Interpretable machine learning analysis of functional metagenomic profiles improves colorectal cancer prediction and reveals basic molecular mechanisms.

Carlos S. Casimiro-Soriguer
Fundacion Progreso y Salud
Carlos Loucera
Fundacion Progreso y Salud
María Peña-Chilet
Fundacion Progreso y Salud
Joaquin Dopazo (joaquin.dopazo@juntadeandalucia.es)
Fundacion Progreso y Salud https://orcid.org/0000-0003-3318-120X

Research

Keywords: Interpretable machine learning, CRC, adenoma, predictor, metagenome, Generalized Additive Models, meta-analysis.

Posted Date: January 22nd, 2020

DOI: https://doi.org/10.21203/rs.2.21634/v1

Abstract

Background Gut microbiome is gaining interest because its links with several diseases, including colorectal cancer (CRC). Results Here we performed a meta-analysis of 851 fecal metagenomic samples from five publicly available studies. We used an interpretable machine learning approach based on functional profiles, instead of the conventional taxonomic profiles, to produce a highly accurate predictor of CRC with better precision than those of previous proposals. Moreover, this approach is also able to discriminate samples with adenoma, which makes this approach very promising for CRC prevention by detecting early stages in which intervention is easier and more effective. In addition, interpretable machine learning methods allows extracting features relevant for the classification, which reveals basic molecular mechanisms accounting for the changes underwent by the microbiome functional landscape in the transition from healthy gut to adenoma and CRC conditions. Conclusion Functional profiles provide superior accuracy in predicting CCR and adenoma conditions than taxonomic profiles and additionally, in a context of explainable machine learning, provide useful hints on the molecular mechanisms operating in the microbiota behind these conditions.

Background

In the last years the study of the microbiome has progressively gained interest, especially in the context of human health [1-4]. Microbial abundance profiles based on 16S rRNA genes have been used to study microbiomes, although whole genome sequencing (WGS) is becoming increasingly popular in recent years due to the decreasing sequencing costs [5, 6]. Contrary to 16S rRNA data, WGS microbiome data provides the real gene composition in the bacterial pool of each sample, which allows identifying strain-specific genomic traits [7, 8]. During the last years, microbiome WGS has been used to explore microbiome–host interactions within a disease context by means of metagenome-wide association studies, that allow studying gut microbiome alterations characteristic of different pathologic conditions [3, 9-17]. In particular, recent evidence suggests that the human gut microbiome could be a relevant factor in human diseases [18, 19]. In fact, the existence of carcinogenic mechanisms mediated by bacterial organisms has recently been proposed [20-22]. And, more specifically, it has been suggested that the gut microbiome could play a relevant role in the development of colorectal cancer (CRC) [15, 16, 23-25]. Due to this, the gut microbiome has been proposed as a potential diagnostic tool for CRC [16, 17, 26, 27]. Nevertheless, its reproducibility and the predictive accuracy of the microbial gene signatures across different cohorts have been questioned [28, 29]. The increasing availability of whole metagenome shotgun datasets of CRC cohorts [15-17, 26, 27] facilitates large-scale multi-population exploratory studies of the CRC-associated microbiome at the resolution level of strain [30, 31]. In two recent studies, a combined analysis of heterogeneous CRC cohorts was able to identify reproducible microbiome biomarkers and accurate disease predictive models that open the door to future clinical prognostic tests [28, 29]. The subsequent meta-analysis of the functional potential in the strains of the signature found gluconeogenesis and putrefaction and fermentation pathways associated with CRC, in coherence with the current knowledge on microbial metabolites implicated in carcinogenesis [32].

It is important to note that the current approaches used to obtain biomarkers with predictive power use microbial strain or gene signatures as features to train a predictive model. Since genes or strains have not a clear interpretability by themselves, the interpretation of the results of the classification produced by the model relies on the analysis of the potential functionalities encoded by these features. In other words, the predictive model is built using features that need to be interpreted a posteriori [33]. In fact, this is a relatively common problem with many current machine learning techniques, which have evolved in recent years to enable robust association of biological signals with measured phenotypes but, in many cases, such approaches are unable to identify causal relationships [34, 35]. However, the interpretability of models, especially in a clinical context, is becoming an increasingly important issue [34-36]. The use of features with a direct functional interpretation has been suggested as crucial for the interpretability of the models [37]. In a recent study, gene profiles derived from WGS of samples of the MetaSub project [38] were initially transformed into functional profiles, which account for bacterial metabolism and other cell functionalities, and have subsequently been used as features to build a city classification machine learning algorithm [39]. Since the features are informative by themselves, their relevance in the classification provides an immediate interpretability to the prediction model built.

Here, we propose an interpretable machine learning approach in which functional profiles of microbiota samples, with a direct interpretation, are first obtained from shotgun sequencing and subsequently used as features for predicting CRC in the patient donor of the sample. Moreover, in the prediction schema proposed, a feature relevance method allows extracting the most important functional features that account for the classification. Thus, any sample is described as a collection of functional modules contributed by the different bacterial species present in it, which account for the potential functional activities that the bacterial population in the sample, as a whole, can perform.

Results

Functional profiles discriminate between CRC, adenoma and normal samples across different cohorts

The application of the machine learning pipeline described in methods using functional profiles across the cohorts used after removing the batch effect produced a clear separation between healthy samples and CRC samples. Moreover, contrarily to the case of previous classification models based on taxonomic profiles [28, 29], the two adenoma types, although with less accuracy than the CRC samples, were also well differentiated (see Figure 1). The figure also showcases the predictive power of the selected features As it has been previously noted [28, 29] sample prediction accuracy is not uniform across projects and some of them present a poorer prediction results when the general predictor model trained with the rest of projects is used (see Supplementary Figure 1). However, many of these misclassified samples are special cases either with comorbilities or with advanced metastases, which probably changes the functional landscape of the bacterial functionality.

The most relevant features for the classification provides a biological interpretation of the model and account for disease mechanisms

The feature selection process used to build the classifier renders a number of features that can be ranked by relevance (see Figure 2A). The threshold based on the elbow point of the medians select a total of 666 orthologs (Supplementary Table 1), that map into 250 pathways (listed in Supplementary Table 2 and represented over a general metabolism map in Supplementary Figure 2) and correspond to 155 KEGG functional modules (see Supplementary Table 3), that are summarized in Supplementary Figure 3. Table 1 shows the ten most relevant modules. Figure 3 shows the distribution of samples along the value of the feature, which allows understanding the contribution of each feature to the disease condition. As described in the Discussion section all these features are related to bacterial activities that have directly or indirectly been linked to CRC.

The most relevant features that differentiate adenomas from CRC and normal samples

In order to estimate the contribution of features to the classification of the adenoma samples, we used the SHAP method. Supplementary Table 5 lists the most relevant features for the classification of large and small adenomas according to the SHAP scores. Supplementary Figure 5 shows the Plots of feature interpretability for the most relevant features used by the model for the discrimination of small and large adenoma. Most of them are specific features of CRC that start to appear probably in the first adenoma stages previous to the cancer. Some of them are known CRC prognostic biomarkers, such as the archaeal *large subunit ribosomal protein L19e* [40], while other are characteristic of the adenoma stage, like the *coenzyme F420 hydrogenase subunit beta*, related to methane metabolism, enhanced in adenoma patients [41] or the *chemotaxis protein methyltransferase CheR*, which has been described as a distinctive feature of patients with adenomas [42].

Comparison of functional and taxonomic profiles

In order to compare the performance obtained in the classification that uses functional profiles as features with the conventional approach in which the features used are taxonomic profiles we obtained per sample taxonomic profiles as described in Methods and repeated the training procedure to obtain another predictor model. Once ranked by relevance (Figure 2B), a total of 132 taxonomic features were over the threshold (listed in Supplementary Table 4 and represented in a phylogenetic tree in Supplementary Figure 4).

Figure 4 shows the relative performance of the prediction using both functional and taxonomic profiles. The predictor based on taxonomic profiles results to be slightly better only in a few cases for the easiest problem of distinguishing between CRC and healthy samples. However, it is interesting to note that in the more difficult problem of distinguishing between adenomas, normal and CRC samples, taxonomic profiles clearly show a poorer performance than functional profiles (compare Figure 5 and Figure 1).

