SIAMCAT: user-friendly and versatile machine learning workflows for statistically rigorous microbiome analyses.

Jakob Wirbel
Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences
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

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

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

Abstract

The human microbiome is increasingly mined for diagnostic and therapeutic biomarkers. However, computational tools tailored to such analyses are still scarce. Here, we present the SIAMCAT R package, a versatile and user-friendly toolbox for comparative metagenome analyses using machine learning (ML), statistical tests, and visualization. Based on a large meta-analysis of gut microbiome studies, we optimized the choice of ML algorithms and preprocessing routines for default workflow settings. Furthermore, we illustrate common pitfalls leading to overfitting and show how SIAMCAT safeguards against these to make statistically rigorous ML workflows broadly accessible. SIAMCAT is available from siamcat.embl.de and Bioconductor.

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

Background

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 rigor 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), 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 (13–15). To overcome the limitations of testing-

based approaches, several researchers have explicitly built ML classifiers to distinguish case and control samples (16–22); 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 104 classification tasks from 50 gut metagenomic studies 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. Additionally, we illustrate how several common pitfalls of ML applications can be avoided using the statistically rigorous approaches implemented in SIAMCAT. Lastly, we showcase how SIAMCAT facilitates meta-analyses in an application to fecal shotgun metagenomic data from several studies of Crohn's disease. SIAMCAT is implemented in the R programming language and freely available from siamcat.embl.de or Bioconductor.

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. In SIAMCAT, design and parameter choices are carefully adapted to metagenomic data analysis to increase ease of use. 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**). The input for the package 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 **Supplementary Note 1**), we analysed a representative dataset (23) consisting of 128 fecal metagenomes from patients with ulcerative colitis (UC) and non-UC controls (**Figure 1**). As input, we used species-level taxonomic profiles available through the *curatedMetagenomicsData* R package (24).

After data preprocessing (unsupervised abundance and prevalence filtering, **Figure 1a** and Methods), the *check.associations* function tests for associations of single species with the disease using the nonparametric Wilcoxon test and visualizes the results. The association plot displays the distribution of microbial relative abundance, the significance of the association, and a generalized fold change as non-parametric measure of effect size (25) (**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 (26–28). 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 flexible 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 (18,19). 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**).

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 (11,29–31). To minimize this risk, SIAMCAT provides a function to optionally examine potential confounders among the provided meta-variables. In the example dataset from (23), 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, the SIAMCAT *check.confounder* function 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**). 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 the *check.association* function of SIAMCAT to compute statistical association for the full dataset (including the Danish control samples) and the reduced dataset containing only samples from Spanish individuals. 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.
In summary, the *check.confounder* function of 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).

Large-scale machine learning meta-analysis
Previous studies that applied ML to microbiome data (16–19) 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 (32) for 26 data sets based on 16S rRNA gene sequencing (19); additionally, taxonomic profiles generated from 14 and 20 shotgun metagenomic data sets using either MetaPhlAn2 (33) or mOTUs2 (34), respectively, as well as functional profiles obtained with HUMAnN2 (35) or with eggNOG 4.5 (36) for the same set of shotgun metagenomic data were included (in total 104 classification tasks, see **Supplementary Table 1** for information about included datasets).
As a first result, we found that given a sufficiently large input data set (with at least 100 samples), SIAMCAT models are generally able to accurately distinguish between cases and controls: the majority (55%) of these datasets in our analysis could be classified with an AUROC of 0.8 or higher – compared to only 12% of datasets with fewer than 100 samples (**Figure 3a-e, Supplementary Figure 3** 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 Figure 4**), 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 $P = 1$, see **Supplementary Figure 5**). When comparing taxonomic and functional profiles derived from the same dataset, we found a high correlation between AUROC values (Pearson's r = 0.91, $P = 1*10^{-19}$), although on average taxonomic profiles performed slightly better than functional profiles (**Supplementary Figure 5**). 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 offers implementations of several ML algorithms. 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 3f** 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 regression (**Supplementary Figure 6**). 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 (18), this analysis indicates that on average Elastic Net regression outperforms LASSO regression and the Random Forest classifier when considering the optimal choice of ML algorithm ($P = 0.004$ comparing Elastic Net to LASSO and $P = 8*10^{-09}$ comparing it to Random Forest, **Figure 3g**). In summary, this large-scale analysis demonstrates the versatility of the ML workflows provided by SIAMCAT and comprehensively validates its default parameters and how sensitive classification accuracy is to deviations from these.

