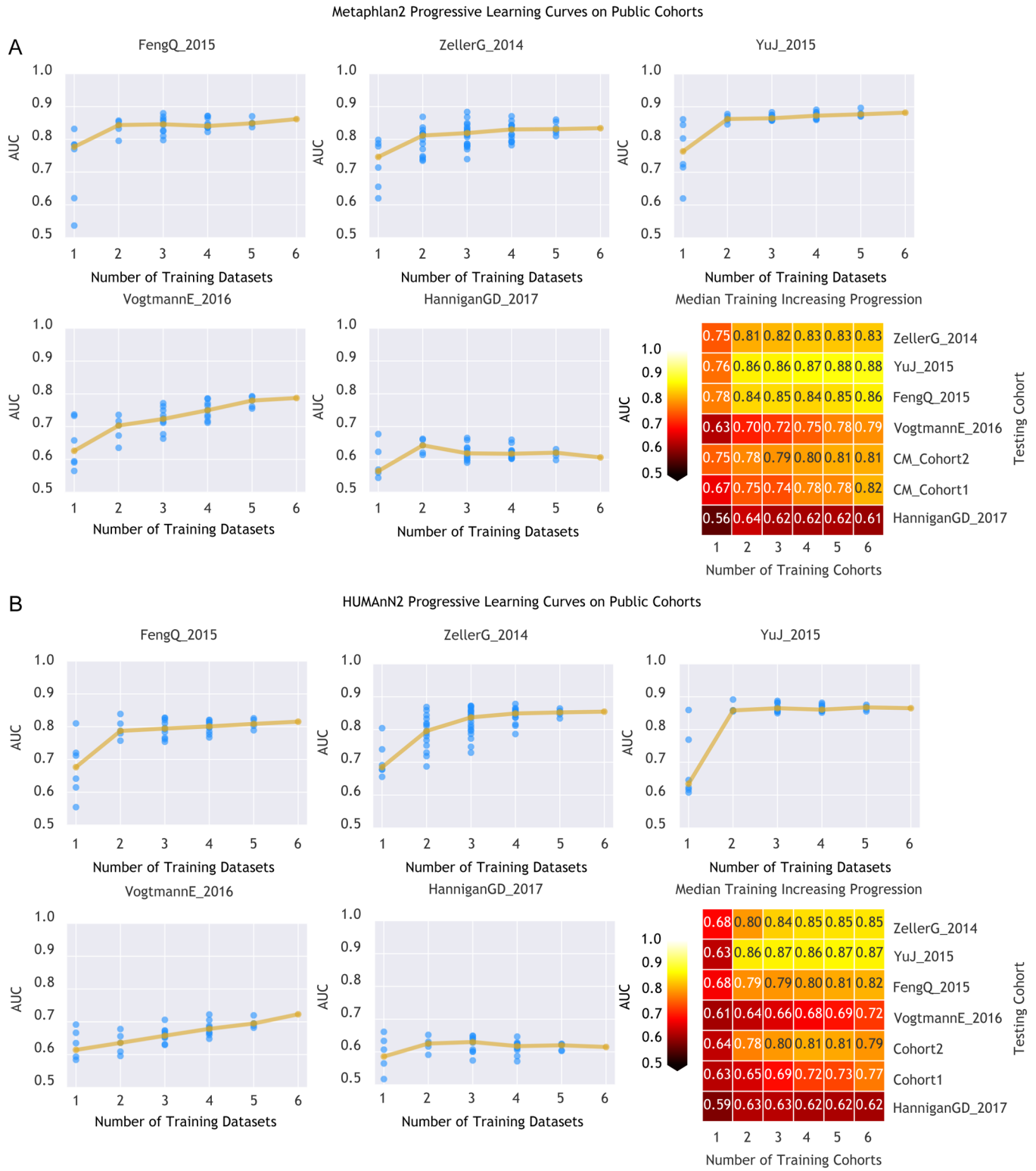
Extended Data Fig. 4 | Analysis of *F. nucleatum* markers and taxonomic meta-analysis of CRC datasets. a, Percentages of *F. nucleatum* clade-specific markers (200 in total) in each dataset. *P* values were obtained by two-tailed Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **b**, Multivariate analysis of meta-analysis species-level abundance biomarkers. Crude and age-, sex- and BMI-adjusted coefficients for species associated with disease status in the meta-analysis of standardized mean differences. **c**, Meta-analysis of CRC datasets using species-level MetaPhlAn2 profiles. Bold lines represent the 95% confidence interval for the random effects model estimate.



