bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

Jakob Wirbel
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
jakob.wirbel@embl.de ORCID: 0000-0002-4073-3562

Konrad Zych
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
konrad.zych@embl.de
Present address: Clinical Microbiomics A/S, Ole Maaløes Vej 3, 2200 København, Danmark

Morgan Essex
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
morgan.essex@mdc-berlin.de ORCID: 0000-0001-8758-7497
Present address: Experimental and Clinical Research Center (ECRC) of the Max Delbrück Center for Molecular Medicine and Charité University Hospital, 13125 Berlin, Germany

Nicolai Karcher
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
Department CIBIO, University of Trento, Trento 38123, Italy
nicolai.karcher@embl.de ORCID: 0000-0001-7894-8182

Ece Kartal
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
ece.kartal@embl.de ORCID: 0000-0002-7720-455X

Guillem Salazar
Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich 8093, Switzerland
guillems@ethz.ch ORCID: 0000-0002-9786-1493

Peer Bork
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
Molecular Medicine Partnership Unit, Heidelberg, Germany
Max Delbrück Centre for Molecular Medicine, Berlin, Germany
Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany
bork@embl.de ORCID: 0000-0002-2627-833X

Shinichi Sunagawa
Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich 8093, Switzerland
ssunagawa@ethz.ch ORCID 0000-0003-3065-0314

Georg Zeller
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstr 1, 69117 Heidelberg, Germany
zeller@embl.de ORCID 0000-0003-1429-7485

Correspondence should be addressed to Georg Zeller: zeller@embl.de

Keywords
Microbiome Data Analysis, Machine Learning, Statistical Modeling, Microbiome-wide Association Studies (MWAS), Meta-analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

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

Disease	Disease abbr.	Datasets	Data type
Ankylosing spondylitis	AS	[7]	Shotgun
Rheumatoid arthritis	ART	[25]	16S rRNA
		[6]	Shotgun
Type 1 diabetes	T1D	[26]	16S rRNA
Crohn's disease	CD	[5,27–30]	Shotgun
Ulcerative colitis	UC	[5,27,30]	Shotgun
Inflammatory bowel disease	IBD	[4,31–33]	16S rRNA
Colorectal cancer	CRC	[8,34–39]	Shotgun
		[8,40–42]	16S rRNA
Advanced colorectal adenoma(s)	ADA	[8,34,38,39]	Shotgun
Atherosclerotic cardiovascular disease	ACVD	[43]	Shotgun
Hypertension Pre-hypertension	HT pHT	[44]	Shotgun
Clostridioides difficile infection	CDI	[45,46]	16S rRNA
enteric diarrheal disease	EDD	[47]	16S rRNA
HIV infection	HIV	[48–50]	16S rRNA
Liver cirrhosis	LIV	[51]	Shotgun
		[52]	16S rRNA
Non-alcoholic fatty liver disease	NAFLD	[53,54]	Shotgun
		[55,56]	16S rRNA
Parkinsons' disease	PAR	[57]	Shotgun
		[58]	16S rRNA
Autism spectrum disorder	ASD	[59,60]	16S rRNA
Obesity	OB	[61]	Shotgun
		[62–65]	16S rRNA
Metabolic syndrome	MS	[66]	Shotgun
Type 2 diabetes	T2D	[67,68]	Shotgun
Impaired glucose tolerance	IGT	[67]	Shotgun

Results

Machine learning and statistical analysis workflows implemented in SIAMCAT

The SIAMCAT R package is a versatile toolbox for analysing microbiome data from case-control studies. The default workflows abstract from and combine many of the complex steps that these workflows entail and that can be difficult to implement correctly for non-experts. To increase ease of use, SIAMCAT interfaces with the popular phyloseq package [69] and

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

design and parameter choices are carefully adapted to metagenomic data analysis. In addition to functions for statistical testing of associations, SIAMCAT workflows include ML procedures, including data preprocessing, model fitting, performance evaluation and visualization of the results and models (**Figure 1a**). Core ML functionality is based on the mlr package [70]. The input for SIAMCAT consists of a feature matrix (abundances of microbial taxa, genes or pathways across all samples), a group label (case-control information for all samples), and optional meta-variables (such as demographics, lifestyle, and clinical records of sample donors or technical parameters of data acquisition).

To demonstrate the main workflow and primary outputs of the SIAMCAT package (see Methods and **SIAMCAT vignettes**), we analysed a representative dataset [27] consisting of 128 fecal metagenomes from patients with ulcerative colitis (UC) and non-UC controls (**Figure 1**). UC is a subtype of inflammatory bowel disease (IBD), a chronic inflammatory condition of the gastrointestinal tract that has been associated with dramatic changes in the gut microbiome [5,71]. As input, we used species-level taxonomic profiles available through the *curatedMetagenomicsData* R package [72].

After data preprocessing (unsupervised abundance and prevalence filtering, **Figure 1a** and Methods), univariate associations of single species with the disease are computed using the nonparametric Wilcoxon test (which has been shown for metagenomic data to reliably control the false discovery rate in contrast to many other tests proposed [73]) and results are visualized (using the *check.associations* function). The association plot displays the distribution of microbial relative abundance, the significance of the association, and a generalized fold change as a non-parametric measure of effect size [37] (**Figure 1b**).

The central component of SIAMCAT consists of ML procedures, which include a selection of normalization methods (*normalize.features*), functionality to set up a cross validation scheme (*create.data.split*), and interfaces to different ML algorithms, such as LASSO, Elastic Net, and Random Forest (offered by the mlr package [70]) [74–76]. As part of the cross-validation procedure, models can be trained (*train.model*) and applied to make predictions (*make.predictions*) on samples not used for training. Based on these predictions, the performance of the model is assessed (*evaluate.predictions*) using the area under the receiver operating characteristic (ROC) curve (AUROC) (**Figure 1d**). SIAMCAT also provides diagnostic plots for the interpretation of ML models (*model.interpretation.plot*) which display the importance of individual features in the classification model, normalized feature distributions as heatmaps, next to sample meta-variables (optionally, see **Figure 1 c,e**).

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

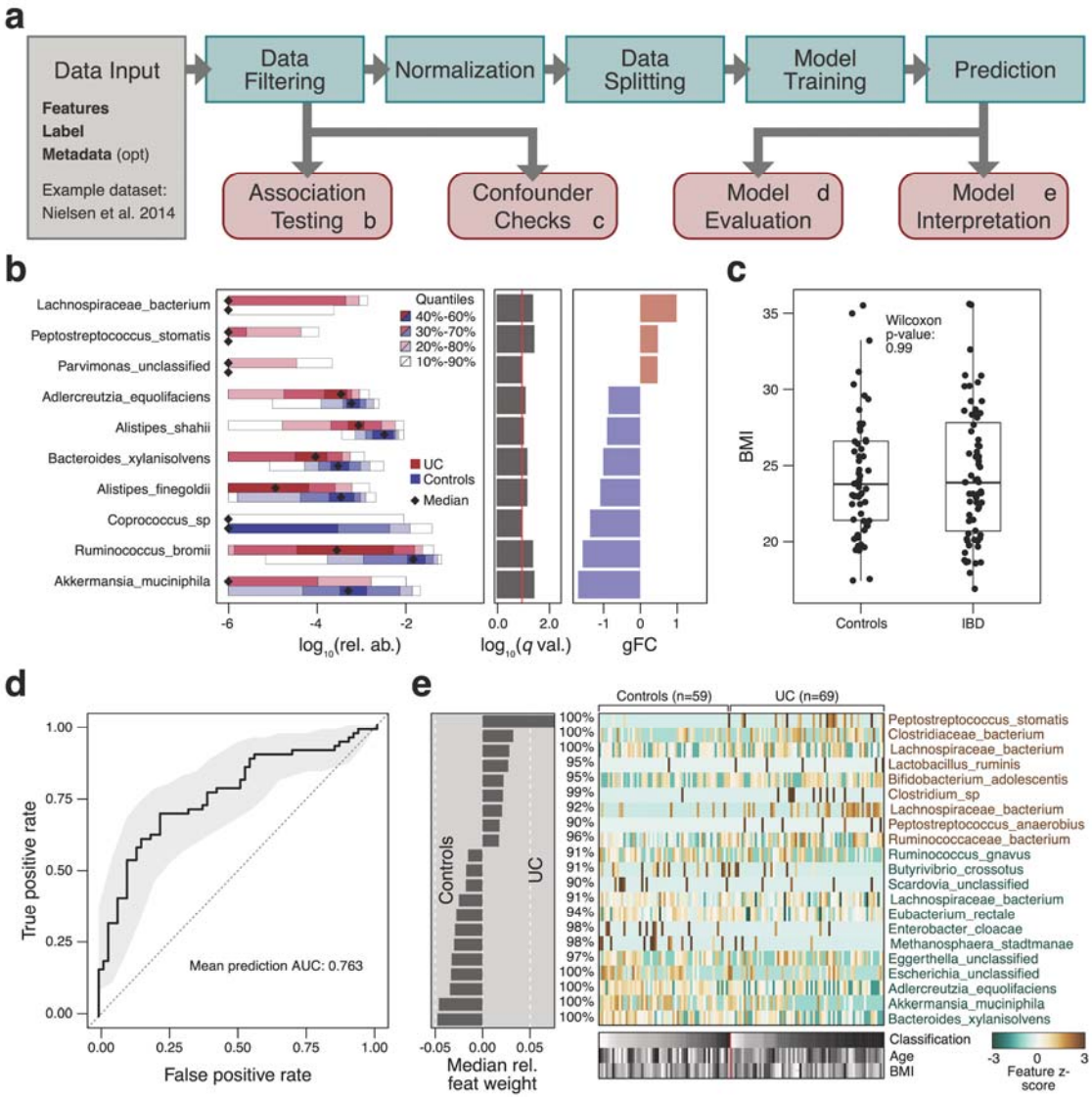


Figure 1. SIAMCAT statistical and machine learning approaches model differences between groups of microbiome samples. (a) Each step in the SIAMCAT workflow (green boxes) is implemented by a function in the R/Bioconductor package (see **SIAMCAT vignettes**). Functions producing graphical output (red boxes) are illustrated in (b)-(e) for an exemplary analysis using a data set from Nielsen et al. [27] which contains ulcerative colitis (UC) patients and non-UC controls. **(b)** Visualisation of the univariate association testing results. The left panel visualizes the distributions of microbial abundance data differing significantly between groups. Significance (after multiple testing correction) is displayed in the middle panel as horizontal bars. The right panel shows the generalized fold change as a nonparametric measure of effect size [37]. **(c)** SIAMCAT offers statistical tests and diagnostic visualisations to identify potential confounders by testing for associations between such meta-variables as covariates and the disease label. The example shows a comparison of body mass index (BMI) between the study groups. The similar distributions between cases and controls suggests that BMI is unlikely to confound UC associations in this dataset. Boxes denote the IQR across all values with the median as a thick black line and the whiskers extending up to the most extreme points within 1.5-fold IQR. **(d)** The model evaluation function displays the cross validation error as a receiver operating characteristic (ROC) curve, with a 95% confidence interval shaded in gray and the area under the receiver operating characteristic curve (AUROC) given below the curve. **(e)** SIAMCAT finally generates visualizations aiming to facilitate the interpretation of the machine learning models and their classification performance. This includes a barplot of feature importance (in the case of penalized logistic regression models, bar width corresponds to coefficient values) for the features that are included in the majority of models fitted during cross validation (percentages indicate the respective fraction

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

of models containing a feature). A heatmap displays their normalized values across all samples (as used for model fitting) along with the classification result (test predictions) and user-defined meta-variables (bottom).

Confounder analysis using SIAMCAT

As many biological and technical factors beyond the primary phenotype of interest can influence microbiome composition [1], microbiome association studies are often at a high risk of confounding, which can lead to spurious results [11,77–79]. To minimize this risk, SIAMCAT provides a function to optionally examine potential confounders among the provided meta-variables. In the example dataset from [27], control samples were obtained from both Spanish as well as Danish subjects, while UC samples were only taken from Spanish individuals (**Figure 2a**). Here, the meta-variable "country" could be viewed as a surrogate variable for other (often difficult-to-measure) factors, which can influence microbiome composition, such as diet, lifestyle, or technical differences between studies. The strong association of the "country" meta-variable with the disease status (SIAMCAT computes such associations using Fisher's exact test or the Wilcoxon test for discrete and continuous meta-variables, respectively; see **Figure 2a**) hints at the possibility that associations computed with the full dataset could be confounded by the country of the sample donor.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In summary, SIAMCAT can help to detect influential confounding factors that have the potential to bias statistical associations and ML model performance (see **Supplementary Figure 2** for additional examples).