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 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 performance estimates (also called overfitted). Two pitfalls that can lead to overfitting and poor generalization to other datasets (**Figure 4a**) are frequently encountered in ML analyses of microbiome and other biological data, even though they are well described in the statistics literature (37–39). 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 4b**).

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 (37). 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 4c**). 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.

The second issue tends to occur when samples are not independent (38). 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 7**). 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 (40). 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 4d**). SIAMCAT offers the possibility to block the cross validation according to meta-variables by simply providing an additional argument to the respective function call.

Meta-analysis of Crohn's disease gut microbiome studies

Microbiome disease associations being reported at an ever-increasing pace have provided opportunities for comparisons across multiple studies of the same disease to assess the robustness of associations and the generalizability of ML models in so-called meta-analyses (18,19,25,41). Meta-analyses will be critical to help resolve the debate about spurious associations and reproducibility issues in microbiome research (42). To demonstrate how SIAMCAT enables cross-study comparisons, we analyzed five metagenomic datasets (5,23,40,43,44) which all included samples from patients with Crohn's disease (CD) as well as controls not suffering from inflammatory bowel diseases (IBD). Raw sequencing data were downloaded from ENA and consistently processed to obtain genus abundance profiles with mOTUs2 (34).

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 5a,** 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 data from (40) 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.

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 (25). For many genera, variation can largely be attributed to heterogeneity among studies; the top five associated genera (cf. **Figure 5a**), however, vary much more with disease status, suggesting that their association with CD is only minimally confounded by differences between studies (**Figure 5b**).

Lastly, 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 5c** 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

(40), which exhibited substantially lower accuracy in both cross validation and external validation.

As many ML algorithms allow for an examination for the most influential predictors, microbiome biomarkers can easily be extracted from these models. In (generalised) linear models, such as LASSO or Elastic Net logistic regression models, the model coefficients directly quantify the importance of microbial predictors. Since the LASSO, and to some extent also the Elastic Net, are sparse (also called regularised or penalised) models, the number of influential predictors (with nonzero coefficients) is kept small. As a consequence, these ML methods tend to omit many statistically significant features when they are correlated to each other in favor of a small subset of features with optimal predictive power. Nonetheless, in our meta-analysis of CD the feature importance values derived from multivariable modeling correspond well to the univariate associations, and also show some consistency across the four studies in which clear CD associations could be detected and an accurate ML model trained (**Figure 5a**, 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.

Discussion

The rising interest in clinical microbiome studies and microbiome-derived diagnostic, prognostic, and therapeutic biomarkers also calls for more standardized 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 flaws. Designed with these objectives in mind, SIAMCAT utilizes a modular architecture, allowing 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 4**), while also providing well-validated workflows built from those same modules for novices to explore. To enable rapid integration into R-based microbiome analysis environments, we optimized the interoperability of SIAMCAT with other widely-used tools for microbiome research. Handover of data from DADA2 (45) or phyloseq (46) is straightforward, as SIAMCAT builds on the phyloseq R object. Furthermore, ML models and procedures are based on the mlr package (47), which makes it easy to extend the selection of ML algorithms interfaced in SIAMCAT.

To showcase the power of ML workflows implemented in SIAMCAT and to select default parameters and to assess the robustness of these choices, we performed a meta-analysis of human gut metagenomic studies at considerably larger scale than previous efforts (16–21) (see **Figure 3**). 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 software. Consequently, these benchmarks are expected to yield much more robust and general results than those from previous studies (16–21). 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, but it required the input data to be appropriately normalized (see **Figure 3** and

Supplementary Figure 6). Compared with the choice of normalization method and classification algorithm, other parameters had a considerably lower influence on the resulting classification accuracy.

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 analyzed using SIAMCAT, e.g. to understand associations between community composition or transcriptional activity of the ocean microbiome with physicochemical environmental properties (see **Supplementary Figure 8** for an example (48)) highlighting that SIAMCAT could be of broad utility in microbiome research.

Conclusion

We developed SIAMCAT to make ML-based microbiome analysis easily accessible to the research community. SIAMCAT workflows will help to improve statistical rigor due to safeguards against issues commonly encountered in ML applications that lead to overestimating the accuracy of microbial signatures. While SIAMCAT is broadly applicable to different types of input data and prediction tasks, we anticipate it to be particularly useful for clinical studies exploring the potential of microbial biomarkers for diagnostics, treatment efficacy and safety.

