Combined metagenomic analysis of colorectal cancer datasets defines cross-cohort microbial diagnostic signatures

Manghi P(1)[†] , Thomas AM(1,2), Asnicar F(1), Pasolli E(1), Armanini F(1), Zolfo M(1), Beghini F(1), Pozzi C(3), Gandini S(3), Serrano D(3), Tarallo S(4), Francavilla A(4), Gallo G(5.6), Trompetto ;(5,6), Cordero F(7), Dias-Neto E(8), Setubal JC(2,9), Pardini B(4,10), Rescigno M(11), Waldron L(12,13), Naccarati A(4,14), Segata N(1)

(1) Centre for Integrative Biology, University of Trento, Trento, Italy.
(2) Biochemistry Department, Chemistry Institute, University of São Paulo, São Paulo, Brazil.
(3) Department of Experimental Oncology, European Institute of Oncology, Milan, Italy.
(4) Italian Institute for Genomic Medicine (IIGM; formerly Human Genetics Foundation - Hugef), Turin, Italy.
(5) Department of Surgical and Medical Sciences, University of Catanzaro, Catanzaro, Italy.
(6) Department of Colorectal Surgery, Clinica S. Rita, Vercelli, Italy.
(7) Department of Computer Science, University of Turin, Turin, Italy
(8) Medical Genomics Laboratory, CIPE/A.C. Camargo Cancer Center, São Paulo, Brazil.
(9) Biocomplexity Institute of Virginia Tech, Blacksburg VA 24061, USA
(10) Department of Medical Sciences, University of Turin, Turin, Italy.
(11) Mucosal immunology and microbiota Unit, Humanitas Research Hospital, Milan, Italy.
(12) Graduate School of Public Health and Health Policy, City University of New York, New York, USA.
(13) Institute for Implementation Science in Population Health, City University of New York, New York, USA.
(14) Department of Molecular Biology of Cancer, Institute of Experimental Medicine, Prague, Czech Republic.

[†] Email: paolomanghi1974@gmail.com

Motivation
Several studies have investigated the link between the gut microbiome and colorectal cancer (CRC), and microbial biomarkers constituting a hypothetical diagnostic signature have been identified. However, these studies have been limited by sample size or used microbiome tools with a limited taxonomic resolution. Moreover, the transferability of such microbiome signatures across cohorts, populations, and other confounding factors have not been comprehensively assessed so far.
Methods
We used whole metagenome sequencing to uniformly quantify taxonomic and functional abundance in fecal metagenomes from 140 participants recruited in two Ital-

ian cohorts along with all 5 available CRC metagenomic datasets, totaling 352 carcinomas and 312 controls. Our analyses exploited 4 types of microbiome quantitative profiles: taxonomic species-level relative abundances and marker presence or absence patterns inferred by MetaPhlAn2, gene-family, and microbial pathway relative abundances estimated by HUMAnN2. Univariate analyses on a per dataset basis was performed using LEfSe to identify features that were statistically different among groups and estimate their effect size. We applied arcsine-square root transformation on the functional and taxonomic relative abundances. We then used the escalc function from the R metafor package that employes Cohen's standardized mean difference statistic to build a random effects model. Predictive machine learning experiments were performed using Random Forest (RF) on distinct learning tasks. Specifically, we measured the inside-dataset prediction capability of the microbiome using 10 fold cross validations. Cross-cohort training-validation predictions were performed on all possible pairs of distinct datasets. A meta-cohort approach was then applied on all datasets except the one used for training (Leave-One-Dataset-Out - LODO - approach) to test the generalizability of multi-cohort models. We also performed experiments at increasingly larger subsets of samples and features (using RF feature selection) to assess the impact of training set size and identify minimal predictive microbial signature.