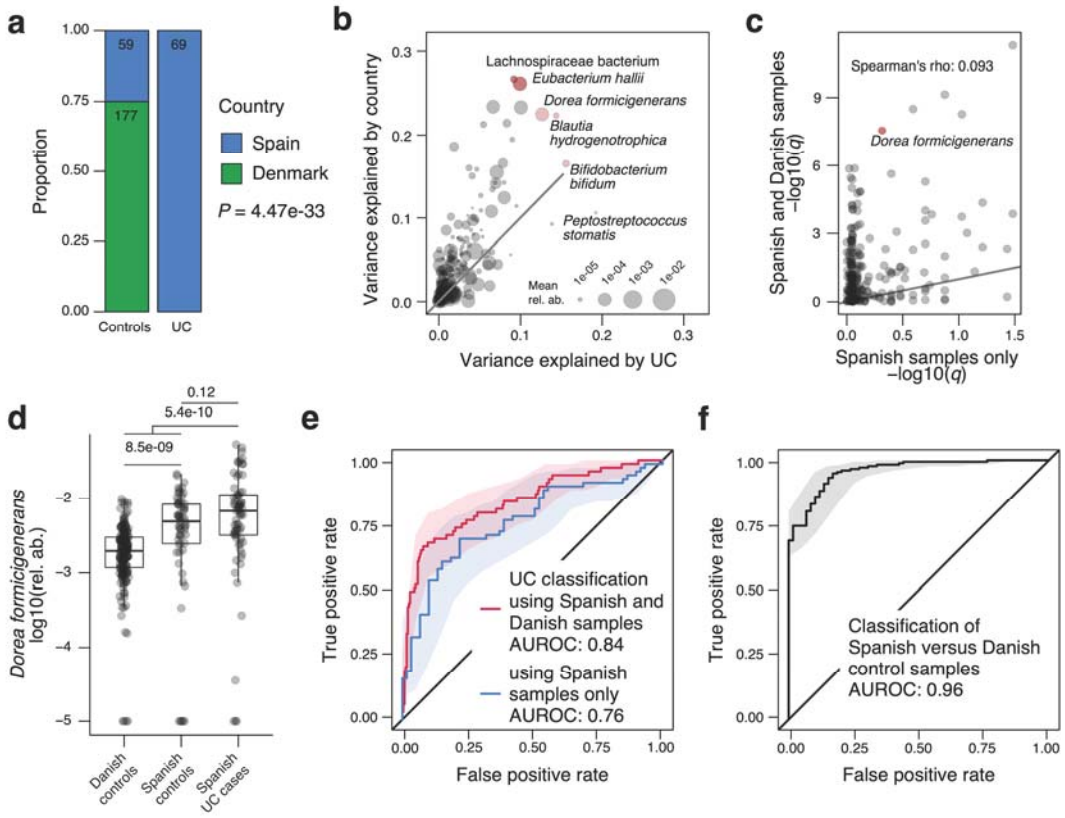


Figure 2. Analysis of covariates that potentially confound microbiome-disease associations and classification models. The UC data set from Nielsen et al. [27] contains fecal metagenomes from subjects enrolled in two different countries and generated using different experimental protocols (data shown is from *curatedMetagenomicData* with CD cases and additional samples per subject removed). **(a)** Visualisations from the SIAMCAT confounder checks reveals that only control samples were taken from Denmark suggesting that any (biological or technical) differences between Danish and Spanish samples might confound a naive analysis for UC-associated differences in microbial abundances. **(b)** Analysis of variance (using ranked abundance data) shows many species to differ more by country than by disease, with several extreme cases highlighted. **(c)** When comparing (FDR-corrected) *P*-values obtained from SIAMCAT's association testing function applied to the whole data set (y-axis) to those obtained for just the Danish samples (x-axis), only a very weak correlation is seen and strong confounding becomes apparent for several species including *Dorea formicigenerans* (highlighted). **(d)** Relative abundance differences for *Dorea formicigenerans* are significantly larger between countries than between Spanish UC cases and controls (*P*-values from Wilcoxon test). See Figure 1C for definition of boxplots. **(e)** Distinguishing UC patients from controls with the same workflow is possible with lower accuracy when only samples from Spain are used compared to the full dataset containing Danish and Spanish controls. This implies that in the latter case the machine learning model is confounded as it exploits the (stronger) country differences (see (c) and (f)), not only UC-associated microbiome changes. **(f)** This is confirmed by the result that control samples from Denmark and Spain can be very accurately distinguished with an AUROC of 0.96 (using SIAMCAT classification workflows).

Advanced machine learning workflows

When designing more complex ML workflows involving feature selection steps or applications to time series data, it becomes more challenging to set up cross validation

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

procedures correctly. Specifically, it is important to estimate how well a trained model would generalize to an independent test set, which is typically a main objective of microbial biomarker discovery. An incorrect ML procedure, in which information leaks from the test to the training set, can result in overly optimistic (i.e. overfitted) performance estimates. Two pitfalls that can lead to overfitting and poor generalization to other datasets (**Figure 3a**) are frequently encountered in ML analyses of microbiome and other biological data, even though the issues are well described in the statistics literature [80–82]. These issues, namely supervised feature filtering and naive splitting of dependent samples, can be exposed by testing model performance in an external validation set, which has not been used during cross validation at all (**Figure 3b**).

The first issue arises when feature selection taking label information into account (supervised feature selection) is naively combined with subsequent cross validation on the same data [80]. This incorrect procedure selects features that are associated with the label (e.g. by testing for differential abundance) on the complete dataset leaving no data aside for an unbiased test error estimation of the whole ML procedure. To avoid overfitting, correct supervised feature selection should always be nested into cross validation (that is, the supervised feature selection has to be applied to each training fold of the cross validation separately). To illustrate the extent of overfitting resulting from the incorrect approach, we used two datasets of colorectal cancer (CRC) patients and controls and performed both the incorrect and correct way of supervised feature selection. As expected, the incorrect feature selection led to inflated performance estimates in cross validation but lower generalization to an external dataset, whereas the correct procedure gave a better estimate of the performance in the external test set; the fewer features were selected, the more the performance in the external datasets dropped (see **Figure 3c**). SIAMCAT readily provides implementations of the correct procedure and additionally takes care that the feature filtering and normalization of the whole data set are blind to the label (therefore called unsupervised), thereby preventing accidental implementation of the incorrect procedure.

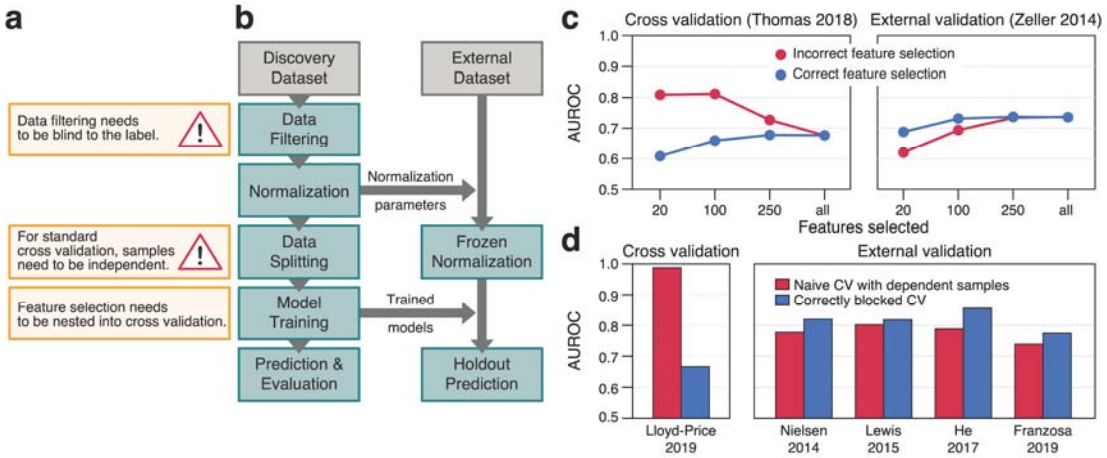


Figure 3. SIAMCAT aids in avoiding common pitfalls leading to poor generalisation of machine learning models. (a) Incorrectly set up machine learning workflows can lead to overoptimistic accuracy estimates (overfitting): the first issue arises from a naive combination of feature selection on the whole data set and subsequent cross validation on the very same data [82]. The second one arises when samples that were not taken independently (as is the case for replicates or samples taken at multiple time points from the same subject) are randomly partitioned in cross validation with the aim to assess the cross-subject generalisation error (see Main text). **(b)** External validation, for which SIAMCAT offers analysis workflows, can expose these issues. The individual steps in the workflow diagram correspond to SIAMCAT functions for fitting a machine learning model

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and applying it to an external data set to assess its external validation accuracy (see SIAMCAT vignette: Holdout testing with SIAMCAT). **(c)** External validation shows overfitting to occur when feature selection and cross validation are combined incorrectly in a sequential manner, rather than correctly in a nested approach. The correct approach is characterized by a lower (but unbiased) cross-validation accuracy, but better generalisation accuracy to external data sets (see header for data sets used). The fewer features are selected, the more pronounced the issue becomes and in the other extreme case ('all'), feature selection is effectively switched off. **(d)** When dependent observations (here by sampling the same individuals at multiple time points) are randomly assigned to cross-validation partitions, effectively the ability of the model to generalise across time points, but not across subjects is assessed. To correctly estimate generalisation accuracy across subjects, repeated measurements need to be blocked, all of them either into the training or test set. Again the correct procedure shows lower cross-validation accuracy, but higher external validation accuracy.

The second issue tends to occur when samples are not independent [81]. For example, microbiome samples taken from the same individual at different time points are usually a lot more similar to each other than those from different individuals (see [12] and **Supplementary Figure 3**). If these dependent samples are randomly split in a standard cross validation procedure, so that some could end up in the training set and others in the test set, it is effectively estimated how well the model generalizes across time points (from the same individual) rather than across individuals. To avoid this, dependent measurements need to be blocked during cross validation, ensuring that measurements of the same individual are assigned to the same test set. How much the naive procedure can overestimate the performance in cross validation and underperform in external validation compared to the correctly blocked procedure is demonstrated here using the iHMP dataset, which contains several samples per subject [30]. Although the cross validation accuracy appears dramatically lower in the correct compared to the naive procedure, generalization to other datasets of the same disease is higher with the correctly blocked model (**Figure 3d**). SIAMCAT offers the possibility to block the cross validation according to meta-variables by simply providing an additional argument to the respective function call (see also **SIAMCAT vignettes**).

Large-scale machine learning meta-analysis

Previous studies that applied ML to microbiome data [17–20] have compared and discussed the performance of several learning algorithms. However, their recommendations were based on the analysis of a small number of data sets which were technically relatively homogeneous. To overcome this limitation and to demonstrate that SIAMCAT can readily be applied to various types of input data, we performed a large-scale ML meta-analysis of case-control gut metagenomic datasets. We included taxonomic profiles obtained with the RDP taxonomic classifier [83] for 26 data sets based on 16S rRNA gene sequencing [20]; additionally, taxonomic profiles generated from 12 and 24 shotgun metagenomic data sets using either MetaPhlAn2 [84] or mOTUs2 [85], respectively, as well as functional profiles obtained with HUMAnN2 [86] or with eggNOG 4.5 [87] for the same set of shotgun metagenomic data were included (in total 130 classification tasks, see **Table 1** and **Supplementary Table 1** for information about included datasets).

Focusing first on intra-study results, we found that given a sufficiently large input data set (with at least 100 samples), SIAMCAT models are generally able to distinguish reasonably well between cases and controls: the majority (58%) of these datasets in our analysis could be classified with an AUROC of 0.75 or higher - compared to only 36% of datasets with fewer than 100 samples (**Figure 4a-c, Supplementary Figure 4 and 5** and Methods). Of note, accurate ML-based classification was possible even for datasets in which cases and controls could not easily be separated using beta-diversity analyses (**Supplementary**

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 6), indicating that a lack of separation in ordination analysis does not preclude ML-based workflows to extract accurate microbiome signatures. In the datasets for which a direct comparison of mOTUs2 and MetaPhlAn2 profiles was possible, we did not find any consistent trend towards either profiling method (paired Wilcoxon P = 0.41, see **Supplementary Figure 7**). When comparing taxonomic and functional profiles derived from the same dataset, we found a high correlation between AUROC values (Pearson's r = 0.92, $P < 2*10^{-16}$), although on average taxonomic profiles performed slightly better than functional profiles (**Supplementary Figure 7**). Taken together this indicates that SIAMCAT can extract accurate microbiome signatures from a range of different input profiles commonly used in microbiome research.