Methods

Implementation

SIAMCAT is implemented as 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) (46). 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 interface to the implementations of different ML algorithms in other R packages (47)**.** 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 in the Supplementary Notes and the online 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 (32) on the basis of 16S rRNA gene sequencing data were downloaded from a recent meta-analysis by Duvallet et al. (19) and summarized at the genus level. MetaPhlAn2 (33) and HUMAnN2 (35) taxonomic and functional profiles were obtained from the *curatedMetagenomicsData* R package (24) 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* (49): after preprocessing with MOCAT2 (50) and filtering for human reads, taxonomic profiles were generated using the mOTUsv2 profiler (34) and functional profiles were calculated by first mapping reads against the integrated gene catalogue (51) and then aggregating the results by eggNOG orthologous groups (36).

Additionally, genus-level taxonomic profiles from the TARA Oceans microbiome project (48) 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 (23) included in the *curatedMetagenomicsData* R package (24). 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 **Supplementary Note 1** 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 **Supplementary Note 1**).

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 normalization parameters, tuning parameters controlling regularization 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 3a-e**), 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 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 (52).

In order to contrast class separation of samples in distance space with the classification performance achieved by ML algorithms (see **Supplementary Figure 4**), 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 4**).

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* (41) obtained through the *curatedMetagenomicsData* R package (24). 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 (40) 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 (5,23,40,43,44) that had been profiled with the mOTUs2 profiler (34) on genus level. While some datasets contained both UC and CD patients (5,23,40), other datasets contained only CD cases (43,44). 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 (23,40,43); specifically for the iHMP dataset (40), 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 **Supplementary Note 2**.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and material
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://10.5281/zenodo.3613794 and the code to reproduce the analysis from https://github.com/zellerlab/siamcat_paper.

Competing interests
No competing interests.

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 a Computational Life Sciences grant no. 031L0181A to G.Z. and P.B.), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project no. 395357507 – SFB 1371 to G.Z.) and the Helmut Horten Foundation (to S.S.).

Authors' 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.

References

1. Schmidt TSB, Raes J, Bork P. The Human Gut Microbiome: From Association to Modulation. Cell. 2018 Mar 8;172(6):1198–215.

2. Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. N Engl J Med. 2016 Dec 15;375(24):2369–79.

3. Garrett WS. The gut microbiota and colon cancer. Science. 2019 Jun 21;364(6446):1133–5.

4. Gevers D, Kugathasan S, Denson LA. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe [Internet]. 2014; Available from: https://www.sciencedirect.com/science/article/pii/S1931312814000638

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. 2019 Feb;4(2):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. 2015 Aug;21(8):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. 2017 Jul 27;18(1):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 [Internet].

2014;10(11). Available from: https://www.embopress.org/doi/abs/10.15252/msb.20145645

9. Tilg H, Cani PD, Mayer EA. Gut microbiome and liver diseases. Gut. 2016 Dec;65(12):2035–44.

10. Zitvogel L, Ma Y, Raoult D, Kroemer G, Gajewski TF. The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. Science. 2018 Mar 23;359(6382):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 Dec 10;528(7581):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 Apr 8;16:73.

13. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011 Jun 24;12(6):R60.

14. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013 Dec;10(12):1200–2.

15. 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. 2015 May 29;26:27663.

16. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. Human-associated microbial signatures: examining their predictive value. Cell Host Microbe. 2011 Oct 20;10(4):292–6.

17. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011 Mar;35(2):343–59.

18. 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 Jul;12(7):e1004977.

19. 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 Dec 5;8(1):1784.

20. 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. 2018 Jun 8;6(1):101.

21. 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. 2019 Jul 15;9(1):10189.

22. Zhou Y-H, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front Genet. 2019 Jun 25;10:579.

23. 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 Aug;32(8):822–8.

24. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017 Oct 31;14(11):1023–4.

25. 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 Apr;25(4):679–89.

26. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Series B Stat Methodol. 1996 Jan 5;58(1):267–88.

27. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol [Internet]. 2005; Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x%4010.1111/%28ISSN%291467-9868.TOP_SERIES_B_RESEARCH

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

29. 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. 2006 Nov 1;17(11):1093–104.

30. Imhann F, Bonder MJ, Vich Vila A, Fu J, Mujagic Z, Vork L, et al. Proton pump inhibitors affect the gut microbiome. Gut. 2016 May;65(5):740–8.

31. 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. 2016 May;65(5):749–56.

32. 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. 2007 Aug;73(16):5261–7.

33. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015 Oct;12(10):902–3.

34. 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. 2019 Mar 4;10(1):1014.

35. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018 Nov;15(11):962–8.

36. 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. 2016 Jan 4;44(D1):D286–93.