It is interesting to note that when the correlation between the most relevant functional features and the most relevant taxonomic features across samples is studied it becomes apparent that it varies among the different experiments (Supplementary Figure 6). This observation suggests that the particular CRC

microbiota profile defined by the taxonomic features can be reached by means of different combination of bacterial species.

Comparison with other approaches

The availability of whole metagenome shotgun datasets of CRC cohorts is relatively recent [15-17, 26, 27] and for this reason there are no much studies on CRC prediction from fecal microbiome data. An early study with 156 samples obtained a classification accuracy, as area under the curve (AUC), of 0.84 [16], although the samples belonged to a unique cohort. A more recent study of 526 samples found seven enriched bacterial markers that classified cases from controls with an accuracy of 0.80 [43] although it has been suggested that could be affected of over fitting issues [29]. The last study published, that outperforms the previous ones, obtained an average AUC 0.84 [29]. Here a more detailed comparison of the results obtained in this last study with the results obtained here is shown. Figure 4 depicts a comparison of the AUC scores obtained in this study and in the last study published [29]. Interestingly, the internal validation of our method, a 20 times repeated tenfold cross-validation over the whole dataset shows a systematic advantage of the approach presented here over the mentioned previous study. It is also interesting to note that both results present a similar trend across all the projects. Moreover, as commented below in the Discussion section, our results are, from a biological point of view, congruent with the findings in the literature.

Discussion

This study uses a comprehensive collection of the cohorts of CRC (listed in Table 2). Only one available dataset was discarded, PRJEB12449, frozen for more than 25 years before it was sequenced [44], which most probably compromised the quality of the results [45], and, actually, was described as technically flawed by previous studies [28, 29].

The interpretable machine learning approach used here has demonstrated to outperform other class predictors previously reported. One of the most interesting properties of this approach is its immediate interpretability. Thus, the features chosen by the model that optimize the discrimination between the conditions compared account for the functionalities that operate differentially among both conditions. The most relevant feature is the *Heptose II phosphotransferase* (K02850). This enzyme, located in the *Lipopolysaccharide biosynthesis* (ko00540) pathway, is associated with CRC in high values (see Figure 3A). Actually, the presence of lipopolysaccharides produced in the surface of Gram- bacteria has been reported to induce an inflammatory response as well as to stimulate the proliferation of colon carcinoma [46, 47]. The second most relevant feature is *Manganese/zinc/iron transport system permease protein* (K11708). According to the model the presence of this feature increases the probability of CRC (see Figure 3B). This transporter increases its number in excess iron conditions that are known to promote colorectal carcinogenesis [48]. *Tagatose 6-phosphate kinase* (KO00917), the third most relevant feature, belongs to the *Galactose metabolism* (ko00052) metabolic pathway. According to the model (Figure 3C), high values seem to be protective (predictor of healthy status). Actually, the metabolism of galactose is

related to diets rich in fiber (fruits and vegetables) that have a protective effect against CRC [49]. *Methyltransferase* (K16168), related to polyketide synthesis, is the next most relevant feature. In this case, the model predicts that low levels of *Methyltransferase* activity are associated to low probability of CRC (Figure 3D). It has recently been described that a class of molecules, colibactins, are produced from the gene cluster called the *polyketide synthase island* that occurs in certain strains of *Escherichia coli* prevalent in the microbiota of CRC patients [50]. Another relevant feature is *Methylaspartate mutase sigma subunit* (K01846), whose high activity is related with high probability of CRC according to the model (Figure 3E). It has been described that cancer cells undergo modifications that include increased glutamine catabolism and over-expression of enzymes involved in glutaminolysis, including glutaminase [51], which is liberated to the gut [52] and promotes the proliferation of bacteria containing this bacterial module. Also, *Kynureninase* (K01556), present in the tryptophan metabolism, is another highly relevant feature. *Kynurenine* is produced in the digestion of red meat [53] and it is known that low tryptophan levels correlate with immune system activation and better prognosis of CRC [54]. Moreover, *Kynurenine* has been proposed as a serum marker of CRC [55].

In the proposed predictor model based on functional features adenoma samples are separated from normal and CRC samples. The application of modern ML techniques like SHAP allows finding what, among the features used by the model, are more relevant for the classification of adenoma samples. Supplementary Figure 5 contains the most relevant features, which include proteins with functionalities related to CCR, such as amino acid metabolism [56], redox metabolism [57] or more generic biomarkers of bacterial proliferation [58]. Also, some specific features of adenoma, like the *coenzyme F420 hydrogenase subunit beta*, related to methane metabolism, enhanced in adenoma patients [41], or the *chemotaxis protein methyltransferase CheR* and the *pilus assembly protein FimV,* both related to chemotaxis and cell motility which have been described as a distinctive feature of adenomas [42], are also listed among the most relevant features. The most relevant feature, the *large subunit ribosomal protein L19e*, is an archaeal protein used as bad prognostic CRC biomarker [40]. Other CRC related proteins relevant in large adenoma are *phosphoglycerate kinase* and *glyceraldehyde-3-phosphate dehydrogenase (NADP)* that account for the altered glucose and glycine metabolism in the microbiota of CRC patients [28], or *heptosyltransferase III*, a virulence gene of pathogenic bacterial taxa, including *Fusobacterium* and *Providencia,* which have been related to CRC [59]. Similarly, in small adenoma some of the most relevant features are: *two-component system, LytTR family, response regulator AgrA*, being *lytTr* a regulator of virulence factors of several pathogens such as *Enterococcus fecalis* or *Clostridium difficile* [60] and AgrA a regulator of virulence of *Staphylococcus aureus* [61], *glutamate N-acetyltransferase / amino-acid N-acetyltransferase* [EC:2.3.1.352.3.1.1], involved in the degradation of glutamate recently linked to CRC [29], *aminoglycoside 3'-phosphotransferase II* [EC:2.7.1.95] that confers resistance to kanamycin [62] or the *L-lactate dehydrogenase (cytochrome)* [EC:1.1.2.3] related to piruvate metabolism, already linked to CRC [28].

An extensive description of the features selected by the model is beyond the scope of this paper, however it is worth noting that fully agree with the findings of functional analysis done in previous reports [28, 29].

Interpretability of the predictive models is becoming a major issue, especially in biomedicine [33, 34, 36]. The idea of using features with full biological meaning to gain interpretability in the machine learning methodology used has recently been proposed as a “white box” strategy [37] and has successfully been used for the first time in the analysis of urban microbiota [39] in the context of the METASub project [38]. Actually, when the relative performance of the predictors based either of functional or on taxonomic features is compared (Figure 4) functional profiles perform better in most of the projects distinguishing between CRC and healthy samples. However, functional profiles clearly overcome the performance of taxonomic profiles in distinguishing between adenoma, normal and CRC samples, which is a more difficult problem (compare Figure 5 and Figure 1). Actually, the correlation plots between functional and taxonomic features show how the same relevant functional features were defined by different combinations of bacterial strains in the different experiments used here (Supplementary Figure 6). All together, these observations suggest that the transition from normal condition to adenoma and CRC is not well defined in terms of strain abundances but there is a clear change at the level of functional activities of the bacteria in the sample that is better captured by functional profiles than by taxonomic profiles, which probably change at later stages, close to the CRC condition.

Conclusions

The interpretable machine learning approach proposed here has demonstrated a superior performance to other approaches previously proposed. Moreover, it demonstrated a better resolution not only with respect to the separation between healthy and CRC samples, but it is also able to discriminate samples with adenoma, being a promising tool for CRC prevention by detecting early stages in which intervention is easier and more effective. And finally, the model has a biological interpretation that provides important clues to better understand the mechanistic implications of the gut microbiota in CRC as well as in the previous stages of adenoma, which can have an interesting potential in preventive medicine and, specifically, in cancer interception [63].

Methods

Data

A total of 851 fecal metagenomic whole genome sequencing (WGS) samples (See Table 2) were analyzed. Sequence data were downloaded from European Nucleotide Archive. The conditions of the samples were obtained from the different supplementary tables of different publications (see Table 2) and complemented in the possible using the R package *curatedMetagenomicData* [64] available in Bioconductor [65].