Results

Our meta-analysis considered stool metagenomes from 352 CRC patients and 312 controls. We first sought robust taxonomic and functional biomarkers for CRC, identifying a panel of confirmed over-represented species including Fusobacterium nucleatum, Parvimonas, and Peptostreptococcus stomatis, and newly strongly associated species such as Streptococcus tigurinus, Streptococcus dysgalactiae, and 3 Campylobacter species (A). Functional potential analysis identified gluconeogenesis and the putrefaction and fermentation pathways to be associated with CRC, whereas the stachyose and starch degradation pathways were associated with controls (B). When we assessed the diagnostic potential of the CRC microbiome, we found that the within cohort cross-validation performances were higher than cross-cohort predictions that were ranging from AUC 0.55 to AUC 0.84 (C). Models trained on the combination of multiple datasets showed on the contrary consistent and high performances on distinct cohorts (avg AUC 0.81, D). Different feature types (markers, gene families) gave similar results (avg AUC 0.80, 0.79, D). Progressively increasing the number of training cohorts gives monotonically improving scores for the newly sequenced cohorts (E). RF features selection highlighted that almost optimal performance is achieved using a panel of only 16 species, with a 1% improvement in the mean AUC value when all the species are considered (F). Altogether, our results highlight the very high clinical predictive potential of the microbiome in CRC. Moreover, we point out that heterogeneity of the meta-cohort and sample size are key factors in the accuracy of the predictive models, and thus motivate the need for sequencing additional CRC-associated cohorts to further refine biomarker discovery and metagenomics-based diagnostic tools.

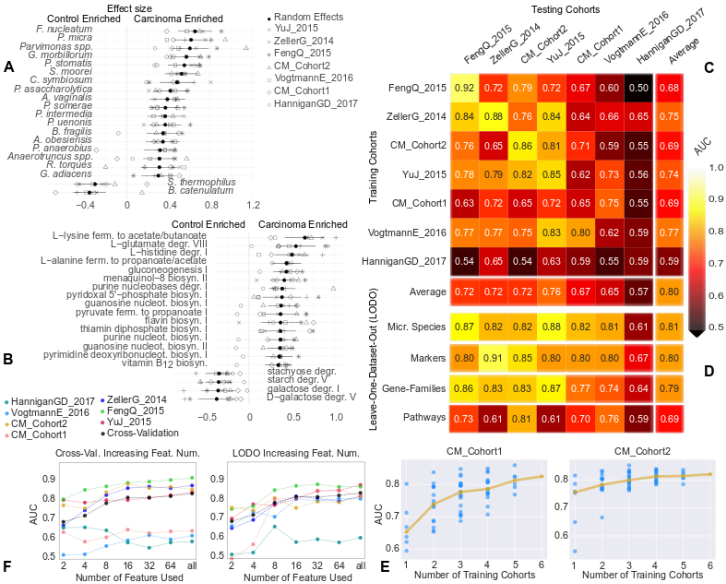

Figure 1 - Reproducible taxonomic and functional microbial signatures for CRC across datasets. (**A**) Pooled effect sizes for the 20 significant features with the largest effect size calculated using a meta-analysis of standardized mean differences and a random effects model on MetaPhlAn2 species abundances and on (**B**) HUMANn2 pathway abundances. Bold lines represent the 95% confidence interval for the random effects model coefficient estimate (marked with a black circle). (**C**) Cross prediction matrix reporting prediction performances as AUC values obtained using a random forest (RF) model on species-level relative abundances. Values on the diagonal refer to 20 times repeated 10-fold 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 on the dataset of the corresponding column. The Leave-One-Dataset-Out (LODO) rows (**D**) refer to the performances obtained by training the model on the species-level abundances, MetaPhlAn2 markers presence-absence, HUMANn2 UniProt90 gene-families and functional pathways abundances, using all but the dataset of the corresponding column and applying it on the dataset of the corresponding column. (**E**) 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. (**F**) Prediction performances at increasing number of microbial species obtained by re-training the RF classifier on the N top ranked features identified with a first RF model training in cross-validation and LODO-setting.