37. Smialowski P, Frishman D, Kramer S. Pitfalls of supervised feature selection. Bioinformatics. 2010 Feb 1;26(3):440–3.

38. 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 . 2017 Aug 3;40(8):913–29.

39. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Math Intelligencer [Internet]. 2005; Available from: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

40. 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 May;569(7758):655–62.

41. 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 Apr;25(4):667–78.

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

43. 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. 2015 Oct 14;18(4):489–500.

44. 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. 2017 Jul 1;6(7):1–11.

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

46. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One. 2013 Apr 22;8(4):e61217.

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

48. 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 Metatranscriptome. Cell. 2019 Nov 14;179(5):1068–83.e21.

49. Coelho LP, Alves R, Monteiro P, Huerta-Cepas J, Freitas AT, Bork P. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. Microbiome. 2019 Jun 3;7(1):84.

50. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics. 2016 Aug 15;32(16):2520–3.

51. 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. 2014 Aug;32(8):834–41.

52. Fox J, Weisberg S. An R Companion to Applied Regression. SAGE Publications; 2018. 608 p.

Figure captions

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 **Supplementary Note 1**). Functions producing graphical output (red boxes) are illustrated in (b)-(e) for an exemplary analysis using a data set from Nielsen et al. (23) which contains ulcerative colitis (UC) patients and non-UC controls.

(b) Visualization 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 (25).

(c) SIAMCAT offers statistical tests and diagnostic visualizations 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.

(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 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).

Figure 2. Analysis of covariates that potentially confound microbiome-disease associations and classification models.

The UC data set from Nielsen et al. (23) 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) Visualizations 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 *D. formicigenerans* are significantly larger between countries than between Spanish UC cases and controls (*P*-values from Wilcoxon test).
(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).

Figure 3. 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 MetaPhlAn2 as available from *curatedMetagenomicData* (24). 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 **Supplementary Table 1** for information about the included datasets and key for disease abbreviations.
(b) Application of SIAMCAT machine learning workflows to functional profiles obtained from HUMAnN2 as provided by *curatedMetagenomicData* (24) for the same datasets as in (a) (see **Supplementary Figure 5** for a direct comparison between taxonomic and functional input data).
(c) Disease classification accuracies, similar to (a), obtained from taxonomic profiles generated with mOTUs2 from fecal shotgun metagenomic data (for a direct comparison between MetaPhlAn2 and mOTUs2 input data see **Supplementary Figure 5**).
(d) Disease classification accuracies, similar to (b), obtained from functional profiles generated with eggNOG4 (see **Supplementary Figure 5** for a direct comparison between taxonomic and functional input data).
(e) Cross-validation accuracy of SIAMCAT machine learning workflows as applied to 16S rRNA gene amplicon data for human gut microbiome case-control studies (19) (see (a) for definitions).
(f) 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 was assessed using an ANOVA (see Methods). Dots show the percentage of variance attributable to each parameter and the 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.
(g) 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 (f)). Generally, the choice of algorithm only has a small effect on classification accuracy, but both the Elastic Net and LASSO performance gains are statistically significant (paired Wilcoxon test: LASSO vs Elastic Net, $P = 0.004$; LASSO vs Random Forest, $P = 0.001$; Elastic Net vs Random Forest, $P = 8e\text{-}09$).

Figure 4. SIAMCAT aids in avoiding common pitfalls leading to poor generalization 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 (39). 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 generalization 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 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 generalization 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 generalize across time points, but not across subjects is assessed. To correctly estimate generalization 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.

Figure 5. 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. 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 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.

