TITLE

Predictive Metagenomic Analysis of Autoimmune Disease Identifies Robust Autoimmunity and Disease Specific Signatures

Angelina Volkova[1], Kelly V. Ruggles*[1,2]

[1]Sackler Institute, Department of Medicine, New York University School of Medicine, New York, NY, USA; [2]Division of Translational Medicine, Department of Medicine and Department of Microbiology, New York University School of Medicine, New York, NY, USA;

ABSTRACT

Within the last decade, numerous studies have demonstrated changes in the gut microbiome associated with specific autoimmune diseases. Due to differences in study design, data quality control, analysis and statistical methods, the results of these studies are inconsistent and incomparable. To better understand the relationship between the intestinal microbiome and autoimmunity, we have completed a comprehensive re-analysis of 29 studies focusing on the gut microbiome in nine autoimmune diseases to identify a specific microbial signature predictive of autoimmune disease using both 16S rRNA sequencing data and shotgun metagenomics data. Despite the heterogeneity of our data set, our approach has allowed us to build robust predictive models for general autoimmunity, as well as models for individual autoimmune diseases. Through this, we identified a number of common features predictive of autoimmune diseases including deficiency in *Alistipes* and *Lachnobacterium*, in addition to 9 inflammatory bowel disease, 7 multiple sclerosis and 7 rheumatoid disease predictive taxa consistently identified across multiple cohort comparison machine learning models. Lastly, we assessed potential metabolomic alterations based on metagenomic/metabolomic correlation analysis, identifying 114 metabolites associated with autoimmunity-predictive taxa.

INTRODUCTION

The human intestine is colonized by millions of microbes, which have been shown to be involved in metabolism[1], immunity[2] and host physiology[3]. This complex ecosystem has been extensively studied in the context of disease[4,5], diet[6–8] and age[9] with the goal of determining how specific taxa and, more recently the gene expression patterns of these taxa, impact human health. The relationship between the microbiome and the immune system has been of particular interest and specific bacteria have been shown to affect the function of both innate and adaptive immunity[10]. Further, an increasing number of inflammatory and autoimmune disorders have been associated with microbial dysbiosis[11], though the precise mechanism for this relationship remains unclear.

Autoimmune diseases are multifactorial and chronic and the term covers nearly 100 distinct disorders[12]. Although there appears to be some genetic component, studies in disease-discordant twins found that concordance rates are incomplete and therefore environmental factors, including the gut microbiome, likely contribute to disease pathogenesis [13,14]. Hundreds of studies have been carried out to better understand the connection between the microbiome and autoimmunity in these diseases including studies specifically focused on inflammatory bowel disease (IBD), multiple sclerosis (MS), rheumatoid arthritis (RA), type 1 diabetes (T1D), and systemic lupus erythematosus (SLE). Despite the extensive study of the human gut microbiome in autoimmune disease, published results are inconsistent, which can be attributed to the differences in origin of samples (e.g. fecal or mucosal), sequencing platforms, sample sizes, therapies administered, patients' age, geographical location, and methods of data analysis. Thus, the question of whether there are common microbial features characterizing general autoimmunity still remains.

Therefore, to better understand the role of specific taxa in autoimmunity, we have reprocessed and reanalyzed 29 16S and metagenomic studies focused on the gut microbiome and autoimmunity. To do this, we have taken advantage of several machine learning

approaches to provide an alternative to the traditional diversity analysis[15–17]. These methods give an advantage of learning functional relationships from the data without a need to define them beforehand. Moreover, many machine learning methods can handle sparse data with a large number of features, ranking them based on importance in their ability to distinguish between health and disease states[18]. These algorithms were used to identify microbial features predictive of general autoimmunity, as well as individual autoimmune diseases through the reanalysis of publicly available data on human gut microbiome in autoimmune diseases from the previous 8 years.

RESULTS

Autoimmunity-associated changes in microbial composition

We used a standardized meta-analysis approach to collect, reprocess and integrate available metagenomics data from case-control autoimmunity studies focusing on changes in the gut microbiome from human fecal samples. Using an expansive literature search we identified a total of 64 total autoimmunity studies fulfilling our criteria. Following filtering based on unique data, age (2 years or older), metadata and raw file availability and sequencing depth we were able to successfully download raw (FASTQ)16S rRNA and/or shotgun metagenomics data from 29 studies, 22 with 16S rRNA sequencing data[19–40] and 5 studies with shotgun metagenomics data[41–45], and 2 studies with both[46,47] (**Supplemental Table 1, Supplementary Fig. 1**). These included studies on Inflammatory Bowel Disease (IBD, *N=12*), Multiple Sclerosis (MS, *N=3*), Rheumatoid Arthritis (RA, *N=2*), Juvenile Idiopathic Arthritis (JIA, *N=3*), Systemic Lupus Erythematosus (SLE, *N=2*), Type 1 Diabetes (*N=2*), Reactive Arthritis/Spondiloarthritis (ReA, *N=1*), Behcet's Syndrome (BS, *N=1*), and Ankylosing Spondylitis (AS, *N=1*). An additional 2 studies with healthy subjects were included to balance the disease and non-diseased cohorts (**Supplemental Table 1**).

Initially, 16S rRNA data was reprocessed using a standard analysis pipeline, which included filtering and taxonomic assignment. Each study was reprocessed individually and final taxonomic abundance tables were then concatenated to a build a final autoimmunity matrix. Disease specific datasets were also created through combining reprocessed data tables for each individual disease type. Each table was then used to build predictive models of general autoimmunity as well as disease-specific models (**Fig. 1**) with the primary goal of identifying the most important features (taxa) involved in autoimmunity across and within disease types. Metagenomics data was also reprocessed using a separate analysis pipeline, providing taxonomic abundance tables (**Supplementary Fig. 2**).

Following quality control (QC) and filtering, 20 16S rRNA studies remained for downstream analysis (**Fig. 1**)[19,20,22–24,26–28,30,31,33–40,46,47]. Notably, 8 out of the 20 studies used investigated the role of the human gut microbiome in IBD, due in part to its relatively high prevalence in 1.3% of US adults[48]. However, we were also able to acquire data from studies of more rare autoimmune diseases including Behçet's Syndrome, which results from inflammation of the blood vessels[19], and Reactive Arthritis. A portion of the studies contained significantly more disease samples than the healthy samples, with Halfvarson *et al.* and Pascal *et al.*, having 10 times more samples from individuals with an autoimmune disease than from healthy controls. For this reason we included healthy samples from two additional studies which investigated non-autoimmune diseases[49,50] , which after QC and preprocessing resulted in 107 additional samples.

While combining of these diverse datasets there were several study-specific characteristics known to impact microbial identification that we paid specific attention to, such as geography, age, sequencing platform and 16S rRNA primers. A majority of the studies were based on populations from North America and Europe, however Manasson et al. investigated the gut microbiome of spondyloarthritis in Guatemalan patients[33]. Further, there was a large range in age across studies, with participants being from 2 to 76 years old. Studies focusing on

newborn children (less than 2 years of age) were removed since it has been well established that the microbial diversity in the first few years of life is significantly lower when compared with adults[51]. DNA was sequenced with one of three sequencing platforms, 454 pyrosequencing, Ion Torrent, or Illumina instruments with both paired and single reads techniques. Description of the characteristics for each study can be found in **Supplementary Table 1** and **Fig. 2**. To assess potential batch effects, we employed a Principal Coordinate Analysis (PCoA)[52] based on the Bray-Curtis distance[53] and investigated non-disease based differences. No significant differences were observed based on autoimmune disease diagnosis, however, there were a subset of non-disease characteristics that were identified as significant based on study, subject characteristics, or sequence methods (**Supplementary Fig. 3**). To combat this, we completed study-based analysis to identify study-specific vs. disease-specific features as part of our downstream analysis (**Supplementary Fig. 6**).

We first examined the taxonomic composition on the genus level of the healthy and diseased samples in each study to verify expected changes based on previously published results. We were able to recapitulate major findings from all studies. For example, we identified disease-specific alterations in multiple studies in *Akkermansia*[39,31]*, Bacteroides*[20,35]*, Blautia*[36,33], *Clostridiaceae*[27], *Faecalibacterium*[28,34], *Lachnospira*[24,28,40], *Parabacteroides*[46], *Prevotella*[40,33,37], *Ruminococcacaea*[24,40,39,33,20] and *Streptococcus*[34] (**Fig. 2**). Interestingly, these previously published results, and our reanalyzed results, varied in the directionality of the change for many of these taxa, with disease specific overabundance occurring in a subset of studies and a reduction in other. These inconsistencies further highlight the need for standardized reanalysis and integration of these valuable datasets to better understand the potential impact of microbial changes in autoimmune disease.

The taxonomic composition of healthy individuals showed clear differences, which can be attributed to several factors. First, it is well established that microbial composition differs by age and geography[54]. Secondly, it is not guaranteed that the “healthy” recruits included in these

studies did not suffer from another pathology impacting the gut microbiome. In most studies, researchers only ensured that healthy controls had not been diagnosed with autoimmune disease of interest and had not taken antibiotics 6 months prior to the sample collection. Thirdly, as these studies were sequenced on different platforms and with differing 16S rRNA hypervariable regions during PCR amplification, we expect a level of variability in the identified taxa even across controls.[55].

Predictive Modeling of Autoimmunity

In order to identify which taxa are most important for distinguishing between healthy controls and subjects with autoimmune disease we employed four independent machine learning disease models: (1) general autoimmunity; which included samples from all the autoimmune diseases identified; (2) IBD specific; (3) MS specific; and (4) Rheumatic Diseases (RD) specific, which included samples from all four rheumatic diseases present in our data set, RA, ReA, JIA, and SLE (**Fig. 1**). Genus level taxonomic abundances were used for the final predictive modeling analyses. Three independent algorithms were used to capitalize on the strengths and limitations of each: Random Forest (RF)[56], Least Absolute Shrinkage and Selection Operator (LASSO)[57], and Support Vector Machine[58] with Recursive Feature Elimination[59] (SVM RFE). Application of three completely independent algorithms capable of feature ranking the same data provided an advantage in robustly identifying the most important features predictive of autoimmunity by multiple models, providing an additional level of confidence.