Bacterial Whole Genome Sequence data processing

Whole genome sequencing data was managed using the NGLess-Profiler [66] package. Raw sequencing data preprocessing and quality control was carried out using a version of the *human-gut.ngl* pipeline. The

substrim built-in function was used to discard reads that do not meet the basic quality filter of being longer than 45 bases and having all bases with a *Phred* score over 25. To prevent potential contaminations with human genome sequences the reads were mapped against the human genome hg19, and those that do not map were conserved. *SAMtools* [67] and *BWA* [68] were used to handle and map reads respectively.

Functional profiles

Strain functional profiles are generated by assessing the gene coverage for KEGG [69] functional orthologs, taken from the KEGG Orthology database of molecular functions, represented in terms of orthology groups of genes [70]. Ortholog genes are the basic features used here, in each sample. Then, the representation of each feature of the profile in any sample is estimated from the number of reads mapping on it. These counts were obtained by mapping the reads that passed the filters against the integrated gene catalog of the human gut [71]. We used the NGLess built-in function *count* with the default values, applying the *scaled* normalization that consists of dividing the raw count by the size of the feature and scaled up so that the total number of counts is similar to the total raw count.

Taxonomic profiles

Strain taxonomic profiles were obtained using the *Centrifuge* application [72]. The *centrifuge-download* command was used to download the reference genomes of archea, bacteria, virus and vertebrate mammalian (human and mouse) taxons. The reads of each sample were mapped over the reference genomes. The taxonomic profile consists of the relative representation of each genome in the sample, which is obtained by normalizing the number of reads mapping on them by the respective lengths of the genomes.

Tumor status prediction

For tumor status prediction a combination of different machine learning methods was used with the aim of attaining a white-box interpretable model [34-36] that improve results previously reported. The white-box model is sequentially constructed by combining robust feature selection techniques with an interpretable yet powerful classification method in the form of Generalized Additive Models (GAM) [73]. GAMs have been used in many fields, including healthcare applications [74, 75], for predictive modeling due to being more flexible than a general linear model without losing the interpretability.

In order to classify a sample as *tumor* or *healthy* a new interpretability-based algorithm we have used: the Explainable Boosting Machine (EBM) [76], a GAM model that uses state-of-the-art machine learning methods, such as bagging and parallel gradient boosting, in order to learn a function for each feature. Note that GAMs are non-parametric learning algorithms with a high flexibility regarding to the data distribution, a very interpretable modeling, due to the additive effects, and the ability to learn complex non-linear relationships between the variables and the outcome. The EBM builds a boosting methodology where each feature is learned in a circular order with a learning rate low enough to avoid ordering-based

issues. These learning cycles allow the algorithm to learn a precise functional for each feature in a profile in such a way that is easy to extract how any of the features contribute towards a given prediction. By summarizing these contributions a global relevance score (i.e. a way to rank each profile feature) is obtained. The learning rate was optimized via ten-fold cross-validation.

Adenoma risk

In addition, from a healthcare point of view, one of the most interesting aspects of the predictions lies in the classifier ability to differentiate between adenomas and normal or CRC samples. In order to check the discriminative power of the classifier we proceeded as follows: (1) perform a split of the *binary* dataset with 30% of the samples in the test set, (2) aggregate the adenoma samples to the test set, (3) train the classifier on the train set and predict the new test set and finally, (4) compute the risk at *k* score defined as the sum the (signed) contributions of the top *k* features: (see Equation 1 in the Supplemental Files)

This risk function can be constructed for any *k*, but here *k* corresponds to the features selected by the feature selection method (see below).

Feature Selection

The feature selection pipeline is based on the biological hypothesis that there are complex relationships between the functional profiles and the tumor status. Since KEGG [77] orthologs, which correspond to biochemical reactions, are taken as features, there is an unbalance between the number of features (in the range of thousands) and the number of samples (in the range of hundreds)

In order to overcome such difficulties a multivariate feature selection technique based on the EBM feature ranking method (discussed below) was developed. The method is similar to the stability selection technique [78] in which a model is fitted across several bootstrapped versions of the data, thus having several rankings (each model produces a feature ranking score). The idea is to select those features that are ranked as top across the majority of the partitions. Instead of selecting the top ranked features, the distribution of the ranking scores of EBM is computed across all data partitions. Then a cut point based on the elbow point of the medians is selected. This method allows using the same multivariate model used for prediction, which has proven to perform well on high dimensional spaces with colinearties by better distributing the feature contributions of similar variables between them. The method is used on the global dataset (sample-wise union of all projects) using bootstrap to try to leverage the problems associated to the inter-project discrepancies.

Feature relevance ranking

To rank each feature EBM computes the feature-wise contribution for each sample in the training set, i.e. how each sample would be scored by a given feature. Then, for each feature the mean of the absolute value of the contributions is taken. Since EBM is an additive model, the ranked features correspond with those that have the most predictive power in the training set. Note that, given the additive properties of the model, the impact of a given feature on any sample can be directly observed. Actually, an

interpretability curve can be built, where the variation across samples of a given feature and its associated model contribution can be directly plotted and observed. The possibility of extracting feature-wise contributions from a complex multivariate classifier is a distinctive property of this approach that allows making more meaningful interpretations.

Feature relevance in the adenoma

In order to infer the relevance of the features for the Adenoma experiment we have used the SHapley Additive exPlanation (SHAP) [79] method: a novel additive feature attribution method that has been used with great success in many domains, such as hypoxaemia prevention during surgery [80] or NO_2 forecasting in ecology [81].

The SHAP method combines local explanation methods, as Local Interpretable Model-agnostic Explanations (LIME) [82], with game theory approaches (Shapley values). In broad terms the SHAP values estimate a given feature contribution by approximating the change between the expected and actual prediction when perturbing that feature. To build the expectation the method needs a set of background samples. In this case all the training samples we can safely be provided to the method since all the adenoma cases belong to the blind test set. Note that the SHAP method provides local explanations and, by combining them, a global relevance score can be built.

Batch effect removal

Batch effect is a known problem in gut metagenomic studies [28, 29]. Here, experimental effects were removed in the possible by filtering those features that contribute to perform a project-wise separation of the healthy samples. Then, project-specific effect removal is done in three steps: first, samples labeled as healthy across all the studies are selected, second, the feature selection pipeline is applied to the healthy subset using a multiclass approach, where the study is the target variable of the underlying classifier. Finally, those features labeled as relevant by the pipeline in all the studies are removed.

Model interpretability

As EBM belong to the additive family of machine learning algorithms, all features contribute to predictions in an adaptive, intelligible way. Thus, we can visualize how a given feature contributes to the prediction of any sample. We can summarize this information by visualizing the feature distribution along its contribution to the EBM prediction.

Software

The machine learning methodology has been implemented in Python 3.6 using the *scikit-learn* library [83] as the basic building block for the pipelines. To train EBM models we have used the *interpretml* package [76].

Declarations

Availability of data and material

The WGS of fecal metagenomic samples data that support the findings of this study are available in the European Nucleotide Archive [https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB10878, https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB27928, https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB6070, https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB7774, https://www.ebi.ac.uk/ena/browser/text-search?query=PRJNA447983].

Competing interests

The authors declare that they have no competing interests

Funding

This work is supported by grants SAF2017-88908-R from the Spanish Ministry of Economy and Competitiveness and “Plataforma de RecursosBiomoleculares y Bioinformáticos” PT17/0009/0006 from the ISCIII, both co-funded with European Regional Development Funds (ERDF) as well as H2020 Programme of the European Union grants Marie Curie Innovative Training Network "Machine Learning Frontiers in Precision Medicine" (MLFPM) (GA 813533) and “ELIXIR-EXCELERATE fast-track ELIXIR implementation and drive early user exploitation across the life sciences” (GA 676559).

Authors' contributions

CCS has carried out the bioinformatic analysis of the samples and contributed to the interpretation of the results; CL has carried out the machine learning analysis of the data; MPC has made the functional interpretation of the results; JD has conceived the work and wrote the paper.