SIAMCAT provides various methods for data filtering and normalization and interfaces to several ML algorithms through mlr [70]. This made it easy to explore the space of possible workflow configurations in order to arrive at recommendations about sensible default parameters. To test the influence of different parameter choices within the complete data analysis pipeline, we performed an ANOVA analysis to quantify their relative importance on the resulting classification accuracy (**Figure 4d** and Methods). Whereas the choice of filtering method and feature selection regime has little influence on the results, the normalization method and ML algorithm explained more of the observed variance in classification accuracy. Analysis of the different normalization methods shows that most of the differences can be explained by a drop in performance for naively normalized data (only total sum scaling and no further normalization) in combination with LASSO or Elastic Net logistic regression (**Supplementary Figure 8**). In contrast, the Random Forest classifier depended much less on optimal data normalization. Lastly, we compared the best classification accuracy for each classification task across the different ML algorithms. Interestingly, in contrast to a previous report [19], this analysis indicates that on average Elastic Net logistic regression outperforms LASSO and Random Forest classifiers when considering the optimal choice of ML algorithm (P = 0.001 comparing Elastic Net to LASSO and $P = 4*10^{-14}$ comparing it to Random Forest, **Figure 4e**). In summary, this large-scale analysis demonstrates the versatility of the ML workflows provided by SIAMCAT and validates its default parameters as well as the robustness of classification accuracy to deviations from these.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

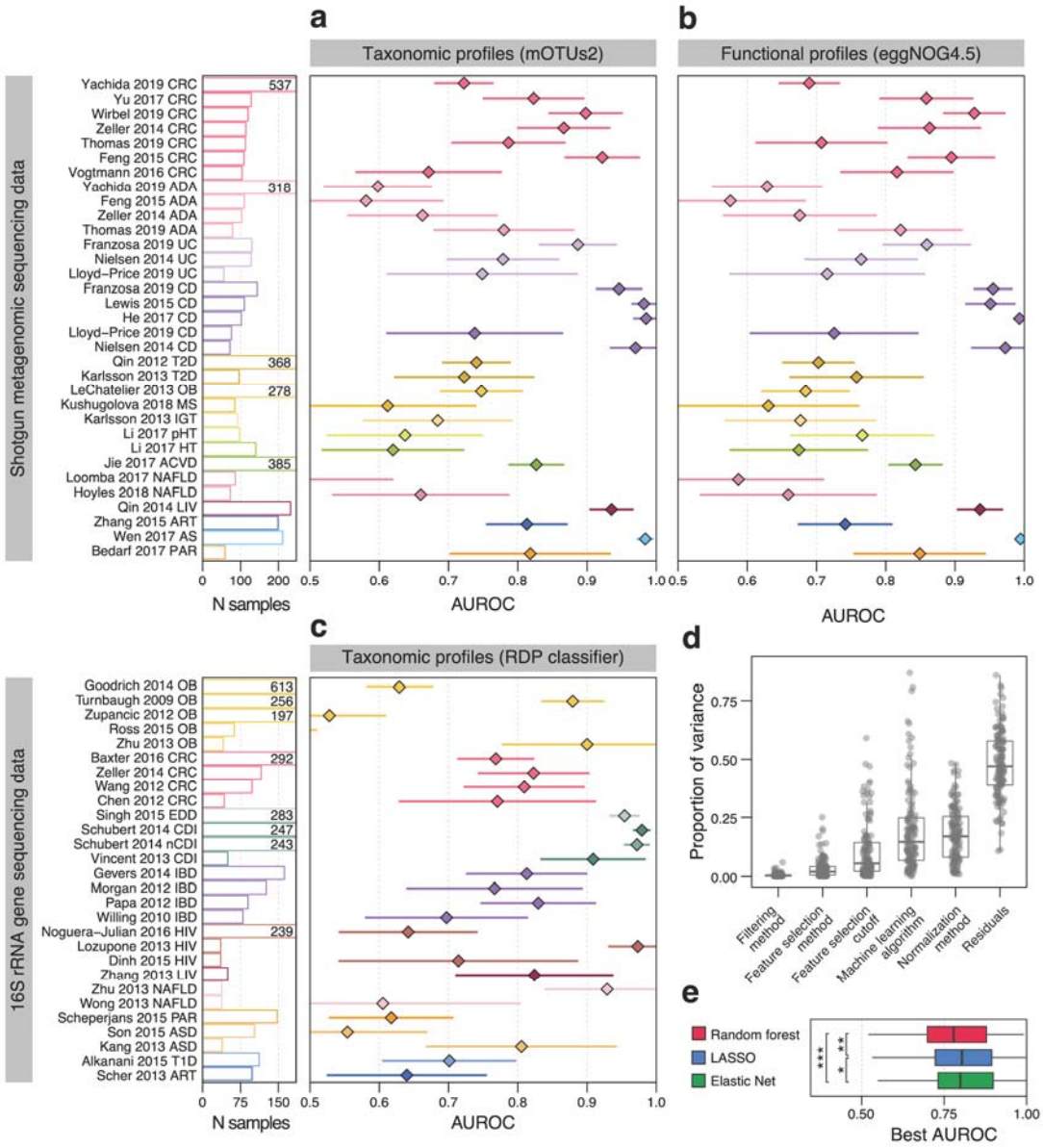


Figure 4. Large-scale application of the SIAMCAT machine learning workflow to human gut metagenomic disease association studies. (a) Application of SIAMCAT machine learning workflows to taxonomic profiles generated from fecal shotgun metagenomes using the mOTUs2 profiler. Cross-validation performance for discriminating between diseased patients and controls quantified by the area under the ROC curve (AUROC) is indicated by diamonds (95% confidence intervals denoted by horizontal lines) with sample size per dataset given as additional panel (cut at N = 250 and given by numbers instead). See **Table 1** and **Supplementary Table 1** for information about the included datasets and key for disease abbreviations. **(b)** Application of SIAMCAT machine learning workflows to functional profiles generated with eggNOG 4.5 for the same datasets as in (a) (see **Supplementary Figures 4, 7** for additional types of, and comparison between taxonomic and functional input data). **(c)** Cross-validation accuracy of SIAMCAT machine learning workflows as applied to 16S rRNA gene amplicon data for human gut microbiome case-control studies [20] (see (a) for definitions). **(d)** Influence of different parameter choices on the resulting classification accuracy. After training a linear model to predict the AUROC values for each classification task, the variance explained by each parameter was assessed using an ANOVA (see Methods). See Figure 1 for definition of boxplots. **(e)** Performance comparison of machine learning algorithms on gut microbial disease association studies. For each machine learning algorithm, the best AUROC values for each task are shown as boxplots (defined as in (D)). Generally, the choice of algorithm only has a small effect on classification accuracy, but both the Elastic Net and LASSO performance gains are statistically

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

significant (paired Wilcoxon test: LASSO vs Elastic Net, *P* = 0.001; LASSO vs Random Forest, *P* = 1e-08; Elastic Net vs Random Forest, *P* = 4e-14).

Cross-study evaluation of microbiome signatures is crucial to establish their validity across patient populations. However, such comparisons are potentially hindered by inter-study differences in sample handling and data generation, with technical variation observed to often dominate over biological factors of interest [88–90]. Additionally, biological and clinical factors can contribute to inter-study differences. These not only include influences of geography, ethnicity, demographics, and life-style, but also how clinical phenotypes are defined and controls selected for clinical studies [91].

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

When evaluating cross-study application of ML models, we specifically focused on two questions (see **Supplementary Figure 9** and Methods). First, we asked to which extent the separation between cases and controls (in terms of prediction scores) would be maintained when control samples of a different study are used. This can be measured via the AUROC, which we rescaled to range between 0 (indicating complete loss of discriminatory power on external data) and 1 (meaning that the ML model could be transferred to another dataset without loss of discrimination accuracy). Conveniently, this measure of cross-study portability can be assessed even on datasets for different diseases, since it is only based on external controls, not cases. Second, we asked how specific an ML model would be to the disease it was trained to recognize, or whether its false-positive rate would be elevated when presented with cases from a distinct condition. This is of interest in the context of an ongoing debate on whether there is a general gut microbial dysbiosis or distinct compositional changes associated with each disease [19,20,92]. Disease-specific classifiers would also be of clinical relevance when applied to a general population: due to large differences in disease prevalence, a model for CRC (a condition with low prevalence) misclassifying many type 2 diabetes (T2D) patients (high prevalence) would in the general population detect many more (false) T2D cases than CRC cases, and thus have very low precision for CRC. To quantify disease specificity of an ML model, we assessed how many samples from a distinct disease would be mispredicted as positive for the disease the ML model was trained on (at a cutoff adjusted to maintain a false-positive rate of 10% on the cross-validation set).

Our evaluations showed low cross-study portability on the majority of external data sets (apparent also from a more than twofold increase in false positives on average) for most models, which suggests that ML model transfer is substantially impacted -- if not rendered impossible -- by biological and technical study heterogeneity (**Supplementary Figure 10**). Accordingly, (false-positive) predictions for other diseases were elevated for most models (by a factor of 2.8 on average), with the extreme case of the ankylosing spondylitis (AS) model predicting more than 90% of cases from other diseases to be AS positive (median across studies, **Supplementary Figure 11**). These evaluations indicate dramatic problems with specificity and cross-disease predictions when naively transferring microbiome ML models.

In order to improve the cross-study portability of ML models, we devised a strategy we call control augmentation, in which randomly selected control samples from independent

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

microbiome population cohort studies [93–95] are added to the training set during model fitting (**Fig 5a**, see Methods). This was motivated by the hypothesis that additional variability from a greater control pool comprising heterogeneous samples from multiple studies would enable classifiers to more specifically recognize disease signals while at the same time minimizing overfitting on peculiarities of a single dataset. We indeed found that control augmentation greatly enhanced cross-study portability uniformly across all ML models (**Fig 5bc**, **Supplementary Figure 10**). At the same time, cross-disease predictions decreased (**Fig 5cd**, **Supplementary Figure 11**) implying that control augmentation is an effective strategy to increase disease specificity of ML models. The control augmentation strategy did not strongly depend on the set of controls used. We found large (>250 samples) cohort studies to work well as a pool for control-augmentation (allowing us to add five times the amount of control samples to each dataset). However, augmentation with fewer controls or with other datasets improved cross-study portability and disease specificity to almost the same effect (**Supplementary Figure 12**).

With cross-study portability greatly improved, we expect the remaining cross-disease predictions to be largely due to biological similarities between diseases rather than due to technical influences. In support of this, we show that CRC signatures have a tendency to cross-predict samples from patients with intestinal adenomas (ADA) or inflammatory bowel disease (CD), both of which are risk factors for CRC development [96]. Similarly, UC models cross-predict CD cases and vice versa, reflecting more general gut microbial changes, i.e. loss of beneficial commensal bacteria, that are shared across both types of inflammatory bowel disease [97]. In summary, we demonstrate that control augmentation is an effective strategy to broadly enable the validation of microbiome disease signatures across different studies, since it can overcome study-specific biases, which preclude the naive transfer of ML models.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

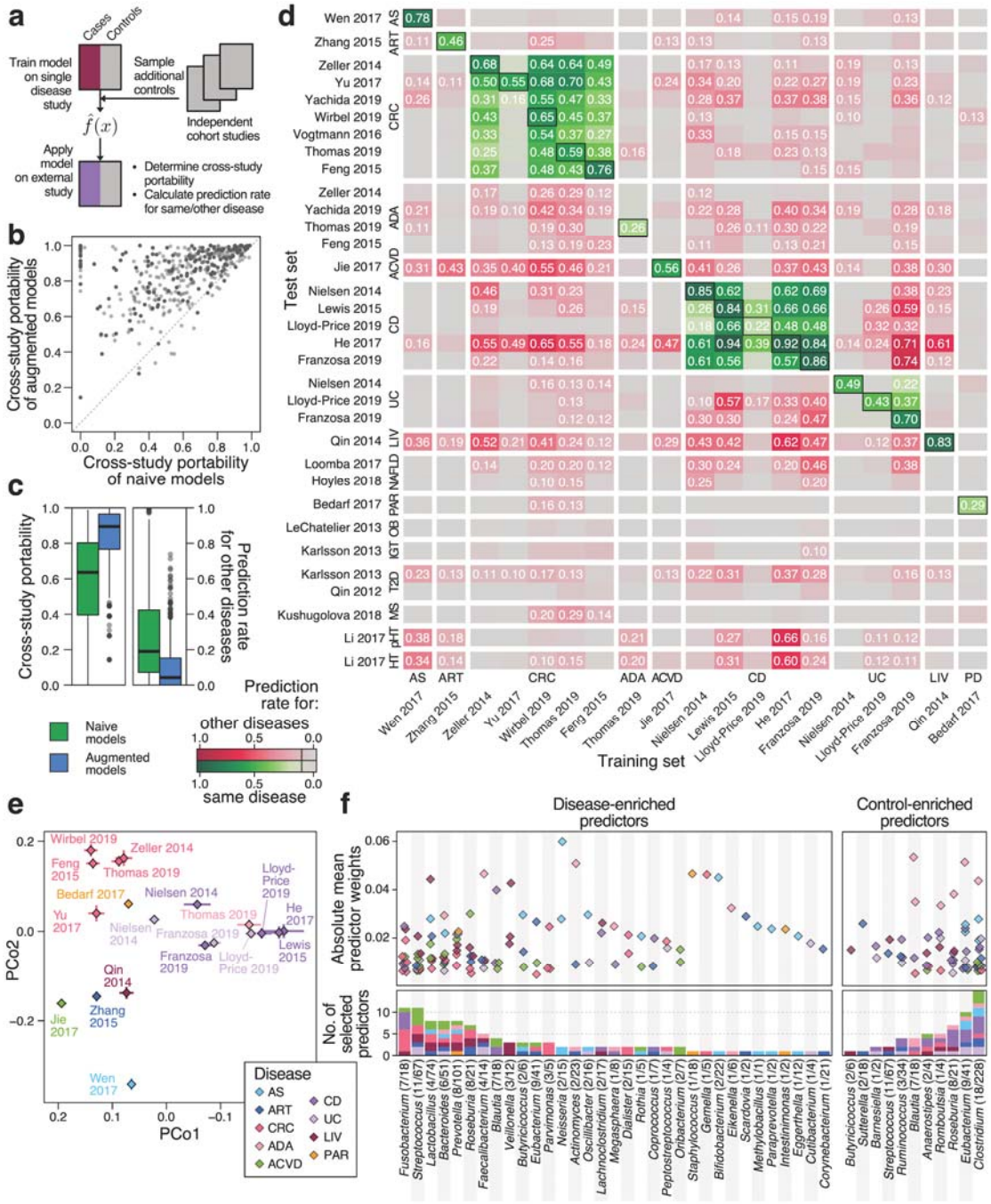


Figure 5. Control augmentation improves ML model disease specificity and reveals shared and distinct predictors. (a) Schematic of the control augmentation procedure: Control samples from external cohort studies are added to the individual cross validation folds during model training. Trained models are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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