Model performance was evaluated using both Area Under the receiver operating characteristics Curve (AUC) and macro F1 score, which reports on the reciprocation between the specificity and sensitivity. Notably, we incorporated near-zero variance feature removal to reduce both computational load and to consider only features with reasonable variation between the samples, as those with little variation likely would not impact disease state. Among the three

algorithms for the autoimmunity model, the best performance was achieved by Random Forest with an AUC of 0.829. The superior performance by this algorithm was not unexpected, as Random Forest has been previously shown to perform well on the microbial data[15]. Disease-specific IBD and MS models reached the AUCs of 0.919 and 0.924, respectively (**Fig. 3**). Interestingly, SVM produced the best AUC of 0.902 for the rheumatic diseases' prediction. Overall, the most stable AUC across the three algorithms was reached on the IBD data set, likely due to the considerably higher number of IBD samples compared with other autoimmune diseases. Notably, we were able to predict autoimmunity based on only microbial composition of the samples, which suggests that there is a common gut microbiome signature present relevant to all autoimmune diseases. In order to determine whether our AUCs could be predicted by chance, we assigned the labels to the samples at random, and computed our models again. The models trained with the random label assignment produced the AUCs of ~0.5 (**Supplemental Fig. 5)**, which is indicative of a true difference between the healthy controls and autoimmune disease subjects based on the gut microbial composition.

Most Predictive Model Features

Since all three of our models employed feature ranking we were able to identify which features were most important for predicting general autoimmunity as well as distinct autoimmune diseases. From this, we identified features that were ranked similarly by all three algorithms. The top 30 features were selected from ranked taxa of interest from each model: 61 for general autoimmunity (**Fig. 4a**), 67 for IBD (**Fig. 4b**), 92 for MS (**Supplementary Fig. 6a**) and 70 for RD (**Supplementary Fig. 6b**). In order to account for potential batch affects occurring due to study population differences (**Supplementary Fig. 3**), we created "mock" models to predict the study a sample came from, regardless of disease status. This allowed us to identify taxa that were able to specifically identify a study population rather than the disease. These models identified *Blautia, Lachnospiraceae, Lachnospiraceae Ruminococcus,* and

Faecalibacterium as consistently able to predict study regardless of disease or healthy status (**Supplementary Fig. 7**). This allowed us to identify taxa that are likely tied to the study population, sequencing platform or experimental method, rather than human health.

The most predictive features for our comprehensive autoimmunity analysis identified *Alistipes, Gemmiger, Clostridales, Veilonella,* and *Enterobactericaea* as the most important features, all showing reduced abundance in autoimmune disease samples for all except *Veilonella* which was increased compared with healthy controls (**Fig. 4a**). *Bacteroidales, Dialister, Barnesiella, Veilonella* and *Enterobacteriaceae* were consistently identified as most important in our IBD model, due to increased *Dialister*, *Barnesiella* and *Enterobacteriaceae* and reduced *Dialister* and *Veilonella* in diseased compared with healthy controls (**Fig. 4b**). MS predictive features included *Lachnospiraceae*, *Coriobcteriaceae*, *Methanobrevibacter* and *Butyricicoccus* (**Supplemental Fig. 6a**). Lastly, the RD model identified *Erysipelotrichaceae*, *Alistipes*, *Facalibacterium* and *Odoribacter* and most predictive of disease state (**Supplemental Fig. 6b**). *Lachnospiraceae and Facalibacterium* were identified in the MS and RD models, respectively, but as these were identified as non-specific to disease status (**Supplementary Fig. 7**) we did not consider them as disease-specific taxa.

Models comparing our three disease types (IBD, MS and RD) to each other were also created to further refine our disease specific predictive taxa from our heterogeneous dataset. To do this, we compared each disease to each other, identifying a new set of predictive taxa, and overlapped these with those identified in the original model created based on healthy controls. The model performance (AUC, F1 score) and overlap of the thirty most predictive taxa from each model is shown in **Figure 5**. This analysis provided us with a list of taxa able to distinguish each disease not only from healthy controls, but from other autoimmune diseases. In IBD, nine features were identified in all three comparisons, including *Veilonella*, *Subdoligranulum*, *Ruminococcaceae* and *Gemmiger* (**Fig. 5c**). *Akkermansia, Bifidobacterium* and *Streptophyta* were three of the seven taxa consistently predicted in our MS models (**Fig. 5d**) and

Barnesiellaceae, Enterobacteriaceae, Odoribacter and *Erysipelotrichaceae* were three of the seven identified in all RD models (**Fig. 5e**).

To validate these findings, we also applied the same machine learning approach to shotgun metagenomics data from 7 studies[41–47] (**Supplemental Fig. 2**). Due to data availability, we were only able to build models for general autoimmunity and IBD. Twelve of the top 30 features most predictive features overlapped in both the 16S autoimmunity model (**Fig. 4a**) and metagenomics autoimmunity model (**Supplementary Fig. 6c**), including *Alistipes, Veilonella, Bacteroidales, and Akkermansia.* Similarly, both 16S (**Fig 4b).** and metagenomics (**Supplementary Fig. 6d**) IBD models had 11 overlapping top features including *Bacteroidales, Alistipes, Parabacteroides,* and *Enterobacteriaceae.*

Correlations between highly ranked taxa and metabolism in IBD

To better understand the potential downstream effects of altered abundance levels of these taxa, we used the Inflammatory Bowel Disease Multiomics Database Metabolomic Dataset (IMDMDB) to identify metabolites which are significantly correlated with our taxa of interest. For this purpose, we chose features that overlapped in at least two of the three disease vs disease models that were identified on the genus level (25 taxa total, **Fig. 5c-e**) and which were present in the IBDMDB shotgun metagenomics dataset. This resulted in a total of 10 genera in common between our dataset and IBDMDB cohort (**Fig. 6, Supplementary Fig. 8**). Investigating correlations between the abundance of these 10 genera with metabolites within the IBDMDB, we identified 114 metabolites that significantly correlated with at least one taxa at an adjusted p-value < 0.05. Nineteen of the 114 were identified as being correlated with two taxa, including bile acid metabolites glycoholate and taurine, unsaturated fatty acids linoleate and oleate and branched chain amino acids alloleucine and leucine.

Two of the 10 genera assessed, *Odoribacter* and *Barnesiella*, were both found to be reduced in IBD (**Fig. 5c**). *Odoribacter* abundance was associated with increased levels in three

short and medium chain acylcarnitines and in panothenate (**Fig. 6f**). This is consistent with a recent study showing a depletion of panothenate (vitamin B5) in the gut of IBD subjects[60] as we would expect to see reduced levels of vitamin B5 with reduced *Odoribacter* levels. However, two pf the acylcarnitines identified as being positively associated with *Odoribacter*, C10 carintine and C12:1 carintine, were both found to be significantly increased in subjects with dysbiotic Crohns Disease[60]. As *Odoribacter* was found to be reduced in our IBD population, this is a contradictory finding but may be due to the variability and unknown levels of dysbiosis in our population. *Barnesiella* was found to be negatively associated with seven metabolites involved in Beta-Alanine metabolism (Beta-Alanine, L-Glutamic Acid, L-Aspartic acid, Anserine, Uracil, 3-Methylhistidine, N-carbamoyl-beta-alanine, FDR enrichment < 0.05, **Fig. 6g**). Further, *Barnesiella* had a negative correlation with a number of additional amino acids, including a number known to be preferred by gut bacteria including leucine, isoleucine, lysine and valine[61,62] and a number of polyamines (diacetylspermine, spermidine, N−Acetylputrescine, N1−Acetylspermidine, N1−Acetylspermine), which are known to play an integral role in immunity regulation[63].

Three genera included in the IMDMDB, *Akkermansia*, *Methanobrevibacter* and *Lactococcus,* were found to be predictive of MS. *Akkermansia* had increased abundance in MS samples (**Fig. 5d**) and showed negative associations with the bile acid component taurocholate, bile acid glycocholate and fatty acid anions 3−hydroxyoctanoate and caproate (**Fig. 6a**). *Methanobrevibacter* and *Lactococcus* were also identified as being increased in MS (**Fig. 5d**) and showed highly positive correlations with the anti-inflammatory metabolite 1-methylnicotinamide (**Fig. 6e**) and nicotinuric acid (**Fig. 6i**) respectively. Interestingly, nicotinuric acid has recently been identified as being exclusively found in the stool of patients with IBD[60] but to our knowledge has not been previous reported in association with any other autoimmune disorder. The final two genera investigated, *Paraprevotella* and *Eggerthella,* were both found to be increased in RD (**Fig. 5e**). These were negatively associated with bile acids glycolithocholate

(**Fig. 6b**) and lithocholate (**Fig. 6h**), respectively. *Paraprevotella* was also found to be associated with increased unsaturated fatty acids oleate and linolate (**Fig. 6b**). Further, *Eggerthella* was found to be significantly associated with histamine pathway metabolite N-acetylhistamine (**Fig 6h**). One genera, *Faecalbacterium*, was excluded from the analysis as this was one of the five taxa that was consistently able to predict study regardless of disease or healthy (**Supplementary Fig. 7**)

DISCUSSION

In this analysis, we used data from 29 studies investigating the role of the human gut microbiome in autoimmune disease, assessing both general autoimmunity and specific diseases. We identified a number of genera that were consistently predictive of diseased vs. non-diseased subjects and were able to filter this list based on control models predicting based on study only. Our analysis has recapitulated several recent articles connecting the microbiome with autoimmunity and has identified a number of novel taxa that may be related to these pathologies. For example, two of the most predictive features from our comprehensive autoimmunity analysis were *Alistipes* and *Lachnobacterium*, both of which were recently identified as being markedly decreased with age in adults greater than 50 years old and, as such, potentially associated with host immunity[64]. We also found a depletion in *Clostridiales* and *Ruminococcacaea* in IBD compared with controls, consistent with other studies of IBD[4] and identified *Akkermansia* as a consistently predictive taxa for MS, an organism which has been shown to interact with spore-forming bacteria to worsen the impact of MS-associated microbiota[65].