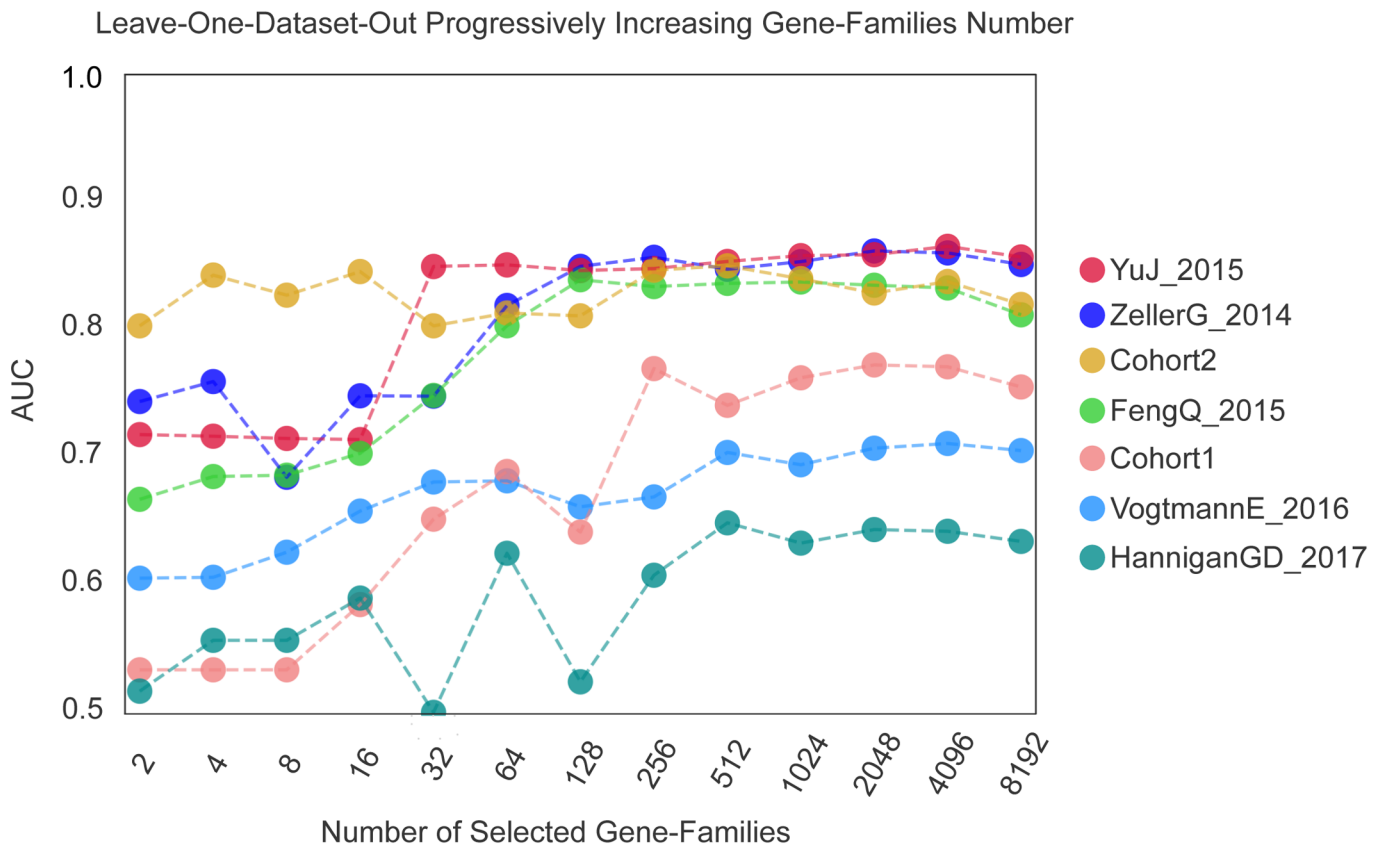
Extended Data Fig. 5 | Analysis of putative oral species abundances in CRC datasets and gene-family richness across CRC datasets. a, Effect sizes of the abundances of significant putative oral species identified using a meta-analysis of standardized mean differences and a random effects model. Bold lines represent the 95% confidence interval for the random effects model estimate. **b**, Total abundance of putative oral species in each gut metagenomic dataset. *P* values were obtained by two-tailed Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **c**, The total number of reads in each sample of each dataset correlates with the total number of gene families identified using HUMANn2. Ellipses represent the 95% confidence level assuming a multivariate *t*-distribution. **d**, Distribution of the total number of gene families identified in the samples of each dataset. *P* values were obtained by two-tailed Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **e**, Distribution of the percentages of unmapped reads across datasets for UniProt90 gene families.