References

1. Cho I, Blaser MJ: **The human microbiome: at the interface of health and disease**. *Nature Reviews Genetics* 2012, **13**(4):260.
2. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS: **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207.
3. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D: **A metagenome-wide association study of gut microbiota in type 2 diabetes**. *Nature* 2012, **490**(7418):55.
4. Findley K, Williams DR, Grice EA, Bonham VL: **Health disparities and the microbiome**. *Trends in microbiology* 2016, **24**(11):847-850.

5. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content**. *Nature genetics* 1999, **21**(1):108.
6. Zaneveld JR, Lozupone C, Gordon JI, Knight R: **Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives**. *Nucleic acids research* 2010, **38**(12):3869-3879.
7. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**(6978):37.
8. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N: **Shotgun metagenomics, from sampling to analysis**. *Nature biotechnology* 2017, **35**(9):833.
9. Börnigen D, Morgan XC, Franzosa EA, Ren B, Xavier RJ, Garrett WS, Huttenhower C: **Functional profiling of the gut microbiome in disease-associated inflammation**. *Genome medicine* 2013, **5**(7):65.
10. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N: **Strain-level microbial epidemiology and population genomics from shotgun metagenomics**. *Nature methods* 2016, **13**(5):435.
11. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG: **Strains, functions and dynamics in the expanded Human Microbiome Project**. *Nature* 2017, **550**(7674):61.
12. Lynch SV, Pedersen O: **The human intestinal microbiome in health and disease**. *New England Journal of Medicine* 2016, **375**(24):2369-2379.
13. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F: **Gut metagenome in European women with normal, impaired and diabetic glucose control**. *Nature* 2013, **498**(7452):99.
14. Bedarf JR, Hildebrand F, Coelho LP, Sunagawa S, Bahram M, Goeser F, Bork P, Wüllner U: **Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients**. *Genome medicine* 2017, **9**(1):39.
15. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z: **Gut microbiome development along the colorectal adenoma–carcinoma sequence**. *Nature communications* 2015, **6**:6528.
16. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N: **Potential of fecal microbiota for early-stage detection of colorectal cancer**. *Molecular systems biology* 2014, **10**(11).
17. Yu J, Feng Q, Wong SH, Zhang D, yi Liang Q, Qin Y, Tang L, Zhao H, Stenvang J, Li Y: **Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer**. *Gut* 2017, **66**(1):70-78.
18. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H: **The gut microbiome in atherosclerotic cardiovascular disease**. *Nature communications* 2017, **8**(1):845.
19. Pasolli E, Truong DT, Malik F, Waldron L, Segata N: **Machine learning meta-analysis of large metagenomic datasets: tools and biological insights**. *PLoS computational biology* 2016,

12(7):e1004977.

20. Cougnoux A, Dalmasso G, Martinez R, Buc E, Delmas J, Gibold L, Sauvanet P, Darcha C, Déchelotte P, Bonnet M: **Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype**. *Gut* 2014, **63**(12):1932-1942.
21. Wu S, Rhee K-J, Albesiano E, Rabizadeh S, Wu X, Yen H-R, Huso DL, Brancati FL, Wick E, McAllister F: **A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses**. *Nature medicine* 2009, **15**(9):1016.
22. Chung L, Orberg ET, Geis AL, Chan JL, Fu K, Shields CED, Dejea CM, Fathi P, Chen J, Finard BB: **Bacteroides fragilis toxin coordinates a pro-carcinogenic inflammatory cascade via targeting of colonic epithelial cells**. *Cell host & microbe* 2018, **23**(2):203-214. e205.
23. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, Clancy TE, Chung DC, Lochhead P, Hold GL: **Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment**. *Cell host & microbe* 2013, **14**(2):207-215.
24. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW: **Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin**. *Cell host & microbe* 2013, **14**(2):195-206.
25. Snoek J, Larochelle H, Adams RP: **Practical bayesian optimization of machine learning algorithms**. In: *Advances in neural information processing systems: 2012*. 2951-2959.
26. Baxter NT, Ruffin MT, Rogers MA, Schloss PD: **Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions**. *Genome medicine* 2016, **8**(1):37.
27. Zackular JP, Rogers MA, Ruffin MT, Schloss PD: **The human gut microbiome as a screening tool for colorectal cancer**. *Cancer prevention research* 2014, **7**(11):1112-1121.
28. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R: **Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer**. *Nature medicine* 2019, **25**(4):679.
29. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C: **Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation**. *Nature medicine* 2019, **25**(4):667.
30. Segata N: **On the road to strain-resolved comparative metagenomics**. *MSystems* 2018, **3**(2):e00190-00117.
31. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N: **Microbial strain-level population structure and genetic diversity from metagenomes**. *Genome research* 2017, **27**(4):626-638.
32. Gerner EW, Meyskens Jr FL: **Polyamines and cancer: old molecules, new understanding**. *Nature Reviews Cancer* 2004, **4**(10):781.
33. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, Srivastava M, Preece A, Julier S, Rao RM: **Interpretability of deep learning models: a survey of results**. In: *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable*

Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI): 2017. IEEE: 1-6.

34. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B: **Definitions, methods, and applications in interpretable machine learning**. *Proceedings of the National Academy of Sciences* 2019, **116**(44):22071-22080.
35. Chen L, Lu X: **Making deep learning models transparent**. *Journal of Medical Artificial Intelligence* 2018, **1**.
36. Michael KY, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T: **Visible machine learning for biomedicine**. *Cell* 2018, **173**(7):1562-1565.
37. Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, Lopatkin AJ, Satish S, Nili A, Palsson BO: **A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action**. *Cell* 2019, **177**(6):1649-1661. e1649.
38. Mason C, Afshinnekoo E, Ahsannudin S, Ghedin E, Read T, Fraser C, Dudley J, Hernandez M, Bowler C, Stolovitzky G: **The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report**. *MICROBIOME* 2016, **4**(1):24.
39. Casimiro-Soriguer CS, Loucera C, Perez Florido J, López-López D, Dopazo J: **Antibiotic resistance and metabolic profiles as functional biomarkers that accurately predict the geographic origin of city metagenomics samples**. *Biology direct* 2019, **14**(1):15-15.
40. Huang CJ, Chien CC, Yang SH, Chang CC, Sun HL, Cheng YC, Liu CC, Lin SC, Lin CM: **Faecal ribosomal protein L19 is a genetic prognostic factor for survival in colorectal cancer**. *Journal of cellular and molecular medicine* 2008, **12**(5b):1936-1943.
41. Scheppach W, Wehner F, Bartram P, Schramel P, Kasper H: **Metabolic and nutritional parameters in patients after colonic polypectomy**. *Digestion* 1988, **41**(2):94-100.
42. Park CH, Han DS, Oh Y-H, Lee A-r, Lee Y-r, Eun CS: **Role of Fusobacteria in the serrated pathway of colorectal carcinogenesis**. *Scientific reports* 2016, **6**:25271.
43. Dai Z, Coker OO, Nakatsu G, Wu WK, Zhao L, Chen Z, Chan FK, Kristiansen K, Sung JJ, Wong SH: **Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers**. *Microbiome* 2018, **6**(1):70.
44. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, Goedert JJ, Shi J, Bork P, Sinha R: **Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing**. *PloS one* 2016, **11**(5):e0155362.
45. Voigt AY, Costea PI, Kultima JR, Li SS, Zeller G, Sunagawa S, Bork P: **Temporal and technical variability of human gut metagenomes**. *Genome biology* 2015, **16**(1):73.
46. Kojima M, Morisaki T, Izuhara K, Uchiyama A, Matsunari Y, Katano M, Tanaka M: **Lipopolysaccharide increases cyclo-oxygenase-2 expression in a colon carcinoma cell line through nuclear factor-κB activation**. *Oncogene* 2000, **19**(9):1225.
47. Yoshioka T, Morimoto Y, Iwagaki H, Itoh H, Saito S, Kobayashi N, Yagi T, Tanaka N: **Bacterial Lipopolysaccharide Induces Transforming Growth Factor β and Hepatocyte Growth Factor through**

Tolllike Receptor 2 in Cultured Human Colon Cancer Cells. *Journal of International Medical Research* 2001, **29**(5):409-420.