Further, many of the taxa we identified as being predictive of autoimmune disease were correlated with metabolites that have been previously found to be associated with autoimmunity and inflammation. Recent publications have identified a number of bile acids, triacylglycerols[66], vitamin B, and acylcarnitine[60] metabolites altered in IBD compared with a control population,

many of which we also found to be significantly associated with our most predictive taxa. Histamine, along with taurine and spermine which were also highlighted by our analysis, have been found to help shape the host-microbiome relationship through the regulation of the NLRP6 inflammasome signaling[67]. Surprisingly, we did not identify many short chain fatty acid (SCFA) species, which have been shown to inhibit histone deacetylases (HDACs) and inhibit immune response through Treg regulation and as ligands for G-protein coupled receptors with downstream anti-inflammatory effects[63,68,69]. The association identified between metabolites and taxa could be either due to the impact of that metabolite on the growth of the taxa, the metabolite being a produced by said taxa, or the metabolite negatively associating growth of an inhibitory species, and thus must be followed up by a more targeted approach to understand the precise biological mechanism.

Duvallet et al., completed a similar meta-analysis study in 2017 looking across 10 disease types (arthritis, autism spectrum disorder, Crohn's disease, *Clostridium difficile* infection, liver cirrhosis, colorectal cancer, enteric diarrheal disease, HIV infection, liver diseases, minimal hepatic encephalopathy, non-alcoholic steatohepatitis, obesity, Parkinson's disease, psoriatic arthritis, rheumatoid arthritis, type I diabetes and ulcerative colitis) to identify disease-specific and shared taxa[4]. They too, identified a number of genera associated with more than one disease, including *Lachnospiraceaea and Ruminococcacaea* families and several members of the *Lactobacillales* order and showed the strengths of cross disease comparison using publicly available data. Studies delving into specific disease subcategories, such as this study focused on autoimmune disease, build upon their original study. Further, our reanalysis focused more acutely on investigation of inter-study batch effects and methods of reducing the impact of these on downstream analysis.

We understand there are several limitations of this study. Firstly, the sample size is relatively small for machine learning reducing model reliability. As additional data is generated on larger cohorts from different ages and different cultural backgrounds we can continue to

develop and run similar models to further elucidate how gut microbiome promotes autoimmune diseases. Additionally, the differences in sequencing platform, geography and subject characteristics provides confounders that are difficult to remove from the dataset *post hoc*. Cautious evaluation of taxa identified by our methods in addition to the use of control models testing the ability to predict by study rather than disease were used to combat this issue, however we are aware that these confounders remain. Future analysis further evaluating how each of these study design techniques and participant make-up effects the results of a microbiome study would be of great benefit to the community.

METHODS

Data acquisition

The PubMed database was searched for publications on 03/01/2018 related to the gut microbiome in autoimmune diseases from the last eight years based on the following criteria: 1) the study was performed on human fecal samples; 2) the subjects in the studies were older than 2 years old; 3) the samples were sequenced with either 16S rRNA sequencing or shotgun metagenomics or both; 4) the raw data in FASTQ format were publicly available; 5) the provided metadata allowed us to distinguish between healthy and control samples, as well as between subjects who were explicitly treated in the study and untreated samples. We identified a total of 29 studies, 22 with 16S rRNA sequencing data, 5 with shotgun metagenomics and 2 studies with both types of data available (**Supplemental Table 1**). In order to balance the number of the subjects with autoimmune disease with the number of healthy controls, we added 2 more 16S rRNA studies, from which we selected only the healthy controls.

16S rRNA data preprocessing

We employed QIIME2[70] (v. 2018.11) to obtain the taxonomic abundances of the samples within each study, which were reprocessed independently. Prior to importing the raw data into QIIME2,

we truncated sequences generated with 454 technology to the maximum length of 300 with Trim Galore[71] (v 0.5.0) to better resemble Illumina and Ion Torrent technologies output, after which the 454 data were imported into QIIME2. The sequences generated with Illumina and Ion Torrent were directly imported into QIIME2 with no preprocessing. Following data input, 454-based data underwent an error correcting step with *qiime dada2 denoise-pyro* command while the rest of the samples were processed with either *qiime dada2 denoise-paired* or *qiime dada2 denoise-single* commands depending on whether the reads were paired or single (**Supplementary Table 1**). During this process the bases with quality less than 20 were removed and the paired reads were merged. The resulted sequences abundance tables were rarefied to the depth of 5000. This depth was selected based on the alpha diversity curves of the studies, in which the plot reached the plateau. Further, we tried to account for 454-specific data since the sequencing depth of 454 samples was significantly lower than that of Illumina or Ion Torrent. As a result, the samples with sequencing depth less than 5000 were excluded from the further analysis (**Supplementary Figure 1**). In the next step we assigned the taxonomy to the sequences by training a Naïve Bayes classifier with *qiime feature-classifier fit-classifier-naive-bayes command* based on the Greengenes database[72]. This was done separately for each pair of primers for each 16S rRNA study as different studies sequenced different hypervariable regions of the 16S rRNA gene. Following taxonomy assignment, the taxonomic abundances tables were collapsed on both genus and species taxonomic levels. Further the resulting abundance tables from each study was merged together to create an “autoimmunity” data matrix or a disease-specific matrix. For further analysis only the first time point was selected from each subject, in cases where there were multiple samples per subject, but the time point information was missing, one sample per subject was randomly selected.

Shotgun metagenomics preprocessing

Sequencing reads were trimmed with Trimmomatic[73] (v. 0.36) to have a quality of 20 or greater. KneadData[74] was used to remove host sequences from reads, which were then supplied to the MetaPhlAn2[75] to obtain relative taxonomic abundance, after which tables from individual studies were merged. One exception was the *Cekanavicute et al*. study, for which only preprocessed tables were available, which were processed in the same way. Again, only the first time point of each subject was selected, and the studies where it was not possible to determine the very first time point, one sample from each subject was selected randomly.

Predictive modeling

Caret package[76] in R was used to build the predictive models and models were built separately for each data type. For 16S rRNA we built 4 types of models: autoimmune disease samples vs healthy controls, IBD samples vs healthy controls, multiple sclerosis samples vs healthy controls and rheumatic diseases (rheumatoid arthritis, spondyloarthritis, reactive arthritis, juvenile idiopathic arthritis, systemic lupus erythematosus) vs healthy controls. In addition, we built predictive models comparing IBD and MS, IBD and RD, and MS and RD. Since we identified only 7 studies with publicly available shotgun metagenomics data, we computed only 2 metagenomics models: an autoimmune diseases vs healthy controls model and IBD vs healthy control model. Since there were significantly more healthy samples than diseased samples when considering the individual disease models, we randomly selected the same number of healthy controls samples to match the number of available as diseased samples. The data were split into training (90%) and test (10%) sets. The predictive models for each dataset were built with three models: Random Forest[56], LASSO[57] and SVM[58] with radial kernel and RFE[59] with a step of 2. Those models were selected due to their ability to rank the features based on the importance for the label prediction. To reduce the computing time before the training step the near zero variance features were identified and removed. In order to avoid overfitting, 9-fold-3-times cross-validation was employed to tune the models during the training step.

Feature Selection

Each of the selected algorithms ranked features based on importance to their classification. Since the three algorithms employ different metrics for the feature ranking, first we sorted the features in the ascending order based on importance in each algorithm and then assigned the least important feature a value of 1, while the most important feature got the maximum score equal to the number of features in a given disease model. For the LASSO results, the features with the weight of zero were assigned zero importance. For SVM RFE the mean weights were reported since it was run 3 times with 9 folds. In the next step features were sorted in the descending order according to the summed rankings produced by all three algorithms, with the most important features being assigned the highest rank. Then we selected the top 30 most important features for each disease model.

Metabolomic analysis

We selected taxa that overlapped between at least one disease vs disease models, were identified on the genus level, and were present in the shotgun metagenomics dataset from The Inflammatory Bowel Disease Multiomics Database (IBDMDB)[41]. This method provided 10 different genera. In the next step we correlated the abundance of 10 obtained genera in the IBDMDB with the metabolomics table from IBDMDB by using pairwise Spearman correlation with Benjamini-Hochberg correction for multiple comparisons and selected metabolites based on correlations with an adjusted p-value cutoff of 0.05. MetaboAnalyst was used for Metabolite Set Enrichment Analysis[77] .

Statistical analysis

Mann-Whitney-Wilcoxon test with Benjamini-Hochberg correction for multiple comparisons was utilized for obtaining p values to detect statistical differences in the Principal Coordinate Analysis (PCoA).

REFERENCES

1. Nicholson, J. K. *et al.* Host-Gut Microbiota Metabolic Interactions. *Science* **336**, 1262–1267 (2012).
2. Belkaid, Y. & Hand, T. Role of the Microbiota in Immunity and inflammation. *Cell* **157**, 121–141 (2014).
3. Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R. & Blaser, M. J. Role of the microbiome in human development. *Gut* **68**, 1108–1114 (2019).
4. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1–10 (2017).
5. Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103 (2016).
6. Carmody, R. N. *et al.* Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host Microbe* **17**, 72–84 (2015).
7. Ruggles, K. V. *et al.* Changes in the Gut Microbiota of Urban Subjects during an Immersion in the Traditional Diet and Lifestyle of a Rainforest Village. *mSphere* **3**, (2018).
8. Singh, R. K. *et al.* Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.* **15**, 73 (2017).
9. O'Toole, P. W. & Jeffery, I. B. Gut microbiota and aging. *Science* **350**, 1214–1215 (2015).
10. Honda, K. & Littman, D. R. The microbiota in adaptive immune homeostasis and disease. *Nature* **535**, 75–84 (2016).

11. Levy, M., Kolodziejczyk, A. A., Thaiss, C. A. & Elinav, E. Dysbiosis and the immune system. *Nat. Rev. Immunol.* **17**, 219–232 (2017).

12. Wang, L., Wang, F.-S. & Gershwin, M. E. Human autoimmune diseases: a comprehensive update. *J. Intern. Med.* **278**, 369–395 (2015).

13. Berer, K. *et al.* Gut microbiota from multiple sclerosis patients enables spontaneous autoimmune encephalomyelitis in mice. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10719–10724 (2017).

14. Horta-Baas, G. *et al.* Intestinal Dysbiosis and Rheumatoid Arthritis: A Link between Gut Microbiota and the Pathogenesis of Rheumatoid Arthritis. *Journal of Immunology Research* (2017). doi:10.1155/2017/4835189

15. Statnikov, A. *et al.* A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **1**, 11 (2013).

16. Mossotto, E. *et al.* Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Sci. Rep.* **7**, 2427 (2017).

17. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359 (2011).

18. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer-Verlag, 2013).

19. Consolandi, C. *et al.* Behçet's syndrome patients exhibit specific microbiome signature. *Autoimmun. Rev.* **14**, 269–276 (2015).

20. Dunn, K. A. *et al.* Early Changes in Microbial Community Structure Are Associated with Sustained Remission After Nutritional Treatment of Pediatric Crohn's Disease. *Inflamm. Bowel Dis.* **22**, 2853–2862 (2016).

21. Eun, C. S. *et al.* Does the intestinal microbial community of Korean Crohn's disease patients differ from that of western patients? *BMC Gastroenterol.* **16**, 28 (2016).

22. Pascal, V. *et al.* A microbial signature for Crohn's disease. *Gut* **66**, 813–822 (2017).

23. Goyal, A. *et al.* Safety, Clinical Response, and Microbiome Findings Following Fecal Microbiota Transplant in Children With Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **24**, 410–421 (2018).

24. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2**, 17004 (2017).

25. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).

26. Shaw, K. A. *et al.* Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med.* **8**, 75 (2016).

27. Di Paola, M. *et al.* Alteration of Fecal Microbiota Profiles in Juvenile Idiopathic Arthritis. Associations with HLA-B27 Allele and Disease Status. *Front. Microbiol.* **7**, 1703 (2016).

28. Stoll, M. L. *et al.* Altered microbiota associated with abnormal humoral immune responses to commensal organisms in enthesitis-related arthritis. *Arthritis Res. Ther.* **16**, 486 (2014).

29. Tejesvi, M. V. *et al.* Faecal microbiome in new-onset juvenile idiopathic arthritis. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **35**, 363–370 (2016).

30. Chen, J. *et al.* Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. *Sci. Rep.* **6**, 28484 (2016).

31. Jangi, S. *et al.* Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.* **7**, 12015 (2016).

32. Miyake, S. *et al.* Dysbiosis in the Gut Microbiota of Patients with Multiple Sclerosis, with a Striking Depletion of Species Belonging to Clostridia XIVa and IV Clusters. *PloS One* **10**, e0137429 (2015).

33. Manasson, J. *et al.* Gut Microbiota Perturbations in Reactive Arthritis and Postinfectious Spondyloarthritis. *Arthritis Rheumatol. Hoboken NJ* **70**, 242–254 (2018).

34. Chen, J. *et al.* An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med.* **8**, 43 (2016).

35. Hevia, A. *et al.* Intestinal Dysbiosis Associated with Systemic Lupus Erythematosus. *mBio* **5**, e01548-14 (2014).

36. Luo, X. M. *et al.* Gut Microbiota in Human Systemic Lupus Erythematosus and a Mouse Model of Lupus. *Appl. Environ. Microbiol.* **84**, (2018).

37. Mejía-León, M. E., Petrosino, J. F., Ajami, N. J., Domínguez-Bello, M. G. & de la Barca, A. M. C. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Sci. Rep.* **4**, 3814 (2014).

38. Jacob, V. *et al.* Single Delivery of High-Diversity Fecal Microbiota Preparation by Colonoscopy Is Safe and Effective in Increasing Microbial Diversity in Active Ulcerative Colitis. *Inflamm. Bowel Dis.* **23**, 903–911 (2017).

39. Kump, P. *et al.* The taxonomic composition of the donor intestinal microbiota is a major factor influencing the efficacy of faecal microbiota transplantation in therapy refractory ulcerative colitis. *Aliment. Pharmacol. Ther.* **47**, 67–77 (2018).

40. Mar, J. S. *et al.* Disease Severity and Immune Activity Relate to Distinct Interkingdom Gut Microbiome States in Ethnically Distinct Ulcerative Colitis Patients. *mBio* **7**, e01072-16 (2016).

41. IBDMDB - Home | IBDMDB. Available at: https://ibdmdb.org/. (Accessed: 8th April 2019)

42. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180 (2016).

43. Wen, C. *et al.* Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* **18**, 142 (2017).

44. Lewis, J. D. *et al.* Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host Microbe* **18**, 489–500 (2015).

45. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).

46. Cekanaviciute, E. *et al.* Gut bacteria from multiple sclerosis patients modulate human T cells and exacerbate symptoms in mouse models. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10713–10718 (2017).

47. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *eLife* **2**, e01202 (2013).

48. Data and Statistics. (2019). Available at: https://www.cdc.gov/ibd/data-statistics.htm. (Accessed: 3rd September 2019)

49. Giloteaux, L. *et al.* Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome* **4**, 30 (2016).

50. Whisner, C. M., Maldonado, J., Dente, B., Krajmalnik-Brown, R. & Bruening, M. Diet, physical activity and screen time but not body mass index are associated with the gut microbiome of a diverse cohort of college students living in university housing: a cross-sectional study. *BMC Microbiol.* **18**, 210 (2018).

51. Early life colonization of the human gut: microbes matter everywhere- ClinicalKey. Available at: https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S1369527418300249?returnurl=https:%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1369527418300249%3Fshowall%3Dtrue&referrer=https:%2F%2Fwww.ncbi.nlm.nih.gov%2F. (Accessed: 3rd September 2019)

52. Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338 (1966).

53. Beals, E. W. Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data. in *Advances in Ecological Research* (eds. MacFadyen, A. & Ford, E. D.) **14**, 1–55 (Academic Press, 1984).

54. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

55. Fredriksson, N. J., Hermansson, M. & Wilén, B.-M. The Choice of PCR Primers Has Great Impact on Assessments of Bacterial Community Diversity and Dynamics in a Wastewater Treatment Plant. *PLoS ONE* **8**, (2013).

56. Breiman, L. Random Forests. *Mach Learn* **45**, 5–32 (2001).

57. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).

58. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

59. Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997).

60. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

61. Macfarlane, G. T., Allison, C., Gibson, S. A. & Cummings, J. H. Contribution of the microflora to proteolysis in the human large intestine. *J. Appl. Bacteriol.* **64**, 37–46 (1988).

62. Kim, C. H. Immune regulation by microbiome metabolites. *Immunology* **154**, 220–229 (2018).

63. Rooks, M. G. & Garrett, W. S. Gut microbiota, metabolites and host immunity. *Nat. Rev. Immunol.* **16**, 341–352 (2016).

64. Shen, X. *et al.* Possible correlation between gut microbiota and immunity among healthy middle-aged and elderly people in southwest China. *Gut Pathog.* **10**, (2018).

65. Cekanaviciute, E. *et al.* Multiple Sclerosis-Associated Changes in the Composition and Immune Functions of Spore-Forming Bacteria. *mSystems* **3**, (2018).

66. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).

67. Levy, M. *et al.* Microbiota-Modulated Metabolites Shape the Intestinal Microenvironment by Regulating NLRP6 Inflammasome Signaling. *Cell* **163**, 1428–1443 (2015).

68. Haase, S., Haghikia, A., Gold, R. & Linker, R. A. Dietary fatty acids and susceptibility to multiple sclerosis. *Mult. Scler. J.* **24**, 12–16 (2018).

69. Smith, P. M. *et al.* The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science* **341**, 569–573 (2013).

70. Bolyen, E. *et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science*. (PeerJ Inc., 2018). doi:10.7287/peerj.preprints.27295v2

71. Babraham Bioinformatics - Trim Galore! Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (Accessed: 8th April 2019)

72. DeSantis, T. Z. *et al.* Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Env. Microbiol* **72**, 5069–5072 (2006).

73. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).

74. KneadData | The Huttenhower Lab. Available at: http://huttenhower.sph.harvard.edu/kneaddata. (Accessed: 12th September 2019)

75. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

76. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

77. Xia, J. & Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **38**, W71–W77 (2010).

ACKNOWLEDGMENTS

This work has used computing resources at the NYU High Performance Computing Facility (HPCF).

AUTHOR CONTRIBUTIONS

A.V. completed the literature search, data downloads and QC and constructed and tested the models, A.V. and K.V.R. developed the study design and created figures and tables, and wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

FIGURE LEGENDS:

Figure 1: Autoimmunity Analysis Workflow. Twenty-nine 16S rRNA sequenced datasets from studies focused on 8 different autoimmune diseases and two datasets representing healthy (non-autoimmune disease) cohorts were reprocessed using QIIME2[70] with data2 denoising and rarified to 5000 sequence depth. The output species and genus level relative abundance matrices were used to create four machine learning models for (1) general autoimmunity; (2) Inflammatory Bowel Disease (IBD); (3) Multiple Schlerosis (MS); and (4) Rheumatic diseases including Rheumetoid Arthritis (FA), Reactive Arthritis Spondiloarthritis (RAS), Juvenile Idiopathic Arthritis (JIA) and Systemic Lupus Erythematosus (SLE). Top ranked features from these models were identified and metabolic changes associated with these taxa of interest were assessed using the IBDMDB dataset[41].

Figure 2. Study overview. Overview of (a) 16S rRNA sequencing and (b) shotgun metagenomics studies included in our analysis. Includes the number of healthy and disease samples in each, geographic location, age group and disease studied. Also includes the

average relative abundance of taxa at the genus level for the healthy and diseased subjects following re-processing. Spondyloarthritis/Ankylosing Spondylitis (S/AS)

Figure 3. Predictive Modeling of Autoimmune Disease. Area under the curve (AUC) (a,c) and F1-scores (b,d) for models predicting any autoimmune disease, irritable bowel disease (IBD), multiple sclerosis (MS) and rheumatoid diseases (RD) for three different machine learning models, random forest, support vector machine (SVM) with recursive feature elimination and least absolute shrinkage and selection operator (LASSO) at the genus, species and strain (for metagenomics only) level.

Figure 4. Taxa Predictive of Disease Top 30 taxa across three predictive models, LASSO, random forest (RF) and support vector machines (SVM) for (a) General Autoimmunity and (b) Inflammatory Bowel Disease. Features ranked by mean rank across the three models and color indicates the rank of each taxa in each model. Log fold change of disease vs. healthy for each taxa.

Figure 5. Disease vs. Disease Comparison Models Model (a) AUCs and (b) F1 scores when predicting diseased samples when compared against other disease. Taxa consistently identified in multiple comparison models for (c) irritable bowel disease (IBD), (d) multiple sclerosis (MS), and (E) rheumatoid diseases (RD).