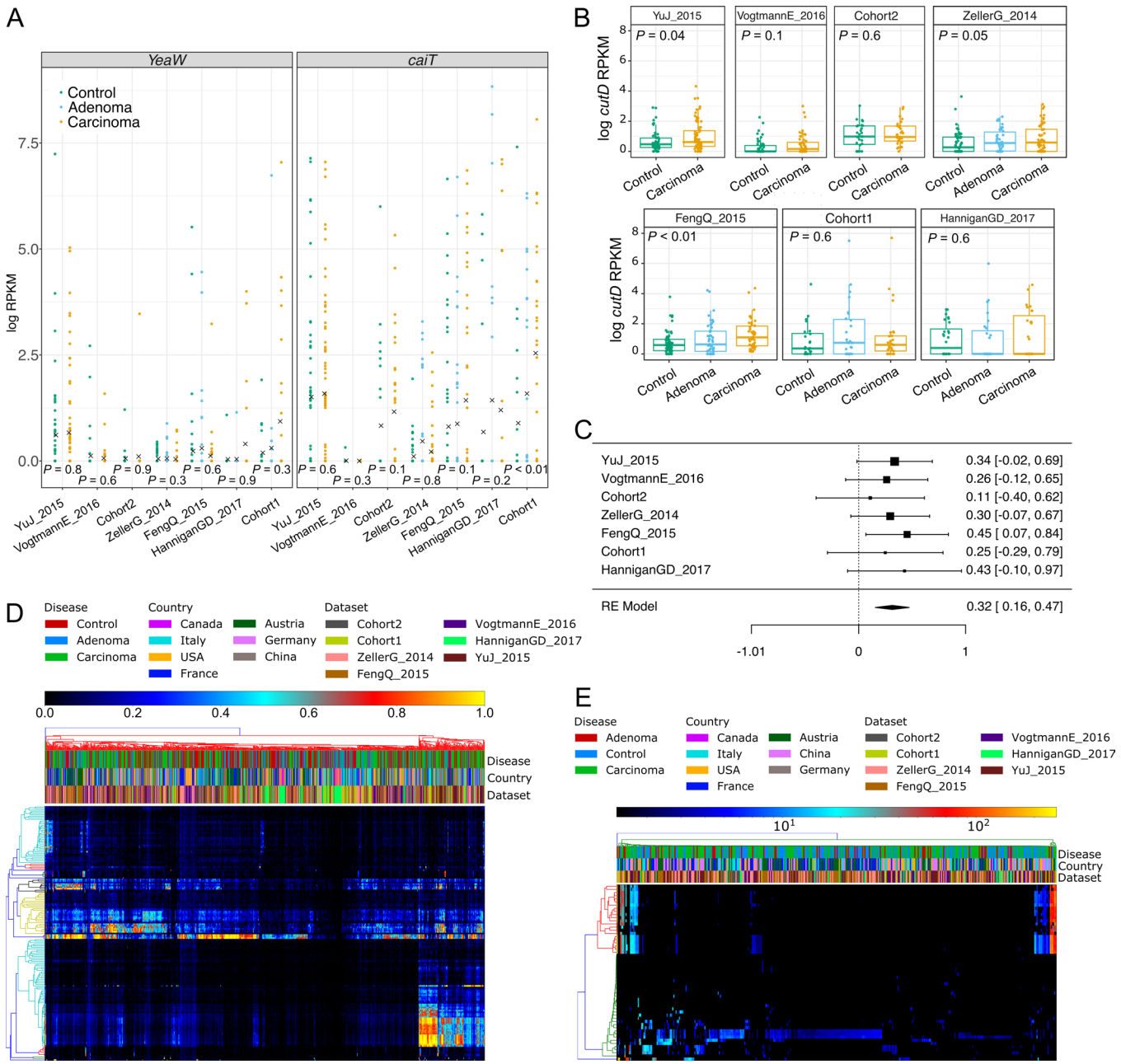
Extended Data Fig. 6 | Cross-validation, cross-cohort and LODO predictions using pathway abundances, species abundances and species-specific markers. a, Prediction matrix reporting prediction performances as AUC values obtained using a random forest model on pathway relative abundances. Values on the diagonal refer to 20 times repeated tenfold stratified cross-validations. Off-diagonal values refer to the AUC values obtained by training the classifier on the dataset of the corresponding row and applying it to the dataset of the corresponding column. The LODO row refers to the performances obtained by training the model on pathway abundances using all but the dataset of the corresponding column and applying it to the dataset of the corresponding column. **b**, Prediction matrix as in **a** but using MetaPhlAn2 marker presence and absence information. **c**, Prediction of samples-to-cohort assignments using species-level relative abundances. Only control samples from each dataset are considered. **d**, Principal coordinate analysis of Bray–Curtis distances computed on MetaPhlAn2 species-level abundances across datasets. Ellipses represent the 95% confidence level assuming a multivariate *t*-distribution. **e**, Cross-prediction matrix for the performances of random forest models in predicting adenomas versus CRC conditions. **f**, Cross-prediction matrix as described in **e** but on the distinction of adenomas versus controls.