When comparing microbiome signatures across diseases in more detail, we also revisited the question of whether microbiome alterations are specific to a disease, or signs of a general dysbiotic state [20]. As many ML algorithms, in particular (generalised) linear models, such as LASSO or Elastic Net logistic regression models, allow for model introspection, microbiome biomarkers can easily be extracted and their weight in the model directly quantified by (normalized) coefficient values. The model weights of the control-augmented models showed a clear clustering by disease in principal coordinate space revealing broad disease-similarity patterns in terms of microbiome predictors that may reflect etiological similarities (**Fig 5e**, not apparent from naively transferred ML models, **Supplementary Figure 13**). To obtain a more nuanced view of gut bacterial taxa underlying these disease similarities, we analysed individual mOTUs (grouped be genus membership) that were selected as predictors in disease models (**Fig 5f**, to minimize bias from multiple studies of the same disease, we used the mean model for each disease and extracted those features whose weights accounted for more than 50% of the model, see Methods for details). We found some disease-enriched predictors to be very specific for a single disease, such as *Veillonella* spp. for LIV, *Bifidobacteria* and *Neisseria* mOTUs for AS, or *Gemella* and *Parvimonas* mOTUs for CRC. In contrast, species from other genera, for example *Lactobacillus*, *Bacteroides* or *Fusobacteria*, appear predictive of several diseases, although species and subspecies belonging to these vary in terms of their disease specificity (**Supplementary Figure 14**). Regarding control-enriched predictors, species from some genera are frequently depleted across multiple diseases (*Anaerostipes* and *Romboutisa*) while some diseases are marked by broad depletion of beneficial microbes, e.g. CD (consistent with [97]).

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

Meta-analysis of Crohn's disease gut microbiome studies

Microbiome disease associations being reported at an ever-increasing pace have also provided opportunities for comparisons across multiple studies of the same disease to assess the robustness of associations and the generalisability of ML models [19,20,37,38]. To demonstrate SIAMCAT's utility in single-disease meta-analyses, we analyzed five metagenomic datasets [5,27–30], all of which included samples from patients with Crohn's disease (CD) as well as controls not suffering from inflammatory bowel diseases (IBD). Raw sequencing data consistently processed to obtain genus abundance profiles with mOTUs2 [85].

Based on SIAMCAT's *check.associations* function, we identified microbial genera that are significantly associated with CD in each study and visualized their agreement across studies (**Figure 6a,** left panel). In line with previous findings [4], the gut microbiome of CD patients is characterised by a loss of diversity and many beneficial taxa. Though our re-analysis of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

data from [30] could not identify any statistically significant genus-level associations, possibly due to the relatively small number of individuals or the choice of control samples obtained from patients with non-IBD gastrointestinal symptoms, the other four studies showed remarkable consistency among the taxa lost in CD patients, in particular for members of the Clostridiales order.

We further investigated variation due to technical and biological differences between studies as a potential confounder using SIAMCAT's *check.confounder* function following a previously validated approach [37]. For many genera, variation can largely be attributed to heterogeneity among studies; the top five associated genera (cf. **Figure 6a**), however, vary much more with disease status, suggesting that their association with CD is only minimally confounded by differences between studies (**Figure 6b**).

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

We lastly assessed the importance of individual microbial predictors in the CD models. The LASSO, and to some extent also the Elastic Net, are sparse models, in which the number of influential predictors (with nonzero coefficients) is kept small. As a consequence, these ML methods tend to omit statistically significant features when they are correlated to each other in favor of a smaller subset of features with optimal predictive power. Nonetheless, in our meta-analysis of CD, the feature weights derived from multivariable modeling corresponded well to the univariate associations, and also showed some consistency across the four studies in which clear CD associations could be detected and an accurate ML model trained (**Figure 6a**, right panel). Taken together, these results demonstrate that SIAMCAT could be a tool of broad utility for consolidating microbiome-disease associations and biomarker discovery by leveraging the large amount of metagenomic data becoming available for ML-based analyses.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

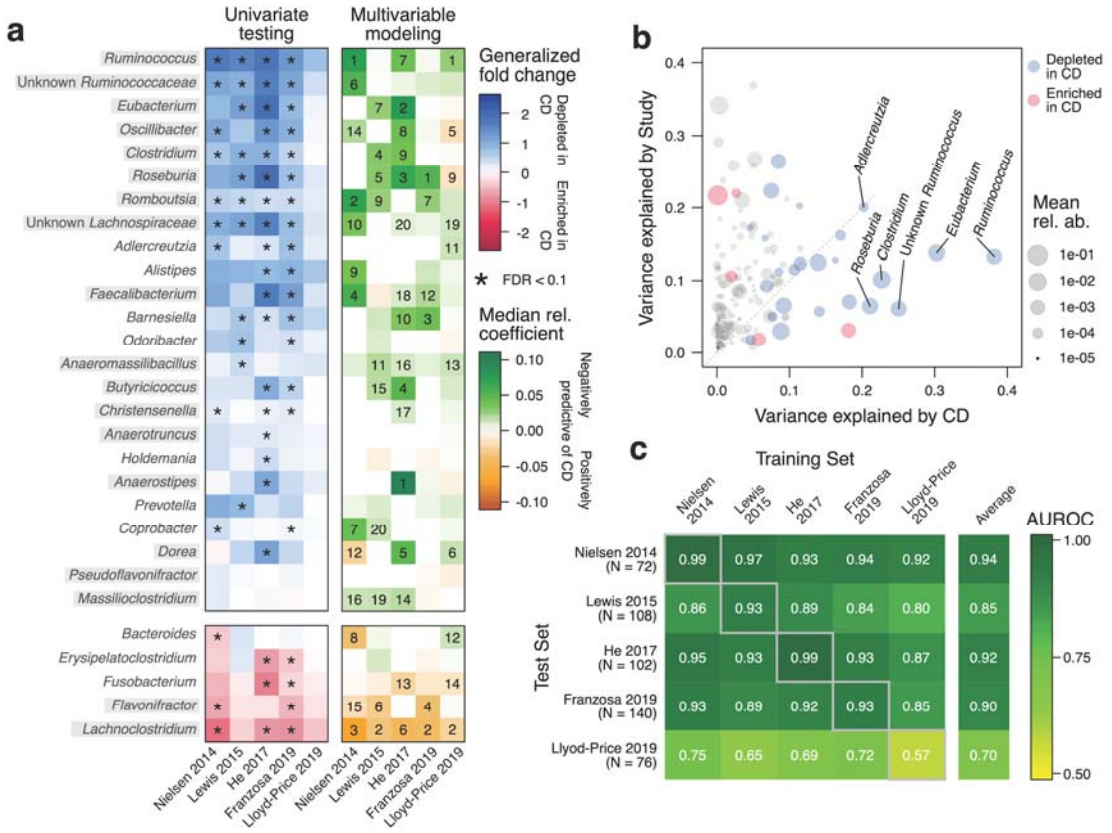


Figure 6. Meta-analysis of CD studies based on fecal shotgun metagenomic data.
(a) Genus-level univariate and multivariable associations with CD across the five included metagenomic studies. The heatmap on the left side shows the generalized fold change for genera with a single-feature AUROC higher than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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Discussion

The rising interest in clinical microbiome studies and microbiome-derived diagnostic, prognostic, and therapeutic biomarkers also calls for more robust analysis procedures. An important step in that direction is the development of freely available, comprehensive, and extensively validated analysis workflows that make complex ML procedures available to non-experts, ideally while safeguarding against statistical analysis pitfalls. Designed with these objectives in mind, SIAMCAT provides a modular analysis framework that builds on the existing R-based microbiome analysis environment: data integration from DADA2 [98] or phyloseq [69] is straightforward since SIAMCAT internally uses the phyloseq object. ML algorithms and procedures in SIAMCAT interface to the mlr package [70], a general-purpose ML library. Since the multitude of ML algorithms, workflow options, and design choices within such a general package can make setting up ML workflows for microbiome research challenging for non-experts, SIAMCAT mainly aims to enable novice ML users to apply robust and validated ML workflows to their data with preprocessing and normalization options tailored to the characteristics of microbiome data. At the same time, SIAMCAT allows advanced users to flexibly set up and customize more complex ML procedures, including non-standard cross-validation splits for dependent measurements and supervised feature selection methods that are properly nested into cross validation (**Figure 5**). Further developments of the package are planned to accommodate the rapidly changing needs of the microbiome research community and updates will be published in accordance with the established Bioconductor release schedule.

To showcase the power of ML workflows implemented in SIAMCAT, we performed a meta-analysis of human gut metagenomic studies at considerably larger scale than previous efforts [17–22] (see **Figure 4**). It importantly encompassed a large number of diseases as well as different taxonomic and functional profiles as input that were derived from different metagenomic sequencing techniques (16S rRNA gene and shotgun metagenomics sequencing) and profiling tools. Consequently, these benchmarks are expected to yield much more robust and general results than those from previous studies [17–22]. In our exploration of more than 7.000 different parameter combinations per classification task (see Methods), we found the Elastic Net logistic regression algorithm to yield highest cross-validation accuracies on average, albeit requiring the input data to be appropriately normalized (see **Figure 4** and **Supplementary Figure 8**). Compared with the choice of normalisation method and classification algorithm, other parameters had a considerably lower influence on the resulting classification accuracy. SIAMCAT's functionality to robustly fit statistical microbiome models and evaluate their performance will enable comparison to established diagnostic biomarkers [99] as an important prerequisite for further translation of microbiome research into the clinic.

To help resolve the debate about spurious associations and reproducibility issues in microbiome research [100], meta-analyses are crucial for the validation of microbiome biomarkers [37,38]. However, we found that ML models have substantial problems with type-I-error control (>2-fold increase in false-positive rate) and disease specificity (>2.5-fold elevated false-positive rate) when naively transferred across studies. We propose measures to detect these issues, which, if more widely adopted, could help to more precisely characterize them and their underlying causes. To address them, we introduce the control augmentation strategy, which greatly improved cross-study portability of ML models. Being a

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

first attempt to overcome study heterogeneity for improved cross-study model application, our work will hopefully stimulate further developments, which could easily be evaluated on the provided datasets. However, all such ML meta-analyses are limited by biological and clinical differences between studies [91], which can only be addressed by better reporting standards [101]. Within these limitations, our ML meta-analysis data sets could become a valuable community resource for method development, systematic assessment of disease similarities, and further exploration of globally applicable microbiome biomarkers including validation of their disease-specificity.

Using model introspection after control-augmentation, we could revisit the question if microbiome alterations are specific to a given disease or more general hallmarks of a dysbiotic state [20]. In general, we found depletion of beneficial bacteria to be more often shared across several diseases (e.g. *Anaerostipes* or *Romboutisa*), in particular in (the subtypes of) IBD. Conversely, disease-enriched bacteria were more often specific for a given disease. This could mean that some disease-specific microbiome alterations may reflect pathogens or pathobionts acting either as etiological agents or exploiting specific disease-related changes in the intestinal milieu. For example, *Parvimonas* spp. are predictive for colorectal cancer, which is consistent with mechanistic work in this area [102]. Similarly, a putative link between oral *Veillonella* spp. and liver cirrhosis severity has been reported in the context of proton-pump inhibitor therapy [103], potentially enabled by increased transmission from the oral to the gut microbiome [78]. Other taxa show a broader disease spectrum, such as *Fusobacterium* spp., which have been studied both in the context of CRC [104] as well as in IBD [105]. However, firmly establishing disease-specificity or disease spectra for microbial biomarkers will require large cohort studies and additional mechanistic work to elucidate the gut microbiome's contribution to disease development at the molecular level. Nonetheless, our analyses suggest both shared and disease-specific gut microbial biomarkers to guide further investigations of specific hypotheses on their ecological roles.