Figure 6. Metabolites significantly correlated with disease-predictive taxa. Spearman correlation coefficient scores plotted and shaded by adjusted p-value for 9 taxa found to be predictive of IBD, MS and RD based on the multiple disease model comparisons.

Supplementary Figure 1: Study collection and filtering

Supplementary Figure 2. Metagenomic Analysis Workflow.

Supplementary Figure 3: PCoA diagrams and statistical differences across 16S datasets showing sample similarity by (a) health status, (b) disease type, (c) age, (d) forward primer, (e) study, (f) geographic location, (g) sequence platform, and (h) reverse primer.

Supplementary Figure 4. PCoA diagrams and statistical differences across metagenomics datasets showing sample similarity by (a) health status, (b) study, (c) disease type and (d) geographic location.

Supplementary Figure 5. Models trained with random label assignment

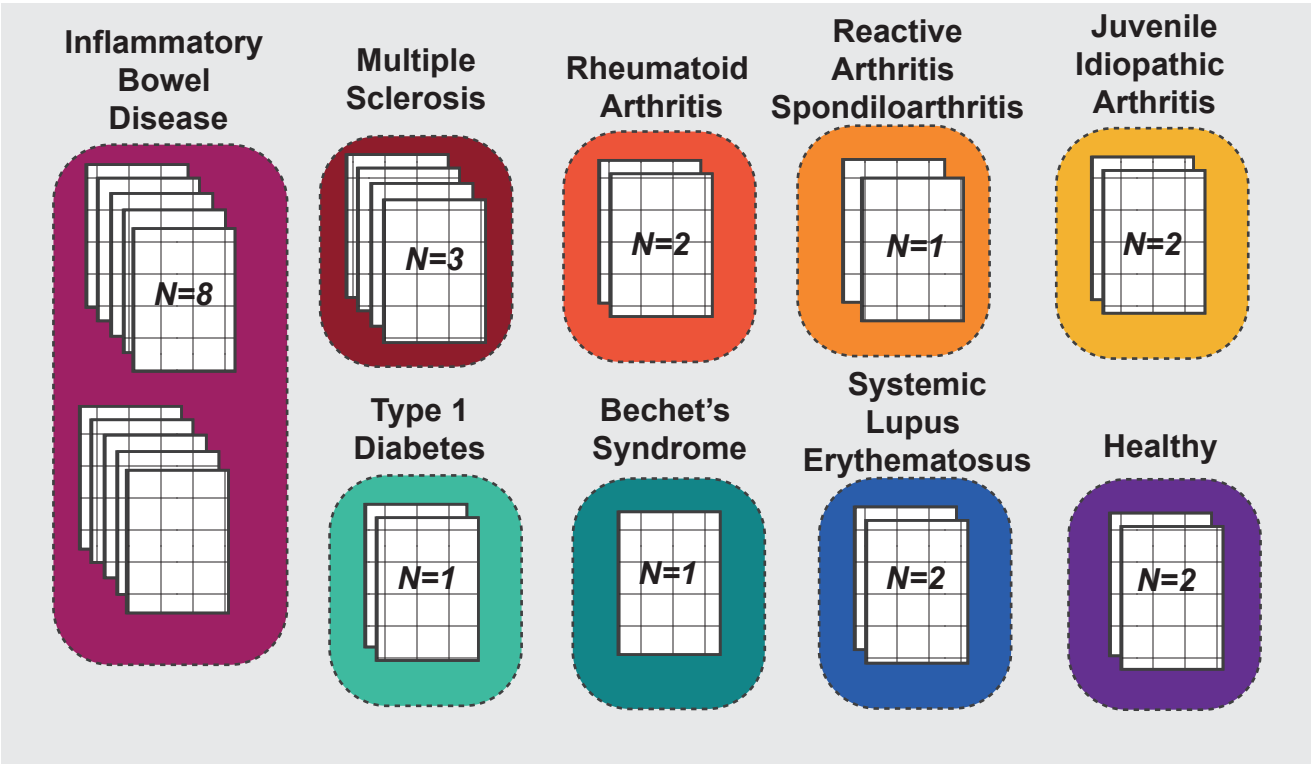
Supplementary Figure 6. Top 30 taxa across three predictive models, LASSO, random forest (RF) and support vector machines (SVM) for (a) General Autoimmunity and (b) Inflammatory Bowel Disease. Features ranked by mean rank across the three models and color indicates the rank of each taxa in each model. Log fold change of disease vs. healthy for each identified taxa also shown.

Supplementary Figure 7. Models and Features Predictive of Study. AUC of random forest model prediction of study and taxa features most predictive of study in those models in 16S (a,c) and metagenomics (b,d) studies.

Supplementary Figure 8. (a) Significant correlations between metagenomic abundance of 10 selected genera and metabolites in IBDMDB dataset. (b) Significant correlation of *Faecalibacterium* abundance, a genus that was predictive of the study, with metabolites from IDBMDB.

TABLES

Supplementary Table 1: Study details including demographics, sample number, sequence primer, sequence platform.