Extended Data Fig. 7 | Prediction performances with increasing numbers of external datasets considered in the training model. a, Prediction performances computed based on MetaPhlAn2 species abundances. The dark yellow line interpolates the median AUC at each number of training datasets considered. **b**, Prediction performances computed based on HUMANn2 gene-family abundances.

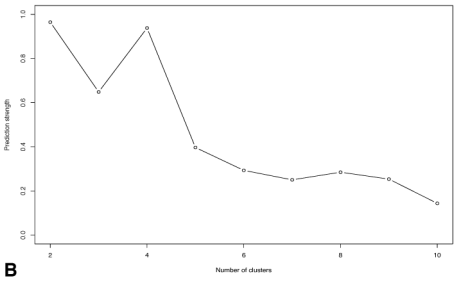


Extended Data Fig. 8 | Identification of a minimal number of microbial gene families for CRC detection. Prediction performances in the LODO settings at increasing numbers of gene families. Each ranking is obtained excluding the testing dataset to avoid overfitting.



Extended Data Fig. 9 | Metagenomic analysis of genes involved in the TMA synthesis pathway. a, ShortBRED analysis of *yeaW* and *caiT* gene abundances. Points represent the log of RPKM for each sample and crosses represent average values per group/dataset. **b**, ShortBRED analysis of *cutD* gene abundances. Boxplots report the RKPM abundances obtained using ShortBRED for the gene of the activating TMA-lyase enzyme *cutD*. *P* values were calculated by two-tailed Wilcoxon rank-sum tests comparing values between controls and carcinomas for each dataset. **c**, Forest plot showing effect sizes calculated using a meta-analysis of standardized mean differences and a random effects model on *cutD* RPKM abundances between carcinomas and controls. **d**, Breadth of coverage of *cutC* gene sequence clusters across CRC datasets. **e**, Depth of coverage of *cutC* gene sequence clusters across CRC datasets.

A



B

Cluster	Carcinomas	Adenomas /Controls	Taxonomy of *cutC* sequences	P-value
Cluster 1	49 (26%)	32 (16%)	*Firmicutes; Megasphaera sp DJF B143* *Proteobacteria; Enterobacter aerogenes* *Proteobacteria; Escherichia coli* *Proteobacteria; Klebsiella oxytoca* *Proteobacteria; Klebsiella pneumoniae* *Proteobacteria; Klebsiella varicola*	
Cluster 2	142 (74%)	172 (84%)	*Actinobacteria; Collinsella sp CAG 289* *Actinobacteria; Collinsella sp MS5* *Actinobacteria; Collinsella tanakaei* *Firmicutes; Clostridiales bacterium VE202 28* *Firmicutes; Clostridium asparagiforme* *Firmicutes; Clostridium botulinum* *Firmicutes; Clostridium clostridioforme* *Firmicutes; Clostridium hathewayi CAG 224* *Firmicutes; Clostridium sp FS41* *Firmicutes; Enterococcus avium* *Firmicutes; Enterococcus faecium* *Firmicutes; Hungatella hathewayi* *Unknown uncultured Eubacterium ssp.* *Proteobacteria; Desulfovibrio desulfuricans*	0.01