Although the analyses presented here are focused on human gut metagenomic datasets with disease prediction tasks, SIAMCAT is not restricted to these. It can also be applied to other tasks of interest in microbiome research, e.g. for investigating the effects of medication (see **Supplementary Figure 2**). Metagenomic or metatranscriptomic data from environmental samples can also be analysed using SIAMCAT, e.g. to understand associations between community composition or transcriptional activity of the ocean microbiome with physicochemical environmental properties (see **Supplementary Figure 15** for an example [106]) highlighting that SIAMCAT could be of broad utility in microbiome research.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

Implementation

SIAMCAT is implemented as an R package with a modular architecture, allowing for flexible combination of different functions to build ML and statistical analysis workflows (see **Code Box**). The output of the functions (for example, the feature matrix after normalization) is stored in the SIAMCAT object, which is an extension of the *phyloseq* object that contains the raw feature abundances, meta-variables about the samples, and other optional information (for example, a taxonomy table or a phylogenetic tree) [69]. The label defining the sample groups for comparison is then derived from a user-specified meta-variable or an additional vector. ML models are trained using the *mlr* infrastructure as an interface to the implementations of different ML algorithms in other R packages [70]**.** SIAMCAT is available under the GNU General Public License, Version 3.

Code Box

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

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

Included datasets and microbiome profiling

In this study, we analyzed taxonomic and functional profiles derived with different profiling tools from several metagenomic datasets (see **Supplementary Table 1**). Taxonomic profiles generated using the RDP classifier [83] on the basis of 16S rRNA gene sequencing data were downloaded from a recent meta-analysis by Duvallet et al. [20] and summarized at the genus level. MetaPhlAn2 [84] and HUMAnN2 [86] taxonomic and functional profiles were obtained from the *curatedMetagenomicsData* R package [72] for all human gut datasets within the package that contained at least 20 cases and 20 controls. MetaPhlAn2 profiles were filtered to contain only species-level microbial taxa.

Additional datasets were profiled in-house with the following pipeline implemented in *NGless* [107]: after preprocessing with MOCAT2 [108] and filtering for human reads, taxonomic

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

profiles were generated using the mOTUsv2 profiler [85] and functional profiles were calculated by first mapping reads against the integrated gene catalogue [109] and then aggregating the results by eggNOG orthologous groups [87].
Additionally, genus-level taxonomic profiles from the TARA Oceans microbiome project [106] were used for two different classification tasks: to classify samples from polar and non-polar ocean regions and to classify samples based on their iron concentration at a depth of 5 meters (high versus low iron content).

Primary package outputs and confounder analysis
To illustrate the main outputs of SIAMCAT, we analyzed the taxonomic profiles from a metagenomic study of IBD [27] included in the *curatedMetagenomicsData* R package [72]. For the analyses presented in **Figure 1**, we restricted the dataset to control samples from Spain and cases with UC, since the two IBD subtypes included in the dataset (ulcerative colitis and Crohn's disease) are very different from one another in terms of the associated changes in gut microbiome composition. See the **SIAMCAT vignettes** for more information or the **Code Box** for an outline of the basic SIAMCAT workflow.
To demonstrate how SIAMCAT can aid in confounder detection, we used the same dataset but this time included the Danish control samples in order to explore potential confounding by differences between samples collected and processed in these two countries. The analyses presented in **Figure 2** have all been conducted with the respective functions of SIAMCAT (see **SIAMCAT vignettes**).

Machine learning hyperparameter exploration
To explore suitable hyperparameter combinations for ML workflows, we trained an ML model for each classification task and each hyperparameter combination. By hyperparameter we mean configuration parameters of the workflow, such as normalisation parameters, tuning parameters controlling regularisation strength or properties of external feature selection procedure in contrast to model parameters fitted during the actual training of the ML algorithms. Specifically, we varied the filtering method (no data filtering, prevalence filtering with 1%, 5%, 10% cutoffs, abundance filtering with 0.001, 0.0001, 0.0001 as cutoffs, and a combination of abundance and prevalence filtering), the normalization method (no normalization beyond the total sum scaling, log-transformation with standardization, rank-transformation with standardization, and centered log ratio transformation), the ML algorithm (LASSO, Elastic Net, and Random Forest classifiers), and feature selection regimes (no feature selection, feature selection based on generalized fold change or based on single-feature AUROC; cutoffs were 25, 50, 100, 200, and 400 features for taxonomic profiles and 100, 500, 1000, and 2000 features for functional profiles). To cover the full hyperparameter space, we therefore trained 7.488 models for taxonomic and 3.168 models for functional datasets for each classification task.
To determine the optimal AUROC across input types (shown in **Figure 4**), we calculated for each individual parameter combination the mean AUROC across all classification tasks with a specific type of input. Different feature filtering procedures could lead to cases in which the feature selection cutoffs were larger than the number of available features after filtering, therefore terminating the ML procedure. For that reason, we only considered those parameter combinations that did produce a result for all classification tasks with the specific type of input data.
To compare the importance of feature filtering, feature selection, normalization method and ML algorithm on classification accuracy, we trained one linear model per classification task

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

predicting the AUROC values from those variables. We then partitioned the variance attributable to each of these variables by calculating type III sums of squares using the Anova function from the car package in R [110].

In order to contrast class separation of samples in distance space with the classification performance achieved by ML algorithms (see **Supplementary Figure 6**), we designed a distance-based measure of separation. For each dataset, we determined the distances between all pairs of samples within a class as well as all pairs of samples between classes and then calculated an AUROC value based on these two distributions. This distance-based measure effectively quantifies to what extent samples are closest to other samples from the same class (i.e. cluster together) and hence corresponds well to the visual separation of classes in ordination space (see **Supplementary Figure 6**).

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

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

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

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

Training Elastic Net models with control augmentation

To train models with the control augmentation strategy, we used the data from the following cohort microbiome studies as additional control samples: [93–95]. Repeated measurements for the same individual were removed in the case of Zeevi et al. For each training set in the repeated cross-validation, we increased the number of control samples 5-fold by randomly sampling the appropriate number of controls (in a balanced manner between datasets to avoid overrepresentation of the larger external cohorts). Before addition, the additional control samples were normalized using the normalization parameters of the discovery set. Due to the introduction of additional variability, the control-augmented Elastic Net models were trained with a pre-set alpha value of 0.5 to ensure stability of model size.

To compare the predictors across different diseases, model weights of the control-augmented models were transformed into relative weights by dividing by the sum of absolute coefficient values. Then, models from the same disease were averaged. Predictors (that is, mOTUs) were selected for display in **Figure 5f**, if they (i) cumulatively contributed more than 50% of the mean relative disease model, (ii) their individual weights were bigger than 1%, and (iii) the genus annotation had an unambiguous NCBI taxonomy.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Illustration of common pitfalls in machine learning procedures

To demonstrate how naive sequential application of supervised feature selection and cross validation might bias performance estimations, we trained LASSO ML models to distinguish colorectal cancer cases from controls based on MetaPhlAn2-derived species abundance profiles using the dataset with the handle *ThomasAM_2018a* [38] obtained through the *curatedMetagenomicsData* R package [72]. For the incorrect procedure of feature selection, single-feature AUROC values were calculated using the complete dataset (inverted for negatively associated features). Then, the features with the highest AUROC values were selected for model training (number depending on the cutoff). In contrast, the correct procedure implemented in SIAMCAT excludes the data in the test fold when calculating single-feature AUROC values; instead, AUROC values are calculated on the training fold only. To test generalization to external data, the models were then applied to another colorectal cancer metagenomic study [8] available through the *curatedMetagenomicsData* R package (also see the SIAMCAT vignette: Holdout testing).

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

Meta-analysis of Crohn's disease metagenomic studies

For the meta-analysis of Crohn's disease gut microbiome studies, we included five metagenomic datasets [5,27–30] that had been profiled with the mOTUs2 profiler [85] on genus level. While some datasets contained both UC and CD patients [5,27,30], other datasets contained only CD cases [28,29]. Therefore, we restricted all datasets to a comparison between only CD cases and control samples, since the two subtypes of IBD are very different from each other.

For training of ML models, we blocked repeated measurements for the same individual when applicable [27,28,30]; specifically for the iHMP dataset [30], we also subsampled the dataset to five repeated measurements per individual to avoid biases associated with differences in the number of samples per individual. For external validation testing, we completely removed repeated measurements in order not to bias the estimation of classification accuracy.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Declarations

Availability of Data and Materials

Raw metagenomics data are available from ENA (see **Supplemental Table 1** for the identifiers of included datasets). All taxonomic and functional profiles used as input for the presented analyses are available from Zenodo (https://doi.org/doi:10.5281/zenodo.3613794) and the code to reproduce the analysis can be found in the dedicated Github repository (https://github.com/zellerlab/siamcat_paper).

The code for SIAMCAT can be found under https://github.com/zellerlab/siamcat and on Bioconductor under https://doi.org/doi:10.18129/B9.bioc.SIAMCAT.

Competing Interests

No competing interests declared.

Funding

We acknowledge funding from EMBL, ETH (PHRT no. 521 to S.S.), the Federal Ministry of Education and Research (BMBF; the de.NBI network no. 031A537B to P.B. and grant no. 031L0181A to G.Z. and P.B.), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation no. 395357507 – SFB 1371 to G.Z.) and the Helmut Horten Foundation (to S.S.).

Author contributions

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

Acknowledgments

We are grateful to Mike Smith, Paul I. Costea, and Kersten Breuer for helpful discussions and advice on the implementation of SIAMCAT. We thank members of the Zeller, Sunagawa and Bork group for fruitful discussions and the EMBL Information Technology Core Facility for support with high-performance computing.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Schmidt TSB, Raes J, Bork P. The Human Gut Microbiome: From Association to Modulation. Cell. Elsevier; 2018;172:1198–215.

2. Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. N Engl J Med. Mass Medical Soc; 2016;375:2369–79.

3. Garrett WS. The gut microbiota and colon cancer. Science. science.sciencemag.org; 2019;364:1133–5.

4. Gevers D, Kugathasan S, Denson LA. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. Elsevier; 2014;

5. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol. nature.com; 2019;4:293–305.

6. Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med. nature.com; 2015;21:895–905.

7. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, et al. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. Genome Biol. genomebiology.biomedcentral.com; 2017;18:142.

8. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. John Wiley & Sons, Ltd; 2014;10.

9. Tilg H, Cani PD, Mayer EA. Gut microbiome and liver diseases. Gut. gut.bmj.com; 2016;65:2035–44.

10. Zitvogel L, Ma Y, Raoult D, Kroemer G, Gajewski TF. The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. Science. science.sciencemag.org; 2018;359:1366–70.

11. Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature. 2015;528:262–6.

12. Voigt AY, Costea PI, Kultima JR, Li SS, Zeller G, Sunagawa S, et al. Temporal and technical variability of human gut metagenomes. Genome Biol. 2015;16:73.

13. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Elife [Internet]. 2019;8. Available from: http://dx.doi.org/10.7554/eLife.46923

14. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. genomebiology.biomedcentral.com; 2011;12:R60.

15. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. nature.com; 2013;10:1200–2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

16. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. Taylor & Francis; 2015;26:27663.

17. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. Human-associated microbial signatures: examining their predictive value. Cell Host Microbe. Elsevier; 2011;10:292–6.

18. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. academic.oup.com; 2011;35:343–59.

19. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. 2016;12:e1004977.

20. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8:1784.

21. Wang J, Kurilshikov A, Radjabzadeh D, Turpin W, Croitoru K, Bonder MJ, et al. Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. Microbiome. microbiomejournal.biomedcentral …; 2018;6:101.

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

23. Zhou Y-H, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front Genet. frontiersin.org; 2019;10:579.

24. Le Goallec A, Tierney BT, Luber JM, Cofer EM, Kostic AD, Patel CJ. A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. PLoS Comput Biol. Public Library of Science San Francisco, CA USA; 2020;16:e1007895.

25. Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. Elife. 2013;2:e01202.

26. Alkanani AK, Hara N, Gottlieb PA, Ir D, Robertson CE, Wagner BD, et al. Alterations in Intestinal Microbiota Correlate With Susceptibility to Type 1 Diabetes. Diabetes. Am Diabetes Assoc; 2015;64:3510–20.

27. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32:822–8.

28. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host Microbe. Elsevier; 2015;18:489–500.

29. He Q, Gao Y, Jie Z, Yu X, Laursen JM, Xiao L, et al. Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. Gigascience. academic.oup.com; 2017;6:1–11.

30. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569:655–62.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

31. Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z, et al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology. 2010;139:1844–54.e1.

32. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. genomebiology.biomedcentral.com; 2012;13:R79.

33. Papa E, Docktor M, Smillie C, Weber S, Preheim SP, Gevers D, et al. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. PLoS One. 2012;7:e39242.

34. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat Commun. nature.com; 2015;6:6528.

35. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. gut.bmj.com; 2017;66:70–8.

36. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, et al. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PLoS One. journals.plos.org; 2016;11:e0155362.

37. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25:679–89.

38. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019;25:667–78.

39. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25:968–76.

40. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J. 2012;6:320–9.

41. Chen W, Liu F, Ling Z, Tong X, Xiang C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PLoS One. journals.plos.org; 2012;7:e39743.

42. Baxter NT, Ruffin MT 4th, Rogers MAM, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. genomemedicine.biomedcentral …; 2016;8:37.

43. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun. nature.com; 2017;8:845.

44. Li J, Zhao F, Wang Y, Chen J, Tao J, Tian G, et al. Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome. 2017;5:14.

45. Schubert AM, Rogers MAM, Ring C, Mogle J, Petrosino JP, Young VB, et al. Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

diarrhea from healthy controls. MBio. 2014;5:e01021–14.

46. Vincent C, Stephens DA, Loo VG, Edens TJ, Behr MA, Dewar K, et al. Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. Microbiome. 2013;1:18.

47. Singh P, Teal TK, Marsh TL, Tiedje JM, Mosci R, Jernigan K, et al. Intestinal microbial communities associated with acute enteric infections and disease recovery. Microbiome. 2015;3:45.

48. Lozupone CA, Li M, Campbell TB, Flores SC, Linderman D, Gebert MJ, et al. Alterations in the gut microbiota associated with HIV-1 infection. Cell Host Microbe. 2013;14:329–39.

49. Dinh DM, Volpe GE, Duffalo C, Bhalchandra S, Tai AK, Kane AV, et al. Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. J Infect Dis. academic.oup.com; 2015;211:19–27.

50. Noguera-Julian M, Rocafort M, Guillén Y, Rivera J, Casadellà M, Nowak P, et al. Gut Microbiota Linked to Sexual Preference and HIV Infection. EBioMedicine. 2016;5:135–46.

51. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513:59–64.

52. Zhang Z, Zhai H, Geng J, Yu R, Ren H, Fan H, et al. Large-scale survey of gut microbiota associated with MHE Via 16S rRNA-based pyrosequencing. Am J Gastroenterol. 2013;108:1601–11.

53. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, et al. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab. 2017;25:1054–62.e5.

54. Hoyles L, Fernández-Real J-M, Federici M, Serino M, Abbott J, Charpentier J, et al. Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. Nat Med. 2018;24:1070–80.

55. Wong VW-S, Tse C-H, Lam TT-Y, Wong GL-H, Chim AM-L, Chu WC-W, et al. Molecular characterization of the fecal microbiota in patients with nonalcoholic steatohepatitis--a longitudinal study. PLoS One. 2013;8:e62885.

56. Zhu L, Baker SS, Gill C, Liu W, Alkhouri R, Baker RD, et al. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH. Hepatology. 2013;57:601–9.

57. Bedarf JR, Hildebrand F, Coelho LP, Sunagawa S, Bahram M, Goeser F, et al. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. Genome Med. 2017;9:39.

58. Scheperjans F, Aho V, Pereira PAB, Koskinen K, Paulin L, Pekkonen E, et al. Gut microbiota are related to Parkinson's disease and clinical phenotype. Mov Disord. 2015;30:350–8.

59. Kang D-W, Park JG, Ilhan ZE, Wallstrom G, Labaer J, Adams JB, et al. Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. PLoS One. 2013;8:e68322.

60. Son JS, Zheng LJ, Rowehl LM, Tian X, Zhang Y, Zhu W, et al. Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Simons Simplex Collection. PLoS One. 2015;10:e0137725.

61. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. nature.com; 2013;500:541–6.

62. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell. Elsevier; 2014;159:789–99.

63. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457:480–4.

64. Zupancic ML, Cantarel BL, Liu Z, Drabek EF, Ryan KA, Cirimotich S, et al. Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. PLoS One. 2012;7:e43052.

65. Ross MC, Muzny DM, McCormick JB, Gibbs RA, Fisher-Hoch SP, Petrosino JF. 16S gut community of the Cameron County Hispanic Cohort. Microbiome. 2015;3:7.

66. Kushugulova A, Forslund SK, Costea PI, Kozhakhmetov S, Khassenbekova Z, Urazova M, et al. Metagenomic analysis of gut microbial communities from a Central Asian population. BMJ Open. 2018;8:e021682.

67. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. nature.com; 2013;498:99–103.

68. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. nature.com; 2012;490:55–60.

69. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One. Public Library of Science; 2013;8:e61217.

70. Bischl B, Lang M, Kotthoff L, Schiffner J. mlr: Machine Learning in R. The Journal of Machine [Internet]. jmlr.org; 2016; Available from: http://www.jmlr.org/papers/volume17/15-066/15-066.pdf

71. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A. National Acad Sciences; 2007;104:13780–5.

72. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017;14:1023–4.

73. Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 2019;20:210–21.

74. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Series B Stat Methodol. Wiley Online Library; 1996;58:267–88.

75. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. Wiley Online Library; 2005;

76. Tin Kam Ho. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. ieeexplore.ieee.org; 1995. p. 278–82 vol.1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

77. Deloris Alexander A, Orcutt RP, Henry JC, Baker J, Bissahoyo AC, Threadgill DW. Quantitative PCR assays for mouse enteric flora reveal strain-dependent differences in composition that are influenced by the microenvironment. Mamm Genome. Springer; 2006;17:1093–104.

78. Imhann F, Bonder MJ, Vich Vila A, Fu J, Mujagic Z, Vork L, et al. Proton pump inhibitors affect the gut microbiome. Gut. gut.bmj.com; 2016;65:740–8.

79. Jackson MA, Goodrich JK, Maxan M-E, Freedberg DE, Abrams JA, Poole AC, et al. Proton pump inhibitors alter the composition of the gut microbiota. Gut. gut.bmj.com; 2016;65:749–56.

80. Smialowski P, Frishman D, Kramer S. Pitfalls of supervised feature selection. Bioinformatics. 2010;26:440–3.

81. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography . Wiley Online Library; 2017;40:913–29.

82. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Math Intelligencer. Springer; 2005;

83. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. Am Soc Microbiol; 2007;73:5261–7.

84. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. nature.com; 2015;12:902–3.

85. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun. nature.com; 2019;10:1014.

86. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. nature.com; 2018;15:962–8.

87. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. academic.oup.com; 2016;44:D286–93.

88. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23:1704–14.

89. Sinha R, The Microbiome Quality Control Project Consortium, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium [Internet]. Nature Biotechnology. 2017. p. 1077–86. Available from: http://dx.doi.org/10.1038/nbt.3981

90. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35:1069–76.

91. Thompson SG. Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. BMJ [Internet]. bmj.com; 1994; Available from: https://www.bmj.com/content/309/6965/1351.full-text

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

92. Olesen SW, Alm EJ. Dysbiosis is not an answer. Nature Microbiology. nature.com; 2016;1:16228.

93. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163:1079–94.

94. Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, et al. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell. Elsevier; 2016;167:1897.

95. Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, et al. Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. Cell Syst. Elsevier; 2016;3:572–84.e3.

96. Bernstein CN, Blanchard JF, Kliewer E, Wajda A. Cancer risk in patients with inflammatory bowel disease: a population-based study. Cancer. Wiley Online Library; 2001;91:854–62.

97. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. gut.bmj.com; 2006;55:205–11.

98. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW. DADA2: high-resolution sample inference from Illumina amplicon data. Nature [Internet]. nature.com; 2016; Available from: https://www.nature.com/nmeth/journal/v13/n7/abs/nmeth.3869.html

99. Zeller G, Tap J, Voigt AY, Sunagawa S. Potential of fecal microbiota for early□stage detection of colorectal cancer. Mol Syst Biol [Internet]. embopress.org; 2014; Available from: https://www.embopress.org/doi/abs/10.15252/msb.20145645

100. Cani PD. Gut microbiota—at the intersection of everything? Nat Rev Gastroenterol Hepatol [Internet]. nature.com; 2017; Available from: https://www.nature.com/articles/nrgastro.2017.54.pdf?origin=ppub

101. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. nature.com; 2011;29:415–20.

102. Yu J, Zhao L, Zhao R, Long X, Coker OO, Sung JJY. The role of Parvimonas micra in intestinal tumorigenesis in germ-free and conventional APCmin/+ mice. J Clin Orthod. American Society of Clinical Oncology; 2019;37:531–531.

103. Horvath A, Rainer F, Bashir M, Leber B, Schmerboeck B, Klymiuk I, et al. Biomarkers for oralization during long-term proton pump inhibitor therapy predict survival in cirrhosis. Sci Rep. nature.com; 2019;9:12000.

104. 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. Elsevier; 2013;14:195–206.

105. Ohkusa T, Okayasu I, Ogihara T, Morita K, Ogawa M, Sato N. Induction of experimental ulcerative colitis by Fusobacterium varium isolated from colonic mucosa of patients with ulcerative colitis. Gut. gut.bmj.com; 2003;52:79–83.

106. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, et al. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.06.931808; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

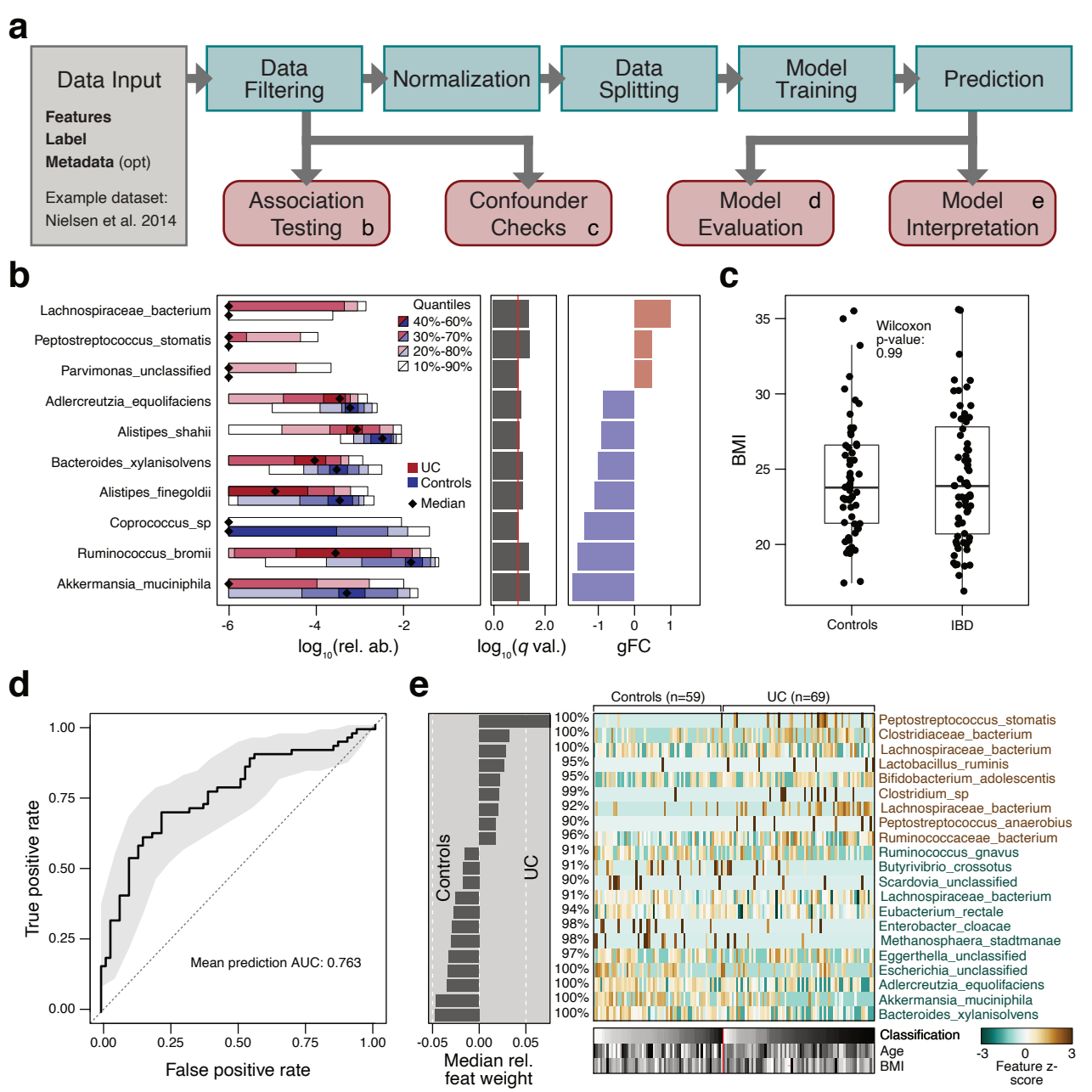
Metatranscriptome. Cell. Elsevier; 2019;179:1068–83.e21.