Quality filtering / Sequence trimming

dada2 denoising

Train classifiers for each primer pair

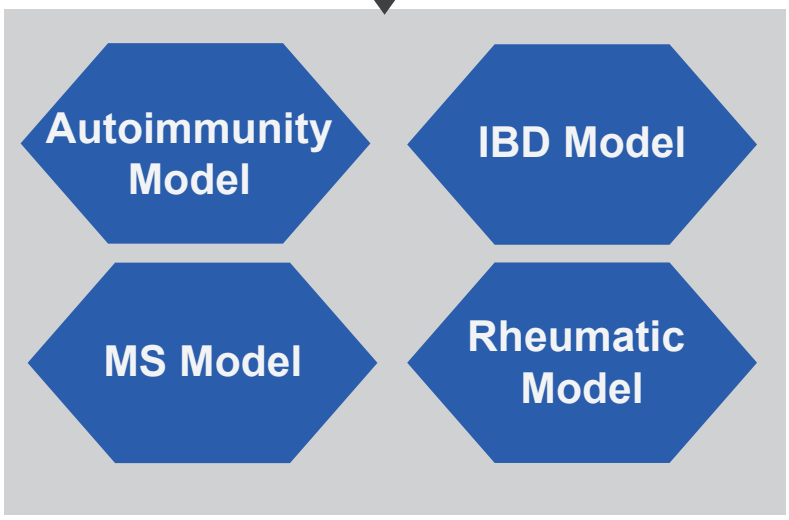
Assign taxonomy

Filter samples < 1000 sequences

Rarify to 5000 sequencing depth

Concatenate tables

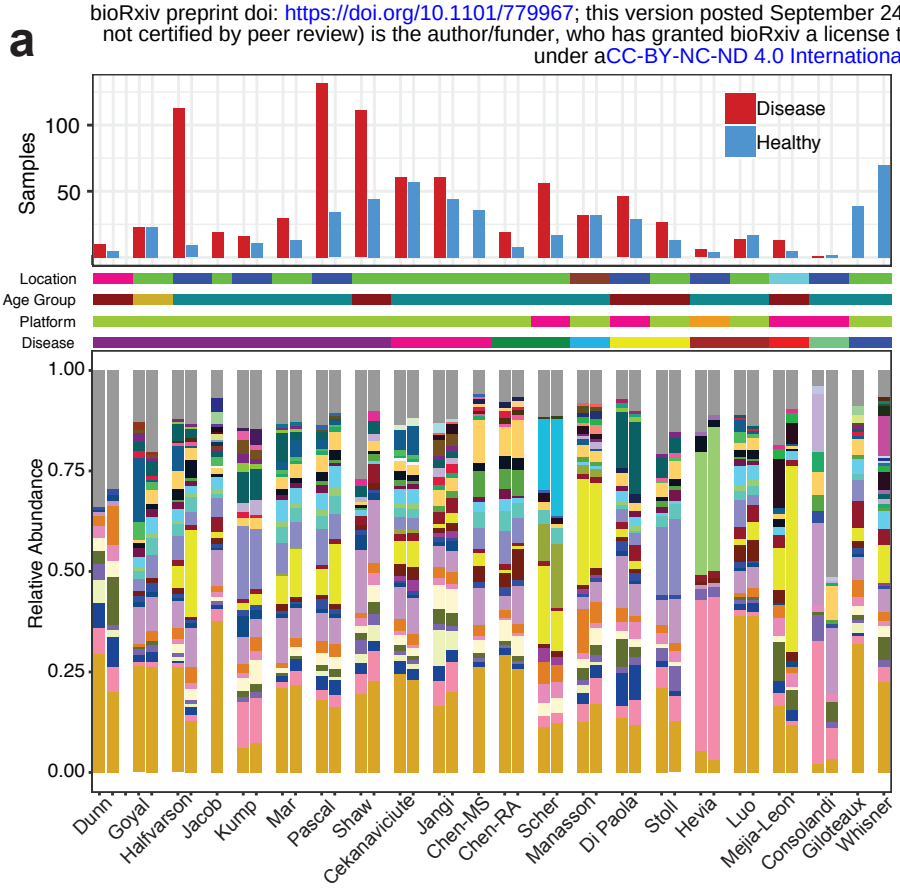
Model Building

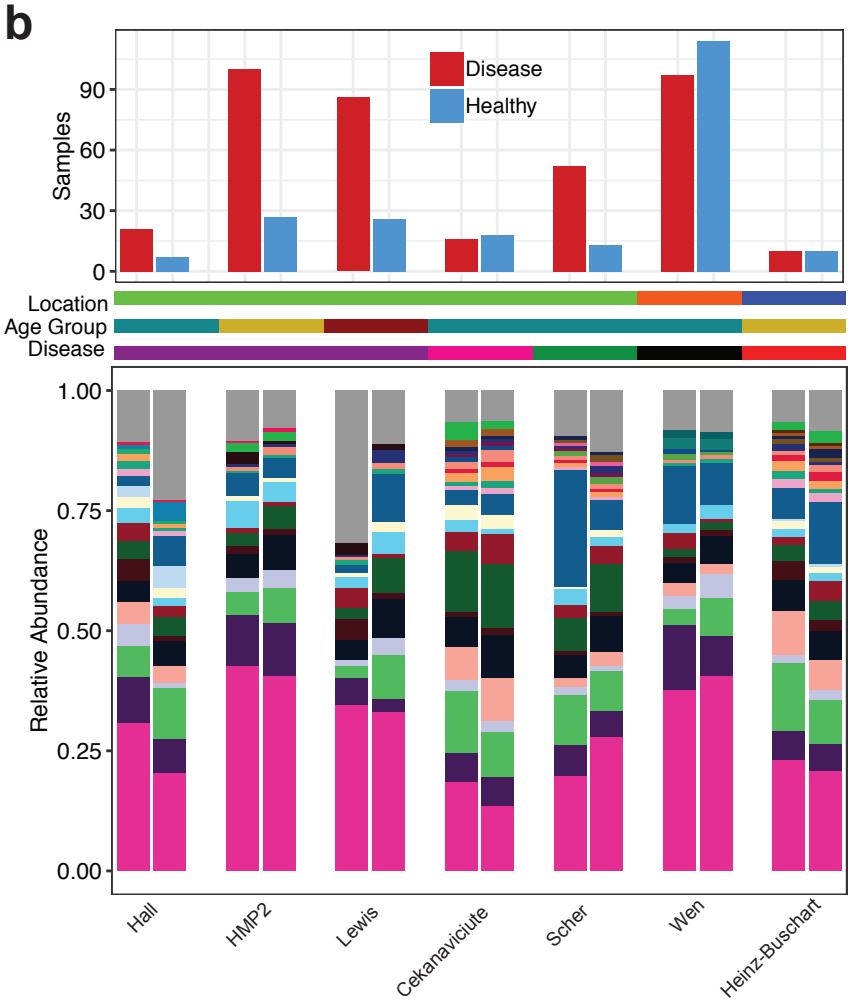


Autoimmunity & disease specific taxa

iHMP IBD dataset

Associated Metabolic Changes





Location
- Canada
- China
- Europe
- Guatemala
- Mexico
- US

Age Group
- Adults
- Both
- Children

Platform
- 454
- Illumina
- IonTorrent

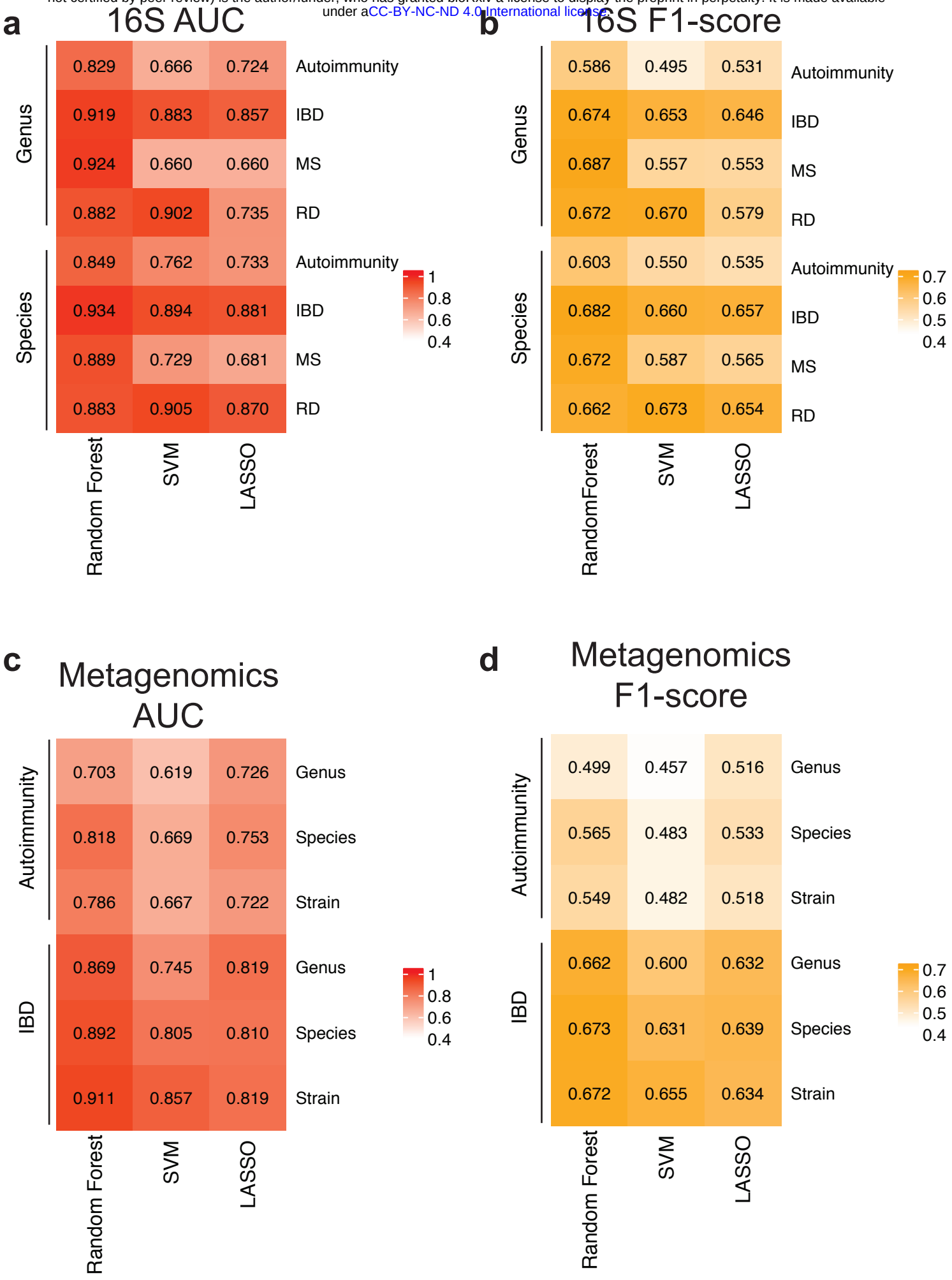
Disease
- Behcets
- IBD
- JIA
- MS
- Other
- RA
- ReA
- SLE
- T1D
- S/AS