48. Ng O: **Iron, microbiota and colorectal cancer**. *Wiener Medizinische Wochenschrift* 2016, **166**(13-14):431-436.
49. Evans RC, Fear S, Ashby D, Hackett A, Williams E, van der Vliet M, Dunstan FD, Rhodes JM: **Diet and colorectal cancer: an investigation of the lectin/galactose hypothesis**. *Gastroenterology* 2002, **122**(7):1784-1792.
50. Bleich RM, Arthur JC: **Revealing a microbial carcinogen**. *Science* 2019, **363**(6428):689-690.
51. Fazzari J, Linher-Melville K, Singh G: **Tumour-derived glutamate: linking aberrant cancer cell metabolism to peripheral sensory pain pathways**. *Current neuropharmacology* 2017, **15**(4):620-636.
52. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP: **Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults**. *PloS one* 2013, **8**(8):e70803.
53. Rombouts C, Hemeryck LY, Van Hecke T, De Smet S, De Vos WH, Vanhaecke L: **Untargeted metabolomics of colonic digests reveals kynurenine pathway metabolites, dityrosine and 3-dehydroxycarnitine as red versus white meat discriminating metabolites**. *Scientific reports* 2017, **7**:42514.
54. Huang A, Fuchs D, Widner B, Glover C, Henderson D, Allen-Mersh T: **Serum tryptophan decrease correlates with immune activation and impaired quality of life in colorectal cancer**. *British journal of cancer* 2002, **86**(11):1691.
55. Nishiumi S, Kobayashi T, Ikeda A, Yoshie T, Kibi M, Izumi Y, Okuno T, Hayashi N, Kawano S, Takenawa T: **A novel serum metabolomics-based diagnostic approach for colorectal cancer**. *PloS one* 2012, **7**(7):e40459.
56. Lukey MJ, Katt WP, Cerione RA: **Targeting amino acid metabolism for cancer therapy**. *Drug discovery today* 2017, **22**(5):796-804.
57. Helfinger V, Schroeder K: **Redox control in cancer development and progression**. *Molecular aspects of medicine* 2018, **63**:88-98.
58. Maharjan RP, Ferenci T: **The impact of growth rate and environmental factors on mutation rates and spectra in Escherichia coli**. *Environmental microbiology reports* 2018, **10**(6):626-633.
59. Burns MB, Lynch J, Starr TK, Knights D, Blekhman R: **Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment**. *Genome medicine* 2015, **7**(1):55.
60. Carter GP, Lyras D, Allen DL, Mackin KE, Howarth PM, O'connor JR, Rood JI: **Binary toxin production in Clostridium difficile is regulated by CdtR, a LytTR family response regulator**. *Journal of bacteriology* 2007, **189**(20):7290-7301.
61. Del Papa MF, Perego M: **Enterococcus faecalis virulence regulator FsrA binding to target promoters**. *Journal of bacteriology* 2011, **193**(7):1527-1532.

62. Umezawa Y, Yagisawa M, Sawa T, Takeuchi T, Umezawa H: **Aminoglycoside 3'-phosphotransferase III, a new phosphotransferase. Resistance mechanism**. *The Journal of antibiotics* 1975, **28**(11):845.
63. Beane J, Campbell JD, Lel J, Vick J, Spira A: **Genomic approaches to accelerate cancer interception**. *The lancet oncology* 2017, **18**(8):e494-e502.
64. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB: **Accessible, curated metagenomic data through ExperimentHub**. *Nature methods* 2017, **14**(11):1023.
65. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T: **Orchestrating high-throughput genomic analysis with Bioconductor**. *Nature methods* 2015, **12**(2):115.
66. Coelho LP, Alves R, Monteiro P, Huerta-Cepas J, Freitas AT, Bork P: **NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language**. *Microbiome* 2019, **7**(1):84.
67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
68. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.
69. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG**. *Nucleic Acids Res* 2014, **42**(Database issue):D199-205.
70. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic acids research* 2015, **44**(D1):D457-D462.
71. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T: **An integrated catalog of reference genes in the human gut microbiome**. *Nature biotechnology* 2014, **32**(8):834.
72. Kim D, Song L, Breitwieser FP, Salzberg SL: **Centrifuge: rapid and sensitive classification of metagenomic sequences**. *Genome research* 2016, **26**(12):1721-1729.
73. Hastie T, Tibshirani R: **Generalized additive models: some applications**. *Journal of the American Statistical Association* 1987, **82**(398):371-386.
74. Hastie T, Tibshirani R: **Generalized additive models for medical research**. *Statistical methods in medical research* 1995, **4**(3):187-196.
75. Fleury M, Charron DF, Holt JD, Allen OB, Maarouf AR: **A time series analysis of the relationship of ambient temperature and common bacterial enteric infections in two Canadian provinces**. *International journal of biometeorology* 2006, **50**(6):385-391.
76. Nori H, Jenkins S, Koch P, Caruana R: **InterpretML: A Unified Framework for Machine Learning Interpretability**. *arXiv preprint arXiv:190909223* 2019.
77. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: **KEGG: new perspectives on genomes, pathways, diseases and drugs**. *Nucleic acids research* 2016, **45**(D1):D353-D361.

78. Meinshausen N, Bühlmann P: **Stability selection**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010, **72**(4):417-473.
79. Lundberg SM, Lee S-I: **A unified approach to interpreting model predictions**. In: *Advances in Neural Information Processing Systems: 2017*. 4765-4774.
80. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J: **Explainable machine-learning predictions for the prevention of hypoxaemia during surgery**. *Nature biomedical engineering* 2018, **2**(10):749.
81. García MV, Aznarte JL: **Shapley additive explanations for NO2 forecasting**. *Ecological Informatics* 2019:101039.
82. Ribeiro MT, Singh S, Guestrin C: **Why should i trust you?: Explaining the predictions of any classifier**. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining: 2016*. ACM: 1135-1144.
83. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V: **Scikit-learn: Machine learning in Python**. *Journal of machine learning research* 2011, **12**(Oct):2825-2830.

Tables

Table 1. The ten most relevant features for the functional predictor.

KEGG ID	DEFINITION	MODULE	PATHWAY
K02850	heptose II phosphotransferase		ko00540 Lipopolysaccharide biosynthesis Ko01100 Metabolic pathways
K11708	manganese/zinc/iron transport system permease protein		ko02010 ABC transporters
K00917	tagatose 6-phosphate kinase		ko00052 Galactose metabolism ko01100 Metabolic pathways
K16168	methyltransferase		
K01846	methylaspartate mutase sigma subunit	M00740 Methylaspartate cycle	ko00630 Glyoxylate and dicarboxylate metabolism ko00660 C5-Branched dibasic acid metabolism ko01100 Metabolic pathways ko01120 Microbial metabolism in diverse environments ko01200 Carbon metabolism
K03653	N-glycosylase/DNA lyase		
K00702	cellobiose phosphorylase		ko00500 Starch and sucrose metabolism Ko01100 Metabolic pathways
K01556	kynureninase	M00038 Tryptophan metabolism, tryptophan => kynurenine => 2-aminomuconate	ko00380 Tryptophan metabolism ko01100 Metabolic pathways
K15584	nickel transport system substrate-binding protein		ko02010 ABC transporters
K04835	methylaspartate ammonia-lyase	M00740 Methylaspartate cycle	ko00630 Glyoxylate and dicarboxylate metabolism ko00660 C5-Branched dibasic acid metabolism ko01100 Metabolic pathways ko01120 Microbial metabolism in diverse environments ko01200 Carbon metabolism

Table 2. Conditions studied

Condition	Samples	Publication	NCBI ID
CRC	74	[17]	PRJEB10878
Healthy	54		
CRC	22	[28]	PRJEB27928
Healthy	60		
CRC	91	[16]	PRJEB6070
Large adenoma	15		
Small adenoma	27		
Healthy	66		
CRC	13	[15]	PRJEB7774
CRC; fatty liver	3		
CRC; fatty liver; hypertension	12		
CRC; hypertension	12		
Advanced adenoma	8		
Advanced adenoma; fatty liver	12		
Advanced adenoma; fatty liver; hypertension	14		
Advanced adenoma; hypertension	8		
Healthy	50		
CRC	47	[29]	PRJNA447983
CRC; cholesterolemia	1		
CRC; hypercholesterolemia	3		
CRC; hypertension	8		
CRC; hypertension; hypercholesterolemia	1		
CRC; metastases	1		
Adenoma	16		
Adenoma; hypercholesterolemia; metastases	1		
Adenoma; hypertension	4		
Adenoma; metastases	1		
Healthy	28		
Hypercholesterolemia	5		
Hypertension	3		
Hypertension; metastases	1		

Supplementary File Legends

Supplementary Figure 1. Risk score distribution across projects. Distribution of the risk score across the healthy, CRC and adenoma samples across the different projects used in this study. Labels are as follows: AdvanceA: Advaced Adenoma; Adenoma+: Adenoma with comorbilities; LargeA: Large Adenoma; SmallA: Small Adenoma; CRC+: CRC with comorbilities; Metastases: CRC with metastases.