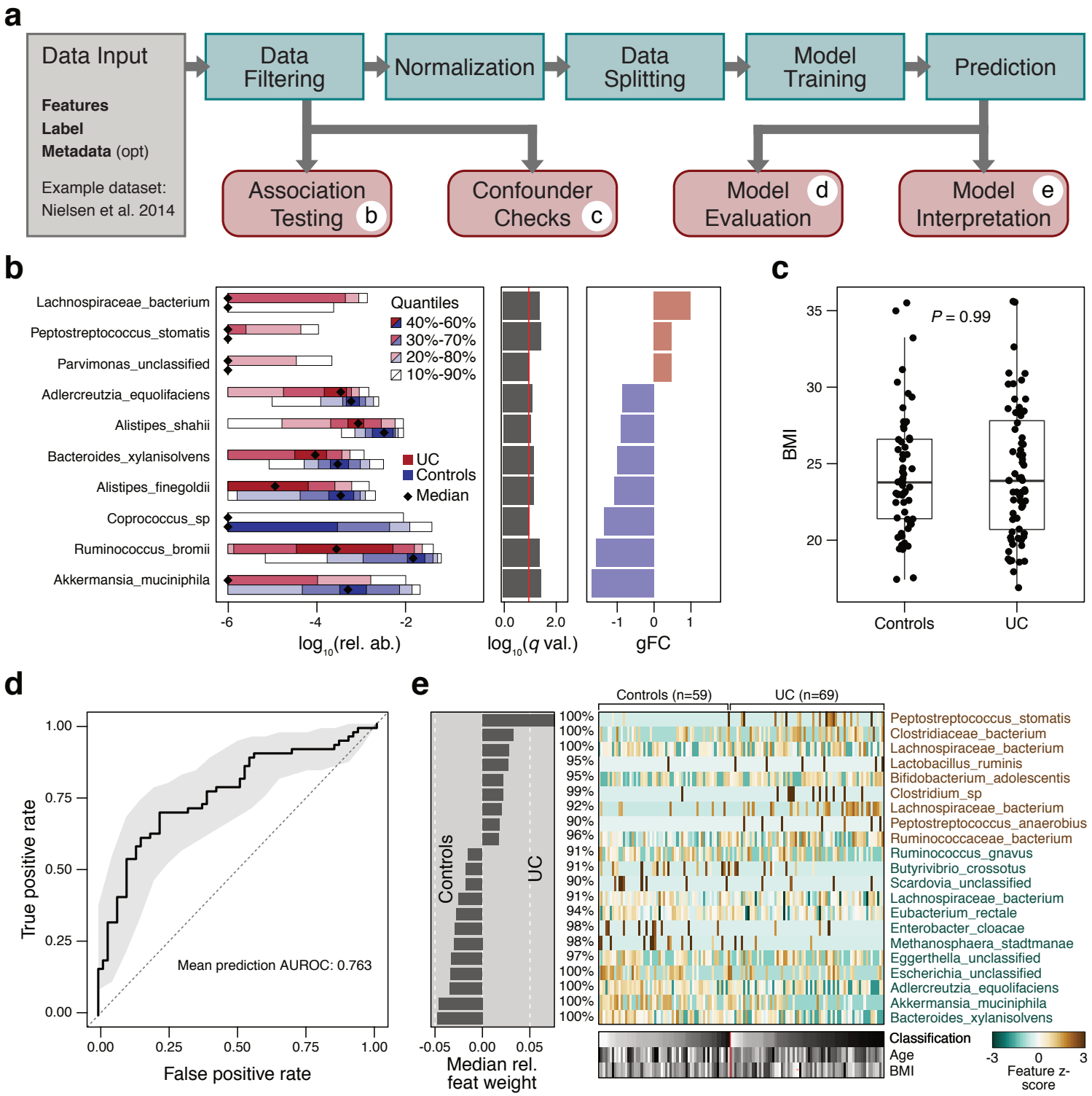

Figure 1