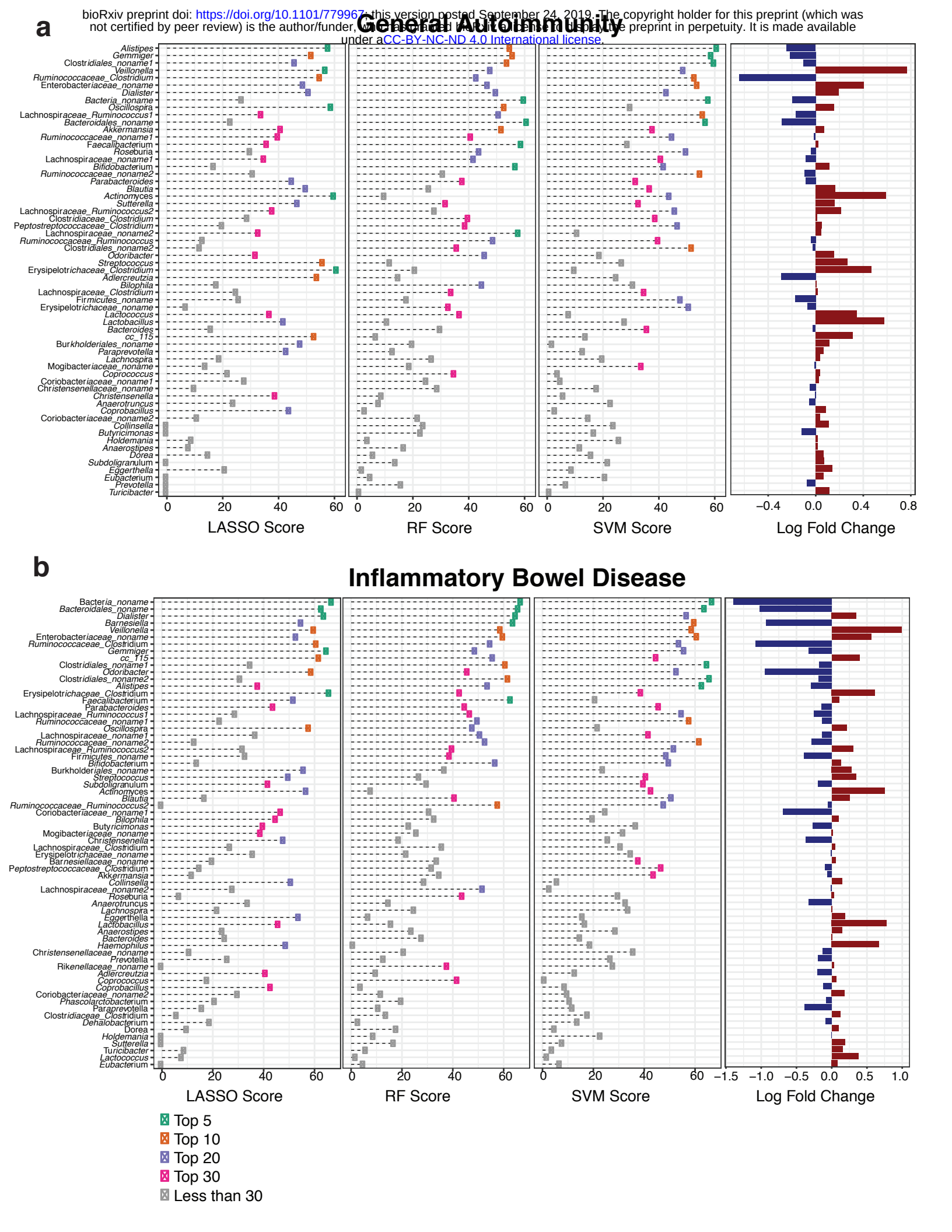
16S Taxa

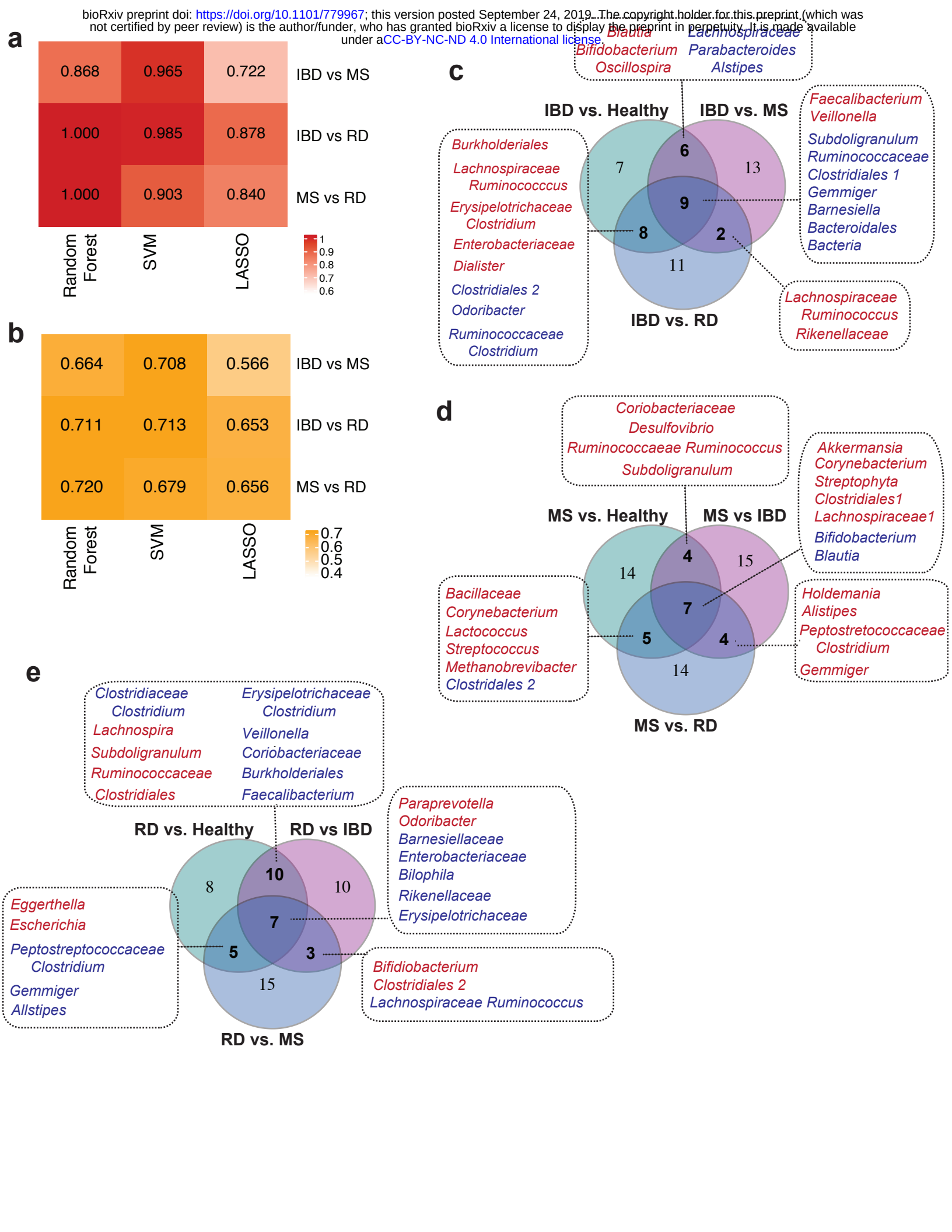
- Bacteroides
- Lachnospiraceae_noname1
- Lachnospiraceae_noname2
- Akkermansia
- Lachnospiraceae_Clostridium
- Ruminococcaceae_noname
- Clostridiales_noname1
- Clostridiales_noname2
- Ruminococcaceae_noname
- Faecalibacterium
- Dialister
- Dorea
- Bacteria_noname
- Parabacteroides
- Prevotellaceae_Prevotella
- Alistipes
- Paraprevotella
- Blautia
- Coprococcus
- Lachnospira
- Roseburia
- Lachnospiraceae_Ruminococcus1
- Gemmiger
- Oscillospira
- Ruminococcaceae_Ruminococcus
- Phascolarctobacterium
- Erysipelotrichaceae_noname
- Sutterella
- Enterobacteriaceae_noname
- Pseudomonas
- Rikenellaceae_noname
- Barnesiellaceae_noname
- Streptococcus
- Christensenellaceae_noname
- Escherichia
- Lachnospiraceae_Ruminococcus2
- Butyricicoccus
- Bifidobacterium
- Collinsella
- Streptococcaceae_noname
- Subdoligranulum
- Clostridiaceae_Clostridium
- Peptostreptococcaceae_Clostridium
- Veillonella
- Barnesiella
- Lactobacillus
- Megamonas
- Bacteroidales_noname
- Paraprevotellaceae_Prevotella
- Veillonellaceae_noname
- Succinispira
- Catenibacterium
- Actinobacillus
- Rikenellaceae_noname
- Burkholderiales_noname
- Eubacterium
- Odoribacter
- Proteobacteria_noname
- Verrucomicrobiaceae_noname
- Turicibacter
- Clostridiaceae_noname
- Peptostreptococcaceae_noname
- Acidaminococcus
- Megasphaera
- Cetobacterium
- Shigella
- Pasteurellaceae_noname
- Marinilabiaceae_noname
- Cytophaga
- Ruminococcaceae_Clostridium
- Sporobacter
- Comamonadaceae_noname
- Oxalobacteraceae_noname
- Less_than_0.5%