Supplementary Figure 2. Most relevant KEGG pathways represented over the general metabolic KEGG pathway.

Supplementary Figure 3. Hierarchical relationships between the most relevant KEGG modules in which the most relevant functional features selected by the classification model map are plotted according to their hierarchical relationships.

Supplementary Figure 4. Phylogeny of the most relevant taxonomic features selected by the classification model arranged according their phylogenetic relationship. Values in red at the tips of the phylogeny are the relevance of the strains in the classification.

Supplementary figure 5. **Plots of feature interpretability of the most relevant features for the classification of adenomas**. The vertical axis shows the feature model scoring of samples (top graph) and the feature real distribution (bottom graph). Contributions above zero denote that the selected feature scores the samples towards the positive class (adenoma in this case) at the values indicated in the horizontal axis, whereas contributions below zero denote a score towards the negative class (being healthy). **Large adenoma most relevant features**: A) K02885, *large subunit ribosomal protein L19e*, B) K00927 *phosphoglycerate kinase* [EC:2.7.2.3]; C) K00575 *chemotaxis protein methyltransferase CheR* [EC:2.1.1.80]; D) K00150 *glyceraldehyde-3-phosphate dehydrogenase (NAD(P))* [EC:1.2.1.59]; E) K02849 *heptosyltransferase III* [EC:2.4.-.-]; F) K07480 *insertion element IS1 protein InsB*. **Small adenoma most relevant features**: G) K07707 *two-component system, LytTR family, response regulator AgrA*; H) K00620 *glutamate N-acetyltransferase / amino-acid N-acetyltransferase* [EC:2.3.1.352.3.1.1]; I) K19300 *aminoglycoside 3'-phosphotransferase II* [EC:2.7.1.95]; J) K08086 *pilus assembly protein FimV*; K) K00101 *L-lactate dehydrogenase (cytochrome)* [EC:1.1.2.3]; L) K00441 *coenzyme F420 hydrogenase subunit beta* [EC:1.12.98.1];

Supplementary Figure 6. Correlations between the presence of the most relevant functional and taxonomic features across samples of individual experiments: A) PRJNA447983; B) PRJEB27928; C) PRJEB12449; D) PRJEB10878; E) PRJEB7774; F) PRJEB6070

Supplementary Table 1. Most relevant functional features (KEGG orthologs) selected by the classification model

Supplementary Table 2. Most relevant KEGG pathways in which the most relevant functional features selected by the classification model map.

Supplementary Table 3. Most relevant KEGG modules in which the most relevant functional features selected by the classification model map.

Supplementary Table 4. Most relevant taxonomic features selected by the classification model

Supplementary Table 5. Most relevant taxonomic features for the classification of adenoma damples selected by the SHAP method

Figures

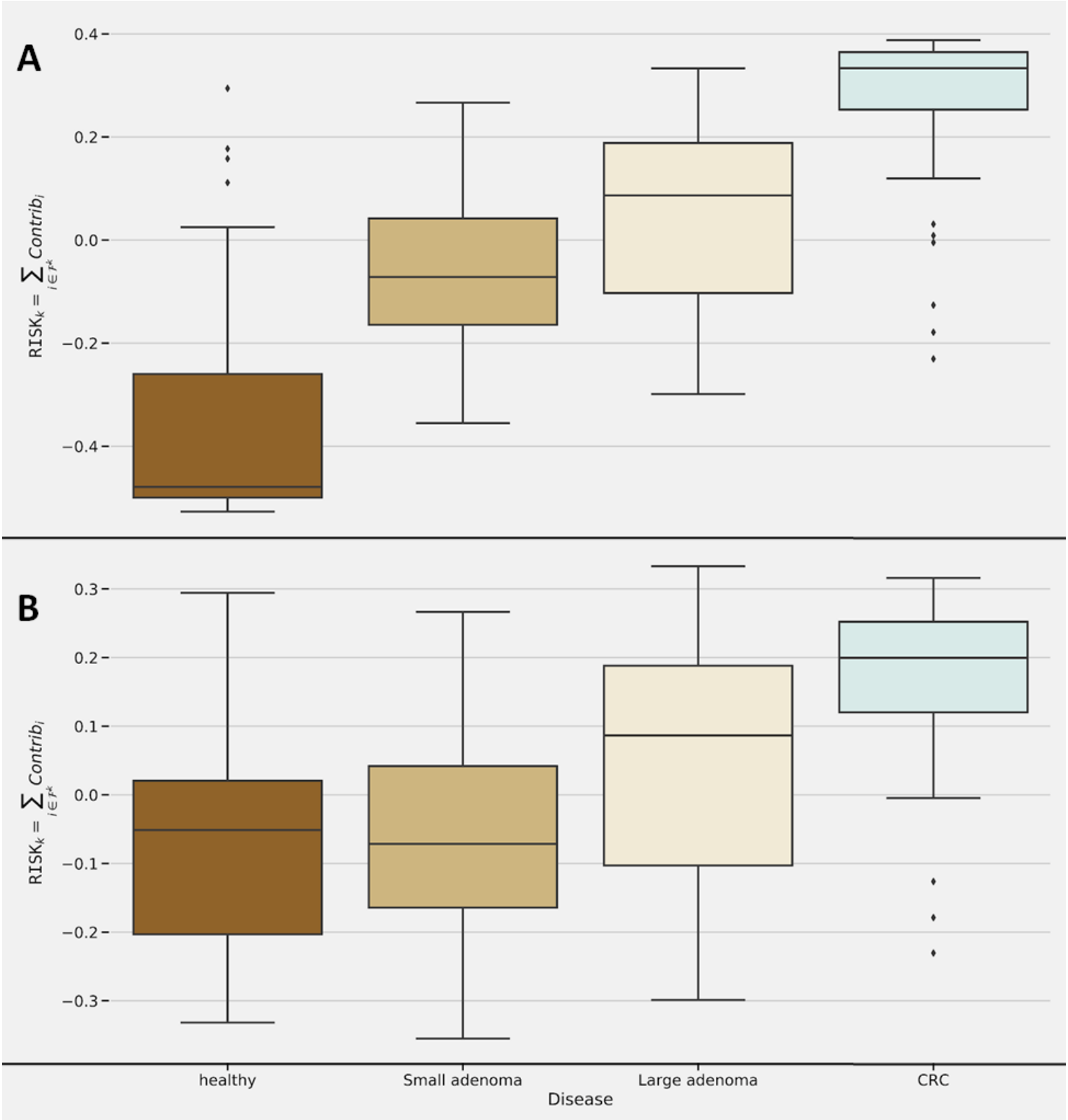

Figure 1

Distribution of the risk score across the healthy, CRC and adenoma samples using functional features in the A) training and B) test splits