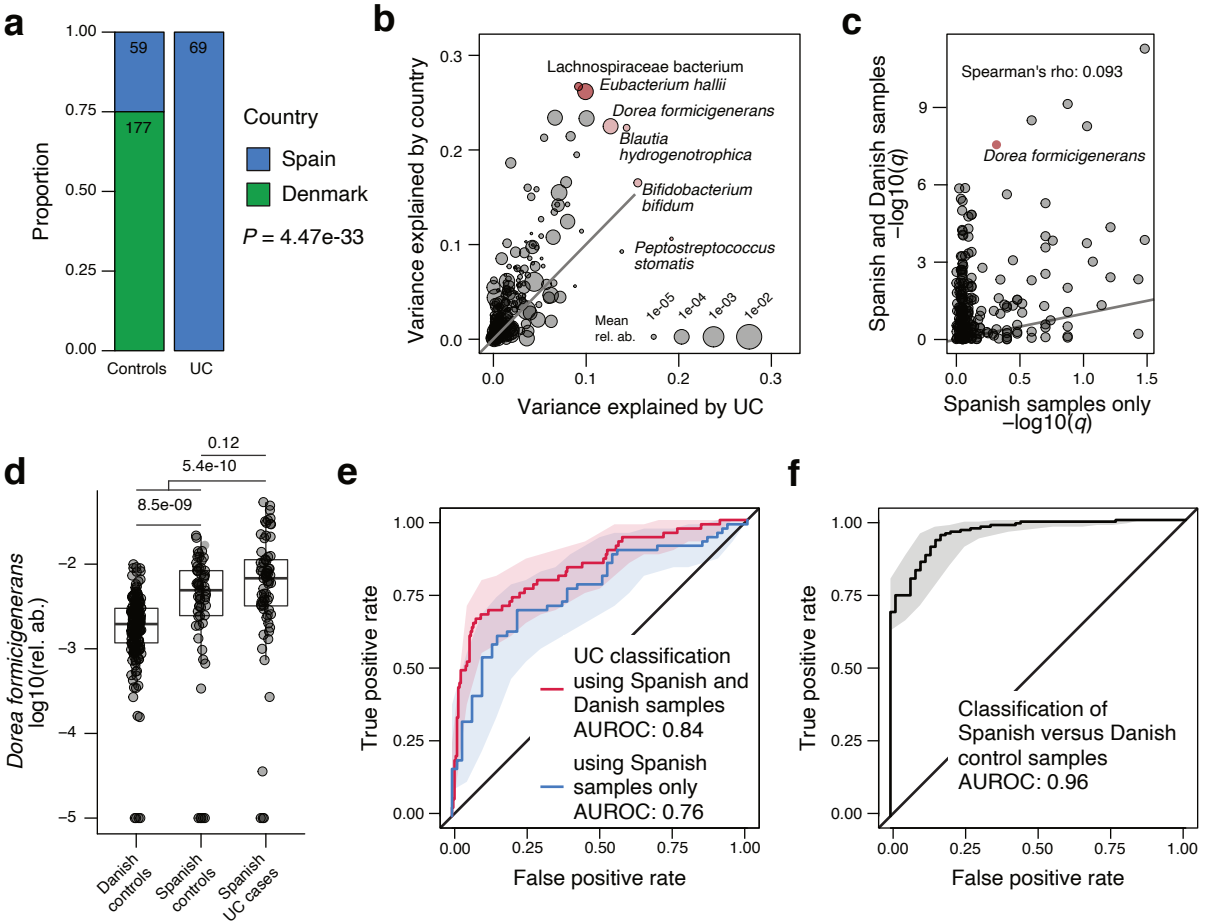
107. Coelho LP, Alves R, Monteiro P, Huerta-Cepas J, Freitas AT, Bork P. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. Microbiome. microbiomejournal.biomedcentral …; 2019;7:84.

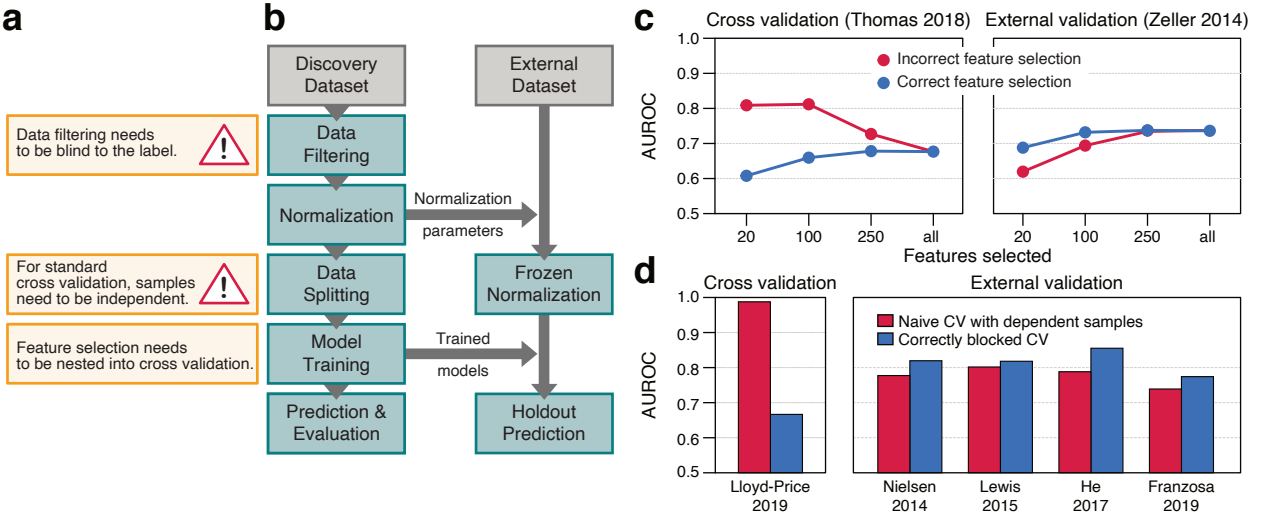
108. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics. academic.oup.com; 2016;32:2520–3.

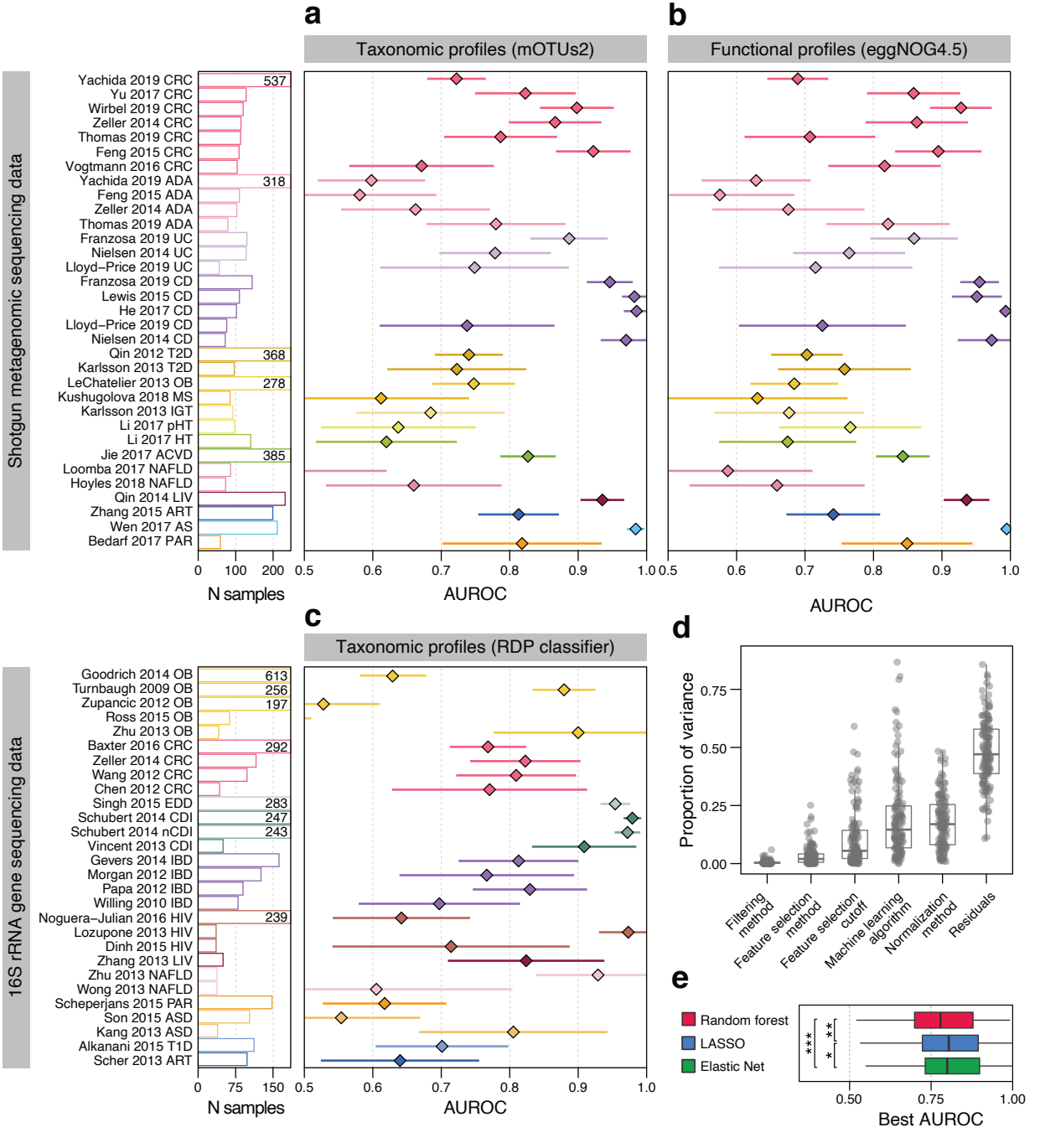
109. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. nature.com; 2014;32:834–41.

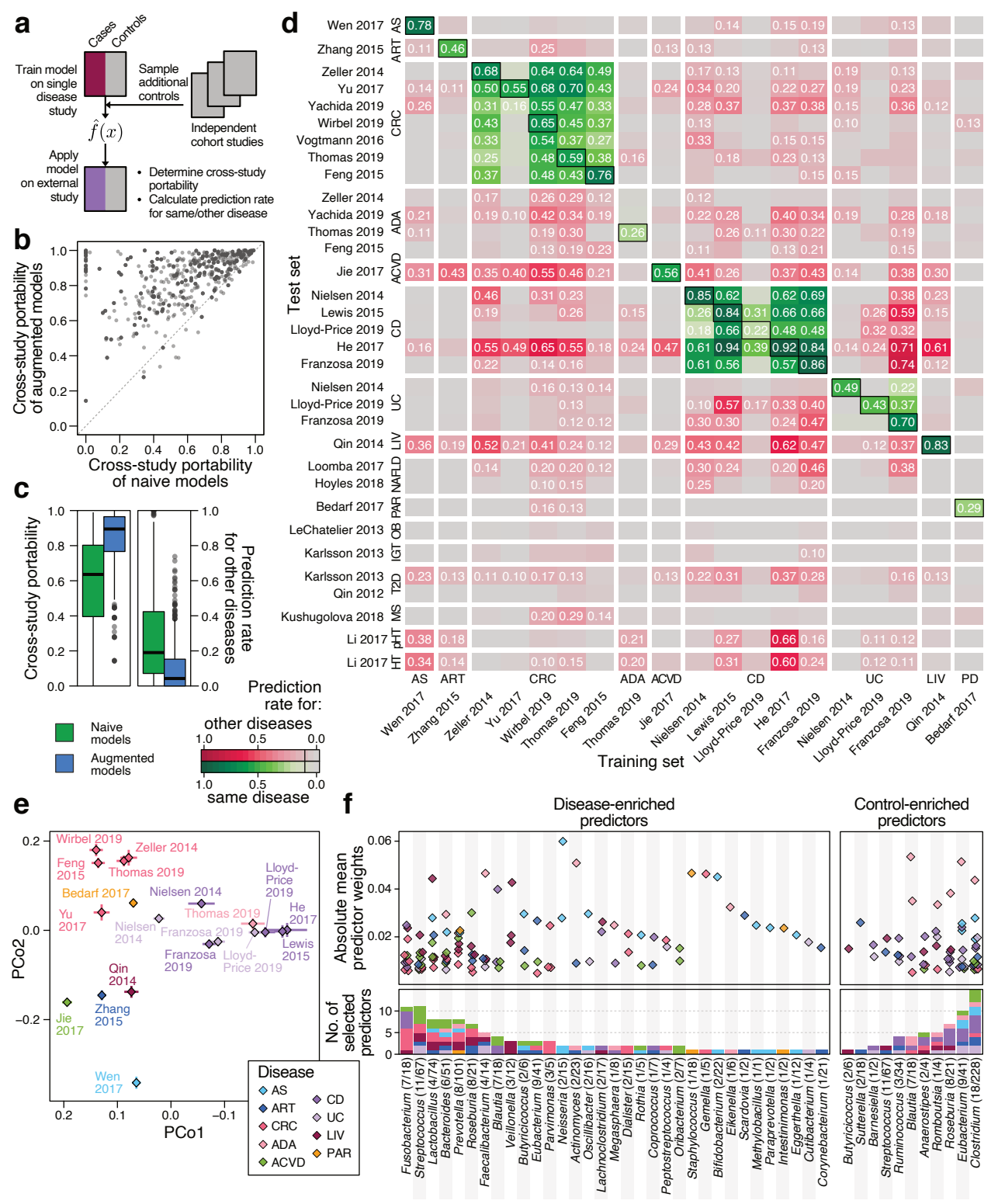
110. Fox J, Weisberg S. An R Companion to Applied Regression. SAGE Publications; 2018.