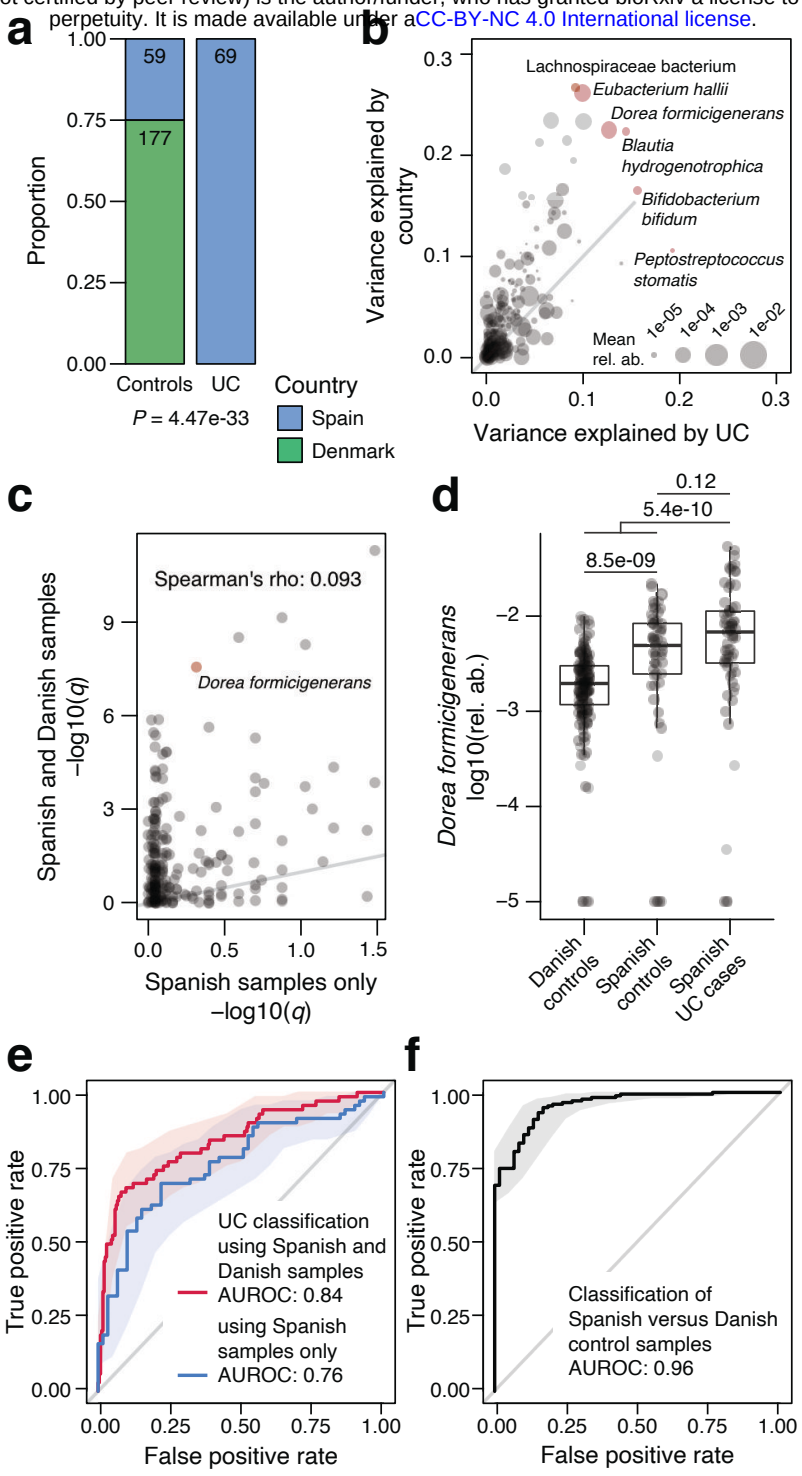

Figure 2

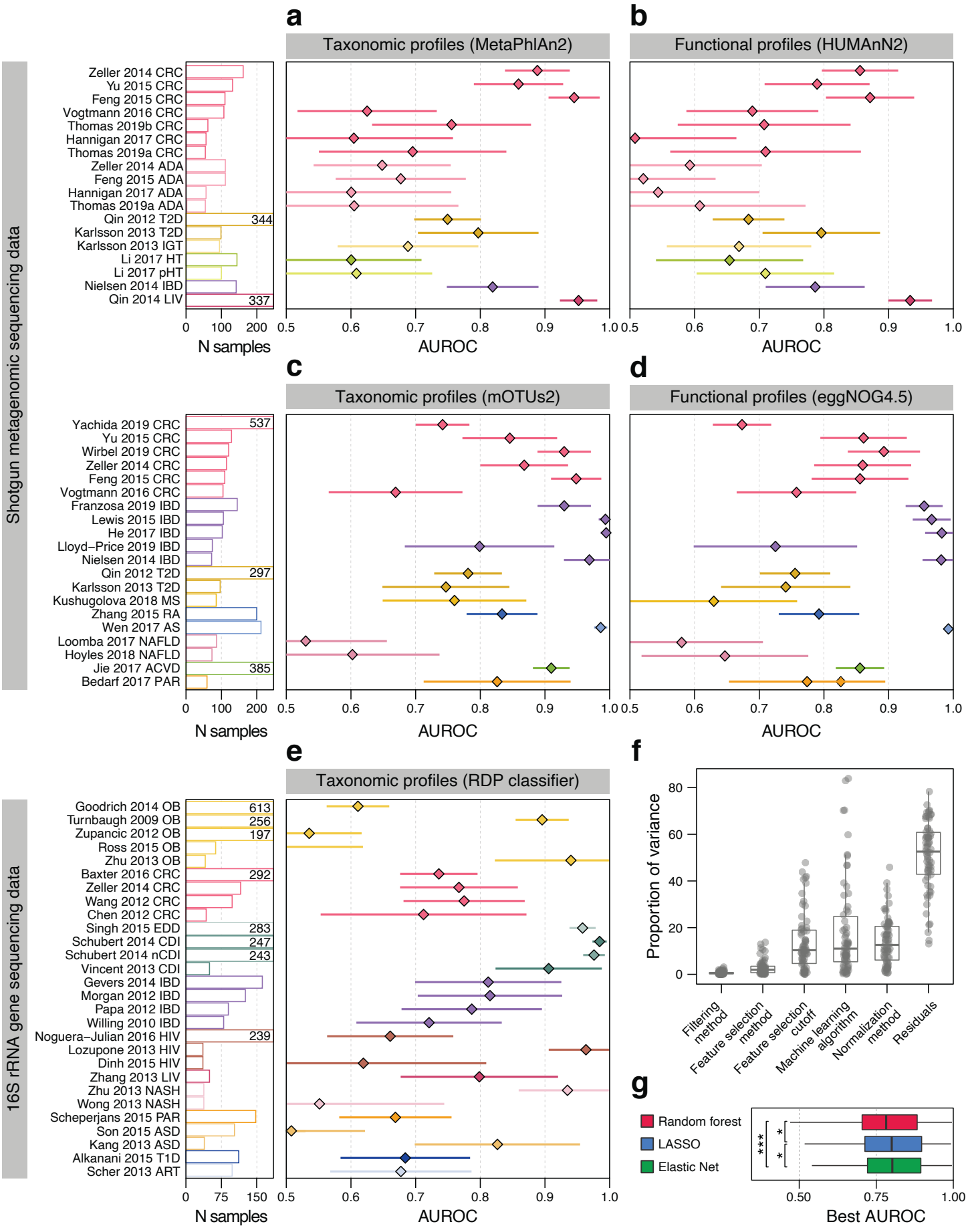