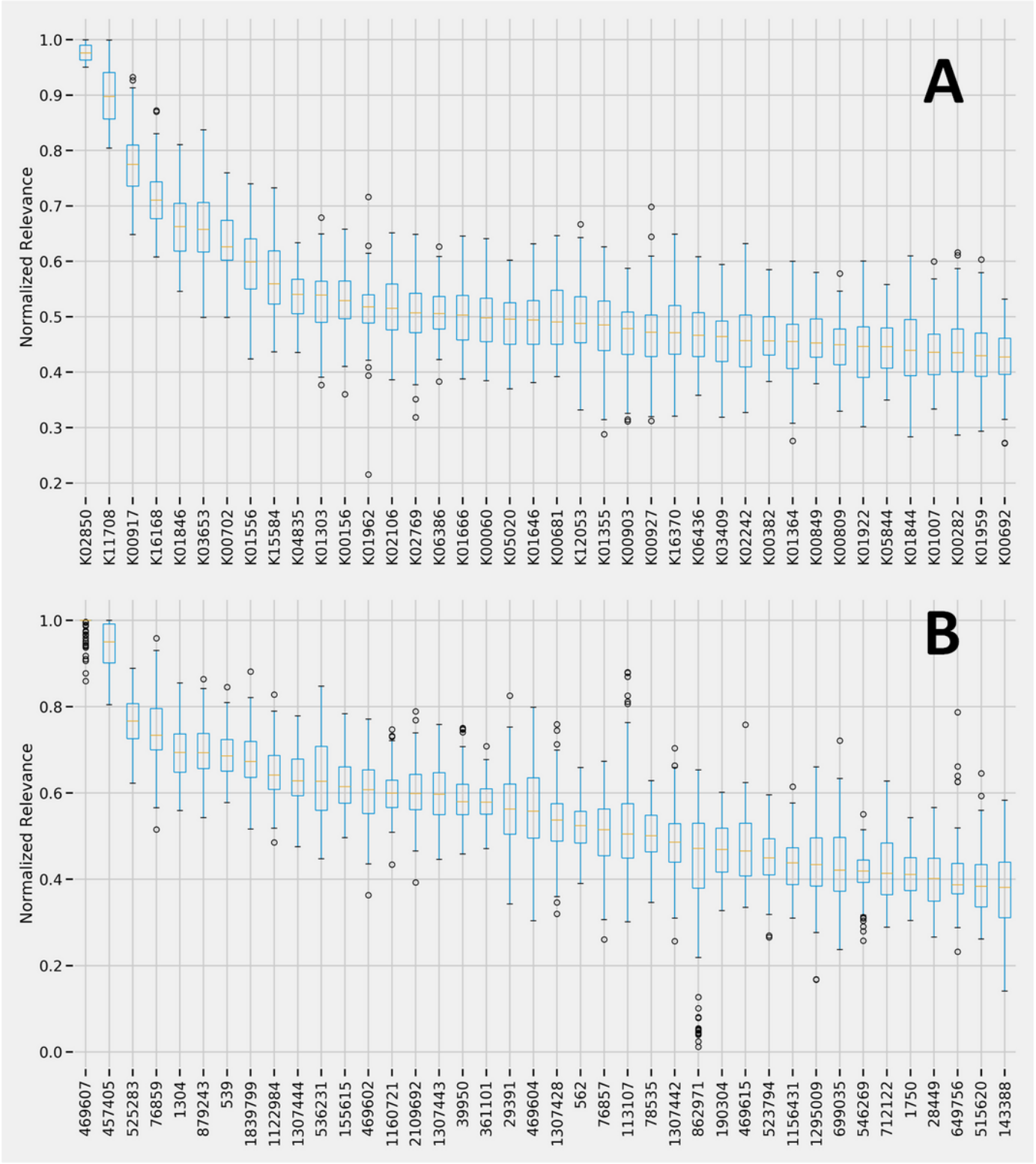

Figure 2

Visual representation of the relevance score distribution for A) the functional and B) taxonomic features. Note that for visualization purposes only the top forty features are shown.

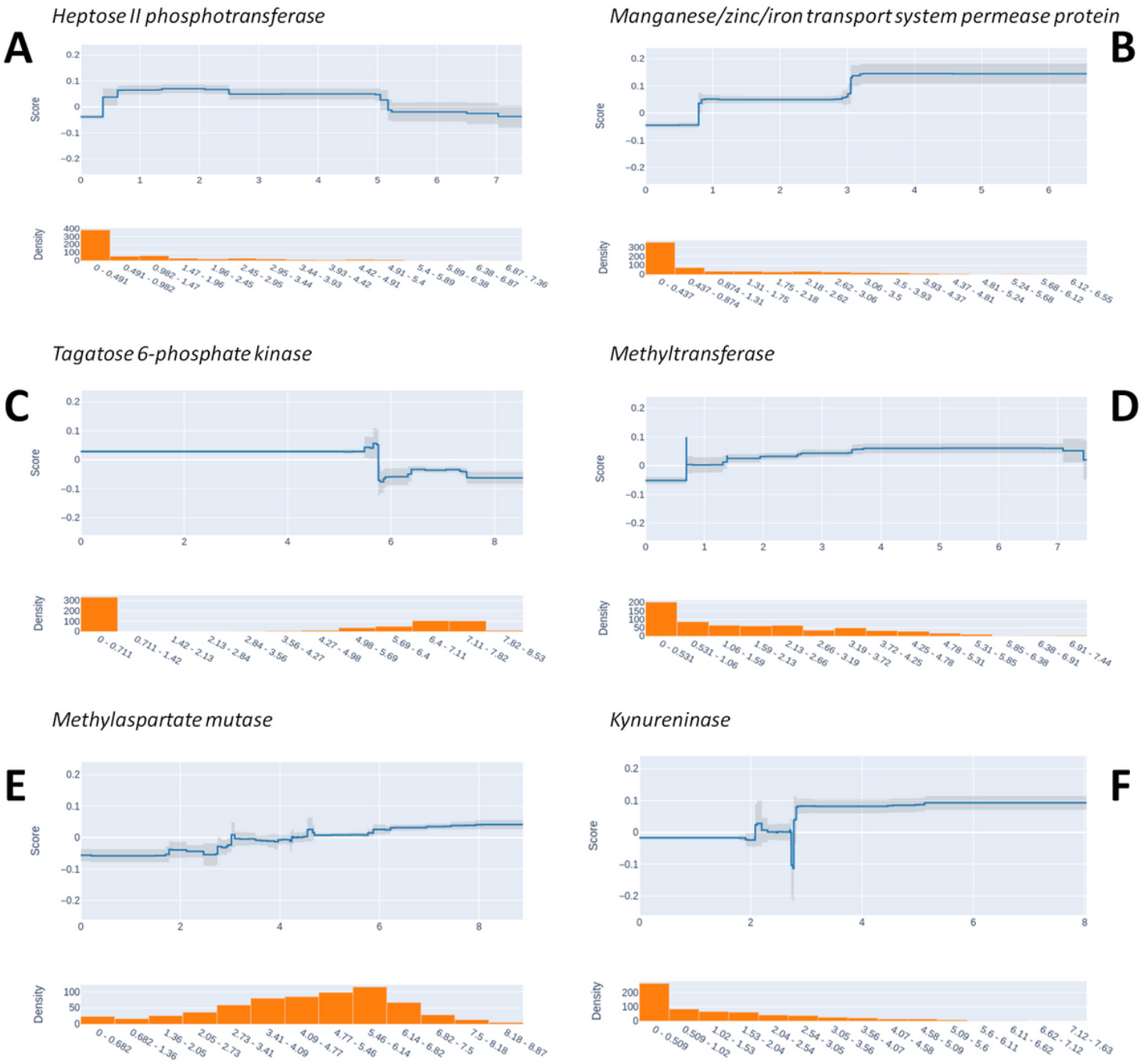

Figure 3

Plot of feature interpretability. The vertical axis shows the feature model scoring of samples (top graph) and the feature real distribution (bottom graph). Contributions above zero denote that the selected feature scores the samples towards the positive class (tumor in our case) at the values indicated in the horizontal axis, whereas contributions below zero denote a score towards the negative class (being healthy). Interpretability of the six most relevant functional features: A) Heptose II phosphotransferase, B) Manganese/zinc/iron transport system permease protein, C) Tagatose 6-phosphate kinase, D) Methyltransferase, E) Methylaspartate mutase sigma subunit, and F) Kynureninase.