Cluster	Carcinomas	Adenomas /Controls	Taxonomy of *cutC* sequences	P-value
Cluster 1**	50 (26%)	29 (14%)	*Firmicutes; Clostridiales bacterium VE202 28* *Firmicutes; Clostridium asparagiforme* *Firmicutes; Clostridium botulinum* *Firmicutes; Clostridium clostridioforme* *Firmicutes; Clostridium hathewayi CAG 224* *Firmicutes; Clostridium sp FS41* *Firmicutes; Enterococcus avium* *Firmicutes; Enterococcus faecium* *Firmicutes; Hungatella hathewayi*	
Cluster 2	26 (13%)	24 (11%)	*Proteobacteria; Desulfovibrio desulfuricans* *Actinobacteria; Collinsella_tanakaei* *Actinobacteria; Collinsella sp CAG 289* *Actinobacteria; Collinsella sp MS5*	2.225e-05
Cluster 3	49 (25%)	32 (15%)	*Firmicutes; Megasphaera sp DJF B143* *Proteobacteria; Enterobacter aerogenes* *Proteobacteria; Escherichia coli* *Proteobacteria; Klebsiella oxytoca* *Proteobacteria; Klebsiella pneumoniae* *Proteobacteria; Klebsiella varicola*	
Cluster 4**	66 (34%)	119 (58%)	*Unknown uncultured Eubacterium ssp.*	

Extended Data Fig. 10 | Cluster analysis of representative cutC sequence variants of samples. a, Prediction strengths at differing numbers of clusters showing optimum numbers at two and four clusters. **b**, Tables showing the number of samples for carcinomas, adenomas and controls with breadth of coverage >80% at two different cluster thresholds. *P* values were calculated using a Fisher *t*-test, and taxonomy was assigned by BLASTn and the *cutC* sequence database (criteria of 80% coverage, >97% identity and minimum 2,000 nt alignment length).

nature research

Corresponding author(s): Nicola Segata

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 *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 *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	No commercial software has been used for data collection, all the software used is reported in the paper
Data analysis	All data analysis has been performed with open source software as comprehensively described in the methods of the paper

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

Nucleotide sequences for the two new Italian cohorts are available in the Sequence Read Archive (SRA) under the accession number SRP136711. MetaPhlAn2 and HUMANn2 profiles for the new cohorts were also added to the curatedMetagenomicData R package along with their corresponding metadata. Validation Cohort1 is available in the European Nucleotide Archive (ENA) under the study identifier PRJEB27928, Validation Cohort2 is available in the DDBJ databases under the accession number DRA006684.

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	This is a meta-analysis of 9 shotgun metagenomic datasets for a total of 969 samples, details on the sample size of each dataset is reported in the paper
Data exclusions	Samples not passing quality-control have been excluded. Details of the exclusion criteria are reported in the paper
Replication	This is a meta-analysis so the main focus is indeed on the reproducibility. We thus used cross validaton, cross-dataset prediction, leave-one-dataset-out validation, and independent validation on additional cohorts. qPCR measurements were also done in triplicates
Randomization	All datasets are from a case/control design as described in the original publications and in our method section
Blinding	Samples were collected from treatment- and diagnosis- naive subjects

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	We report this information in Table 1 and Suppl. Table 1
Recruitment	All subjects meeting the inclusion criteria were enrolled until the required sample size was reached
Ethics oversight	The two new clinical studies performed here were approved by the relevant ethics committees (Cohort1: Ethics committee of Azienda Ospedaliera "SS. Antonio e Biagio e C. Arrigo" of Alessandria, Italy, protocol N. Colorectal_miRNA_CEC2014 and Cohort2: Ethics committee of European Institute of Oncology of Milan, Italy, protocol N. R107/14-IEO 118) and informed consent was obtained from all participants

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