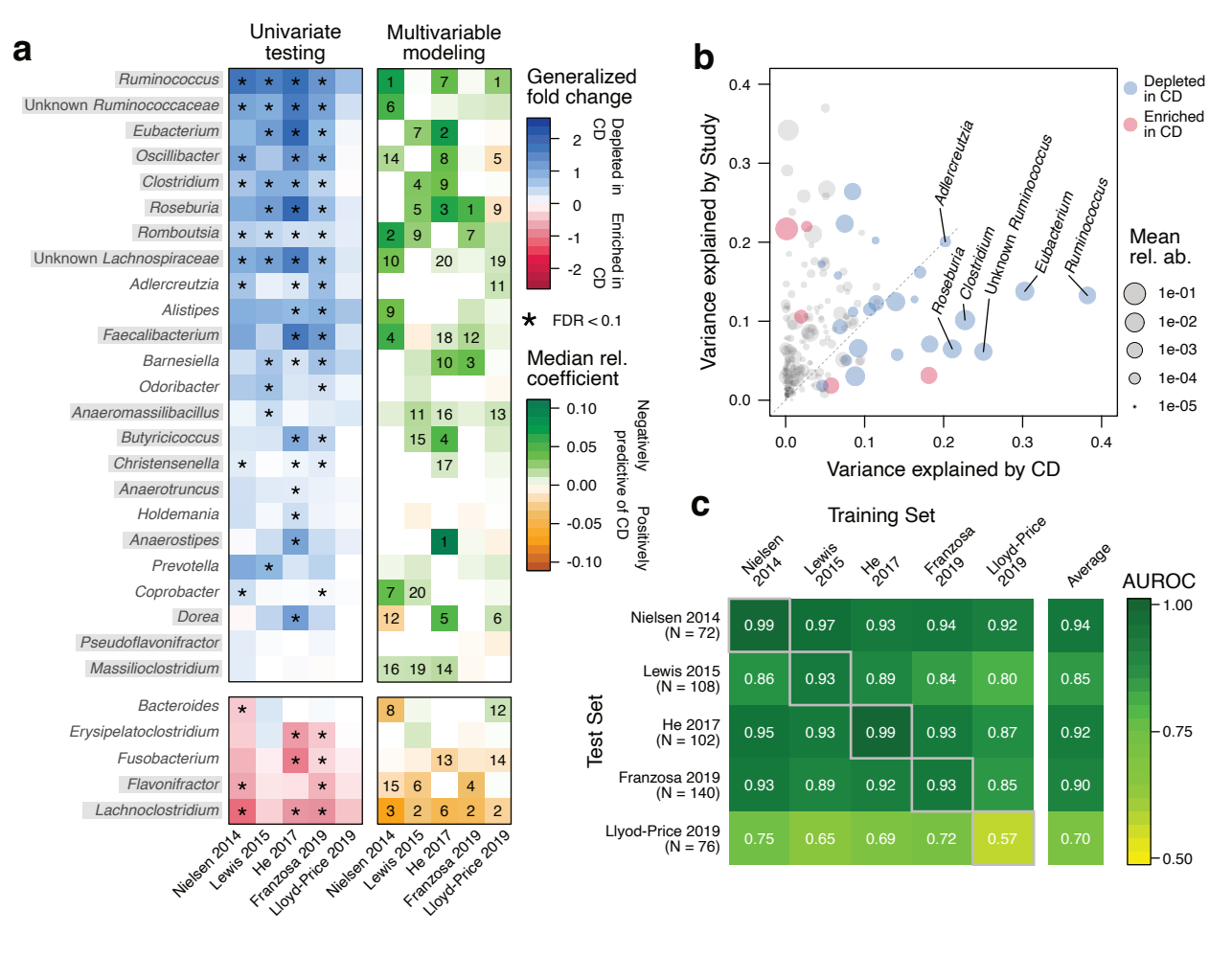

a
Data Input
Features
Label
Metadata (opt)
Example dataset:
Nielsen et al. 2014
Data Filtering
Normalization
Data Splitting
Model Training
Prediction
Association Testing b
Confounder Checks c
Model Evaluation d
Model Interpretation e
b
Lachnospiraceae_bacterium
Peptostreptococcus_stomatis
Parvimonas_unclassified
Adlercreutzia_equolifaciens
Alistipes_shahii
Bacteroides_xylanisolvens
Alistipes_finegoldii
Coprococcus_sp
Ruminococcus_bromii
Akkermansia_muciniphila
Quantiles
40%-60%
30%-70%
20%-80%
10%-90%
UC
Controls
Median
log10(rel. ab.)
log10(q val.)
gFC
c
Wilcoxon p-value: 0.99
BMI
Controls
IBD
d
True positive rate
False positive rate
Mean prediction AUC: 0.763
e
Controls
UC
Median rel. feat weight
Controls (n=59)
UC (n=69)
Peptostreptococcus_stomatis
Clostridiaceae_bacterium
Lachnospiraceae_bacterium
Lactobacillus_ruminis
Bifidobacterium_adolescentis
Clostridium_sp
Lachnospiraceae_bacterium
Peptostreptococcus_anaerobius
Ruminococcaceae_bacterium
Ruminococcus_gnavus
Butyrivibrio_crossotus
Scardovia_unclassified
Lachnospiraceae_bacterium
Eubacterium_rectale
Enterobacter_cloacae
Methanosphaera_stadtmanae
Eggerthella_unclassified
Escherichia_unclassified
Adlercreutzia_equolifaciens
Akkermansia_muciniphila
Bacteroides_xylanisolvens
Classification
Age
BMI
Feature z-score


a
Proportion
1.00
0.75
0.50
0.25
0.00
59
177
69
Controls
UC
Country
Spain
Denmark
P = 4.47e-33
b
Variance explained by country
Variance explained by UC
Lachnospiraceae bacterium
Eubacterium hallii
Dorea formicigenerans
Blautia hydrogenotrophica
Bifidobacterium bifidum
Peptostreptococcus stomatis
Mean rel. ab.
1e-05
1e-04
1e-03
1e-02
c
Spanish and Danish samples −log10(q)
Spanish samples only −log10(q)
Spearman's rho: 0.093
Dorea formicigenerans
d
Dorea formicigenerans log10(rel. ab.)
0.12
5.4e-10
8.5e-09
Danish controls
Spanish controls
Spanish UC cases
e
True positive rate
False positive rate
UC classification
using Spanish and Danish samples
AUROC: 0.84
using Spanish samples only
AUROC: 0.76
f
True positive rate
False positive rate
Classification of Spanish versus Danish control samples
AUROC: 0.96


a
b
c
d
Data filtering needs to be blind to the label.
For standard cross validation, samples need to be independent.
Feature selection needs to be nested into cross validation.
Discovery Dataset
External Dataset
Data Filtering
Normalization
Normalization parameters
Data Splitting
Frozen Normalization
Model Training
Trained models
Prediction & Evaluation
Holdout Prediction
Cross validation (Thomas 2018)
External validation (Zeller 2014)
Incorrect feature selection
Correct feature selection
AUROC
1.0
0.9
0.8
0.7
0.6
0.5
20
100
250
all
Features selected
Cross validation
External validation
Naive CV with dependent samples
Correctly blocked CV
Lloyd-Price 2019
Nielsen 2014
Lewis 2015
He 2017
Franzosa 2019


a
Taxonomic profiles (mOTUs2)
b
Functional profiles (eggNOG4.5)
Shotgun metagenomic sequencing data
Yachida 2019 CRC
Yu 2017 CRC
Wirbel 2019 CRC
Zeller 2014 CRC
Thomas 2019 CRC
Feng 2015 CRC
Vogtmann 2016 CRC
Yachida 2019 ADA
Feng 2015 ADA
Zeller 2014 ADA
Thomas 2019 ADA
Franzosa 2019 UC
Nielsen 2014 UC
Lloyd−Price 2019 UC
Franzosa 2019 CD
Lewis 2015 CD
He 2017 CD
Lloyd−Price 2019 CD
Nielsen 2014 CD
Qin 2012 T2D
Karlsson 2013 T2D
LeChatelier 2013 OB
Kushugolova 2018 MS
Karlsson 2013 IGT
Li 2017 pHT
Li 2017 HT
Jie 2017 ACVD
Loomba 2017 NAFLD
Hoyles 2018 NAFLD
Qin 2014 LIV
Zhang 2015 ART
Wen 2017 AS
Bedarf 2017 PAR
537
318
368
278
385
0 100 200
N samples
0.5 0.6 0.7 0.8 0.9 1.0
AUROC
c
Taxonomic profiles (RDP classifier)
16S rRNA gene sequencing data
Goodrich 2014 OB
Turnbaugh 2009 OB
Zupancic 2012 OB
Ross 2015 OB
Zhu 2013 OB
Baxter 2016 CRC
Zeller 2014 CRC
Wang 2012 CRC
Chen 2012 CRC
Singh 2015 EDD
Schubert 2014 CDI
Schubert 2014 nCDI
Vincent 2013 CDI
Gevers 2014 IBD
Morgan 2012 IBD
Papa 2012 IBD
Willing 2010 IBD
Noguera−Julian 2016 HIV
Lozupone 2013 HIV
Dinh 2015 HIV
Zhang 2013 LIV
Zhu 2013 NAFLD
Wong 2013 NAFLD
Scheperjans 2015 PAR
Son 2015 ASD
Kang 2013 ASD
Alkanani 2015 T1D
Scher 2013 ART
613
256
197
292
283
247
243
239
0 75 150
N samples
0.5 0.6 0.7 0.8 0.9 1.0
AUROC
d
Proportion of variance
0.00 0.25 0.50 0.75
Filtering method
Feature selection method
Feature selection cutoff
Machine learning algorithm
Normalization method
Residuals
e
Random forest
LASSO
Elastic Net
***
**
*
0.50 0.75 1.00
Best AUROC


a
Cases
Controls
Train model on single disease study
Sample additional controls
Independent cohort studies
$\hat{f}(x)$
Apply model on external study
• Determine cross-study portability
• Calculate prediction rate for same/other disease
b
Cross-study portability of augmented models
Cross-study portability of naive models
c
Cross-study portability
Prediction rate for other diseases
Naive models
Augmented models
Prediction rate for:
other diseases
same disease
d
Test set
Training set
AS
ART
CRC
ADA
ACVD
CD
UC
LIV
PD
Wen 2017
Zhang 2015
Zeller 2014
Yu 2017
Yachida 2019
Wirbel 2019
Vogtmann 2016
Thomas 2019
Feng 2015
Jie 2017
Nielsen 2014
Lewis 2015
Lloyd-Price 2019
He 2017
Franzosa 2019
Qin 2014
Loomba 2017
Hoyles 2018
Bedarf 2017
LeChatelier 2013
Karlsson 2013
Qin 2012
Kushugolova 2018
Li 2017
e
PCo1
PCo2
Disease
AS
ART
CRC
ADA
ACVD
CD
UC
LIV
PAR
f
Disease-enriched predictors
Control-enriched predictors
Absolute mean predictor weights
No. of selected predictors


a
Univariate testing
Multivariable modeling
Ruminococcus
Unknown Ruminococcaceae
Eubacterium
Oscillibacter
Clostridium
Roseburia
Romboutsia
Unknown Lachnospiraceae
Adlercreutzia
Alistipes
Faecalibacterium
Barnesiella
Odoribacter
Anaeromassilibacillus
Butyricicoccus
Christensenella
Anaerotruncus
Holdemania
Anaerostipes
Prevotella
Coprobacter
Dorea
Pseudoflavonifractor
Massilioclostridium
Bacteroides
Erysipelatoclostridium
Fusobacterium
Flavonifractor
Lachnoclostridium
Nielsen 2014
Lewis 2015
He 2017
Franzosa 2019
Lloyd-Price 2019
Generalized fold change
Depleted in CD
Enriched in CD
* FDR < 0.1
Median rel. coefficient
Negatively predictive of CD
Positively predictive of CD
b
Variance explained by Study
Variance explained by CD
Depleted in CD
Enriched in CD
Mean rel. ab.
1e-01
1e-02
1e-03
1e-04
1e-05
Adlercreutzia
Roseburia
Clostridium
Unknown Ruminococcus
Eubacterium
Ruminococcus
c
Training Set
Test Set
Nielsen 2014 (N = 72)
Lewis 2015 (N = 108)
He 2017 (N = 102)
Franzosa 2019 (N = 140)
Llyod-Price 2019 (N = 76)
Average
AUROC