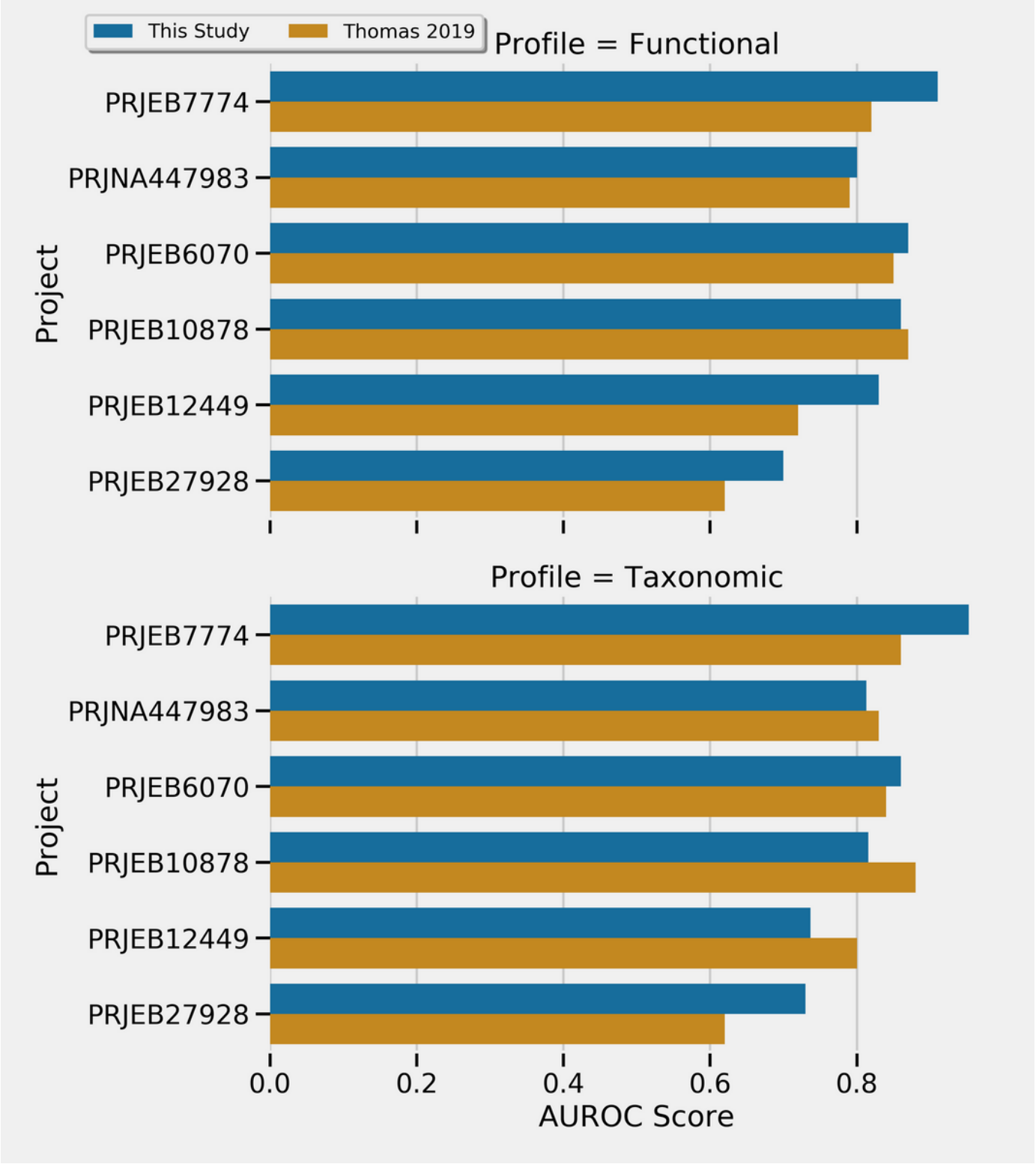

Figure 4

Comparison between the ML approach presented in [29] and the approach proposed here using both functional and taxonomic profiles.

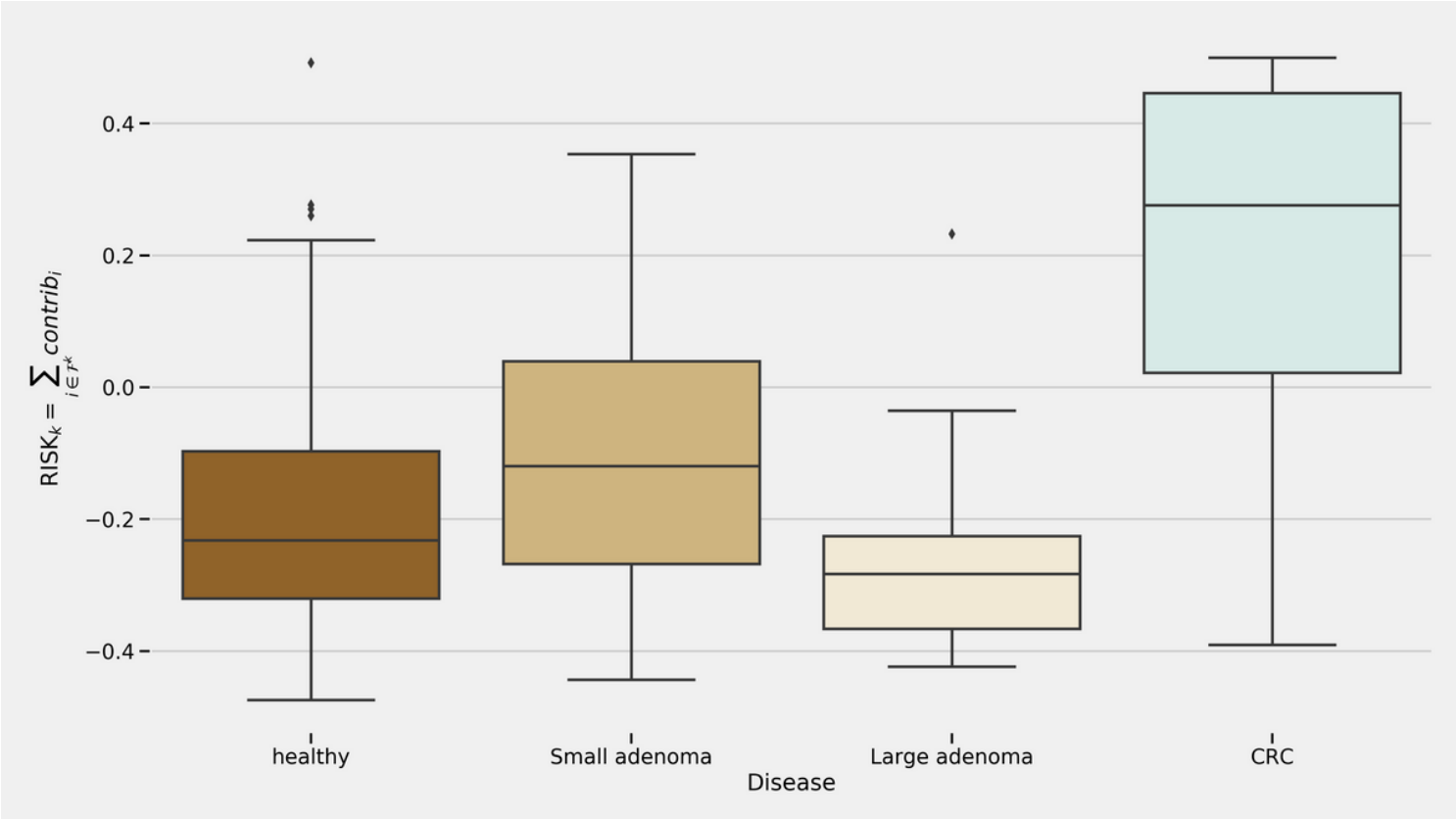

Figure 5

Distribution of risk score across the healthy, CRC and adenoma samples using taxonomic features.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Equation1.jpg
- SupplementaryTable1.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable5.xlsx
- SupplementaryFigures.pdf
- SupplementaryTable3.xlsx
- SupplementaryTable2.xlsx