Figure 3

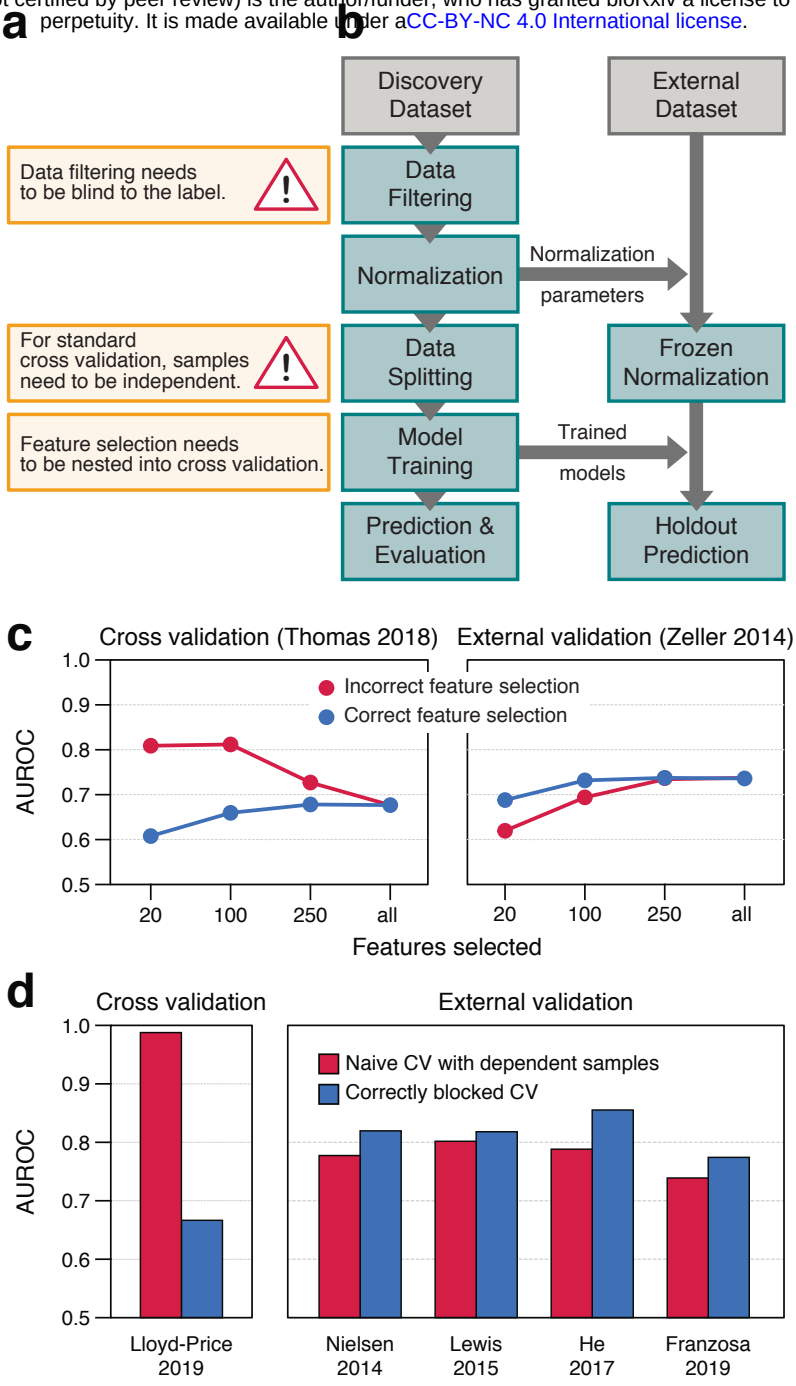

Figure 4

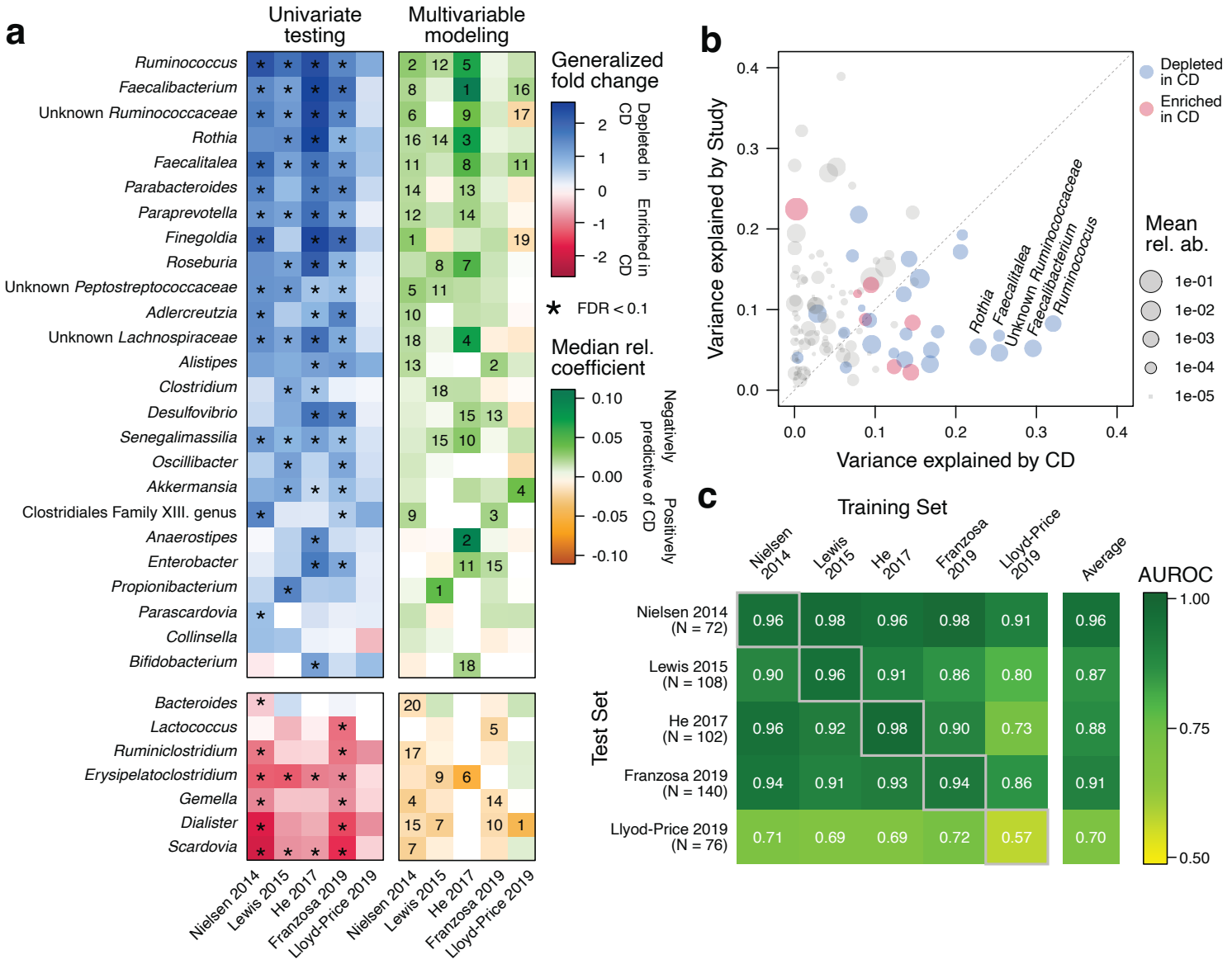

Figure 5