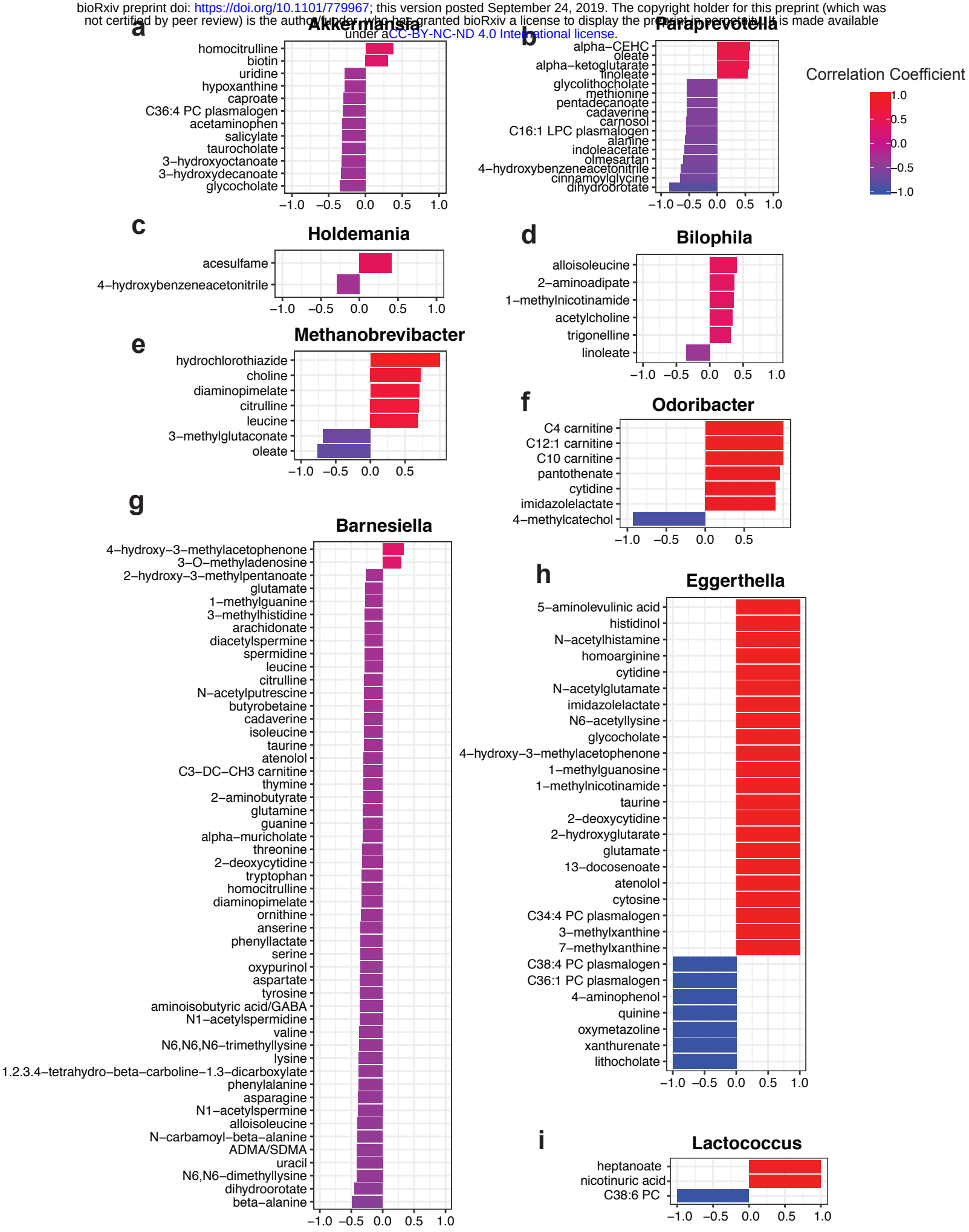
Metagenomics Taxa

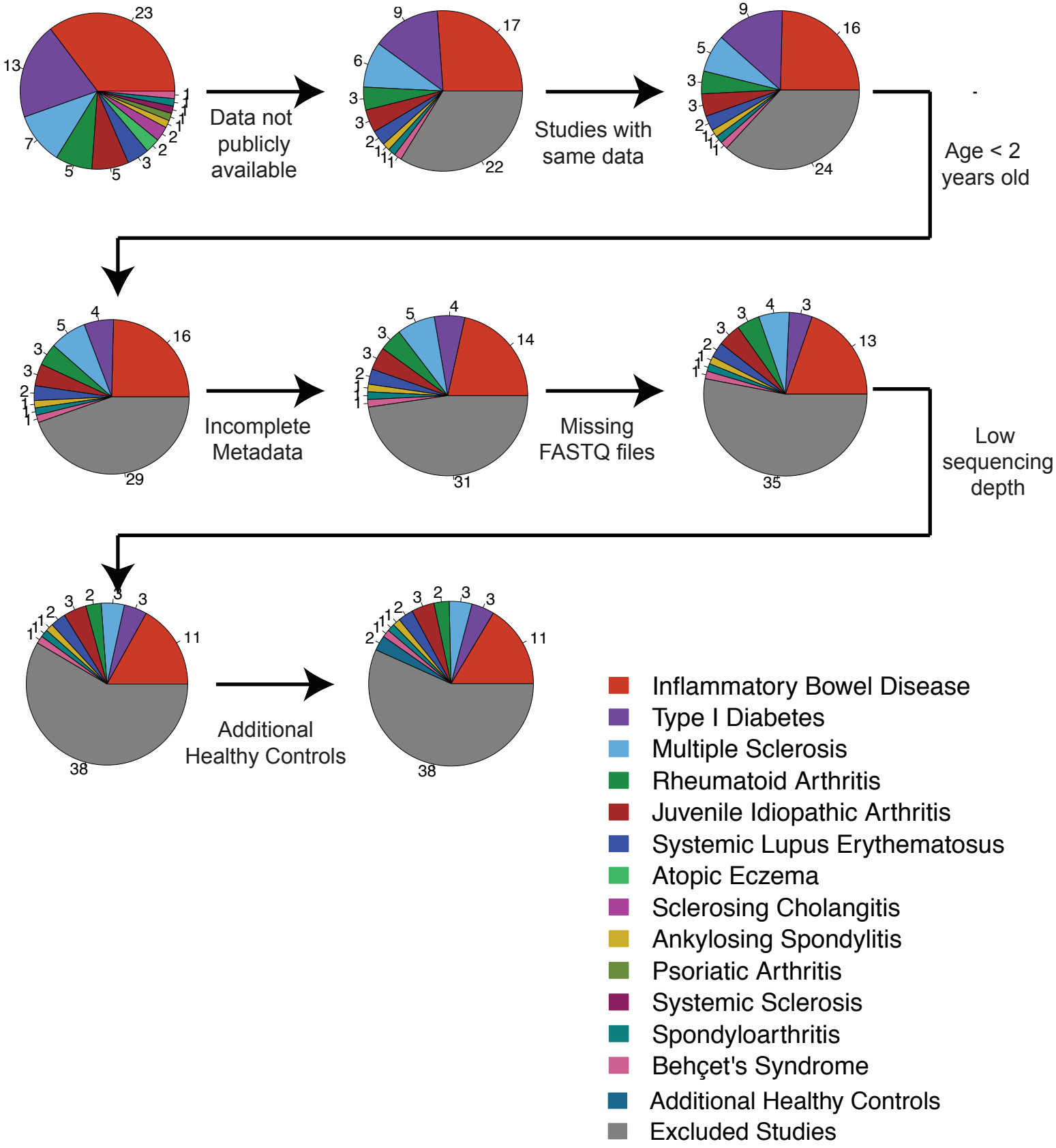
- Bacteroides
- Faecalibacterium
- Eubacterium
- Roseburia
- Ruminococcus
- Alistipes
- Blautia
- Subdoligranulum
- Bifidobacterium
- Parabacteroides
- Dialister
- Streptococcus
- Prevotella
- Dorea
- Lachnospiraceae_noname
- Coprococcus
- Veillonella
- Haemophilus
- Collinsella
- Barnesiella
- Paraprevotella
- Oscillibacter
- Megamonas
- Escherichia
- Bacteroidales_noname
- Odoribacter
- Lactobacillus
- Erysipelotrichaceae_noname
- Phascolarctobacterium
- Clostridium
- Methanobrevibacter
- Catenibacterium
- Akkermansia
- Sutterella
- Less_than_0.5%

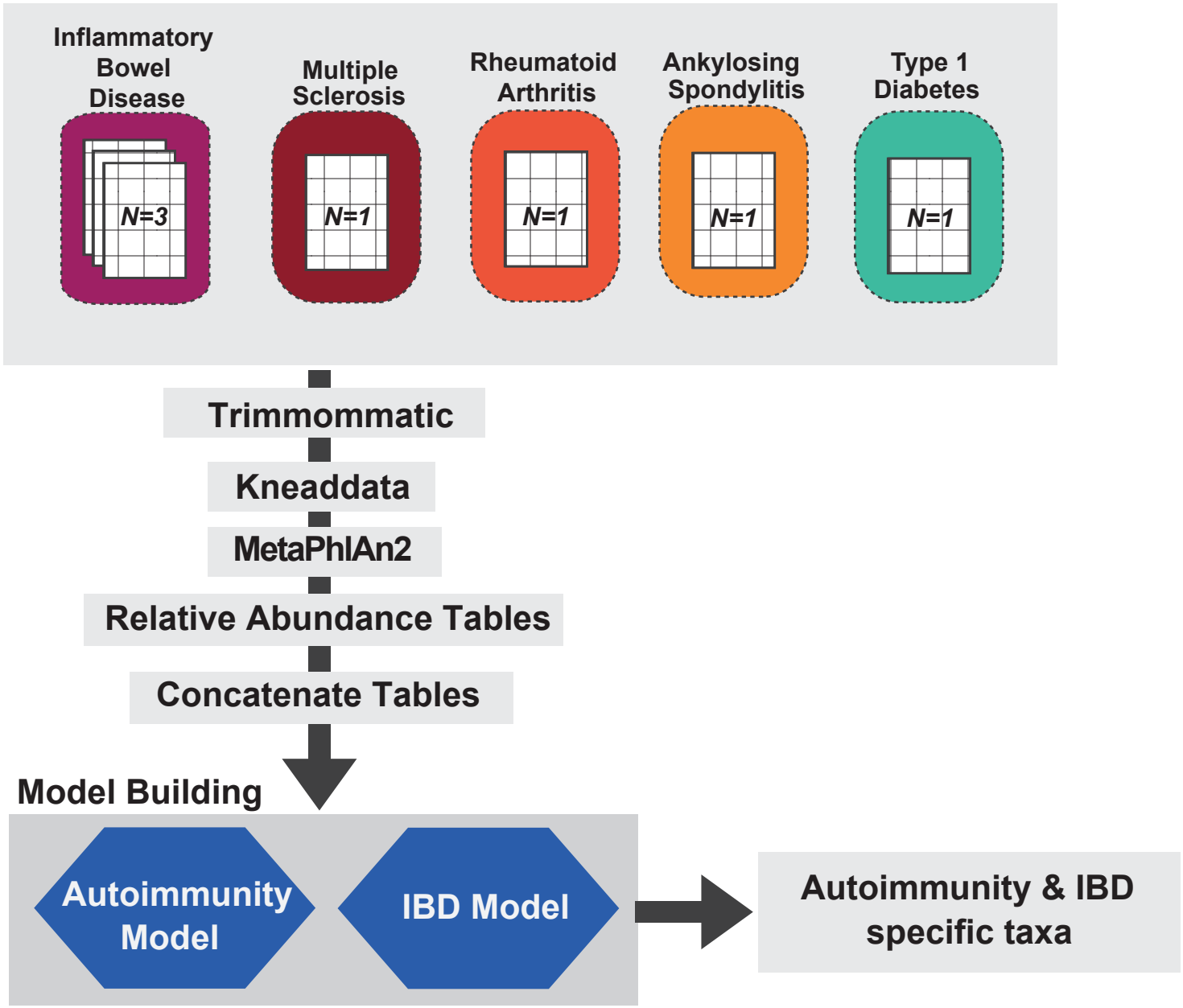




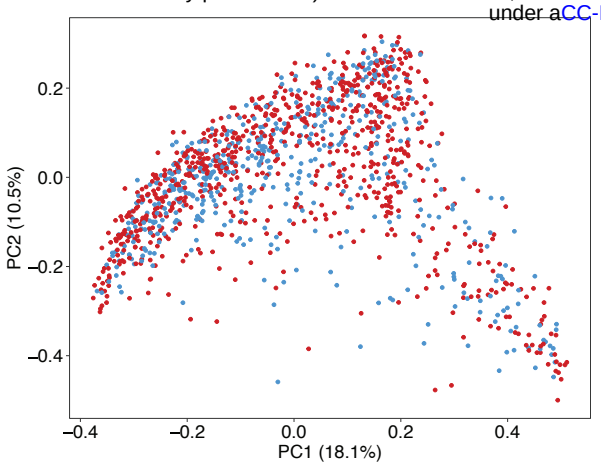






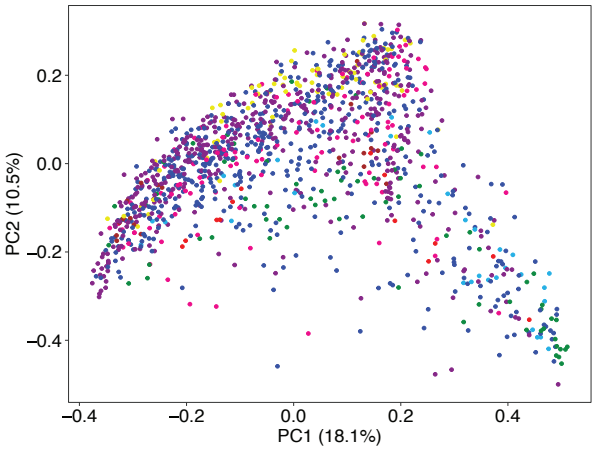


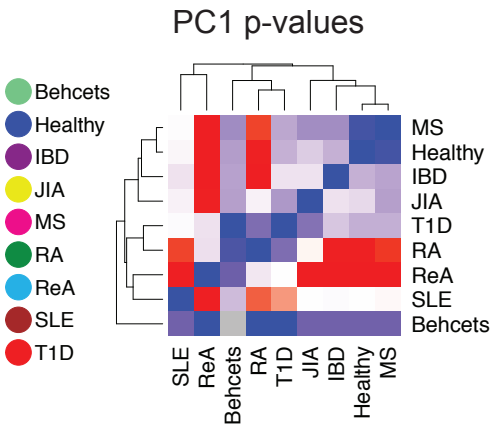
a

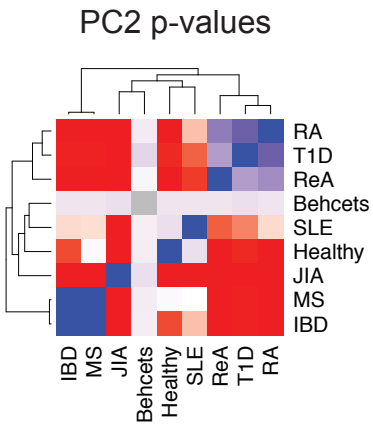


PC1 p-value = 0.33
PC2 p-value = 0.07

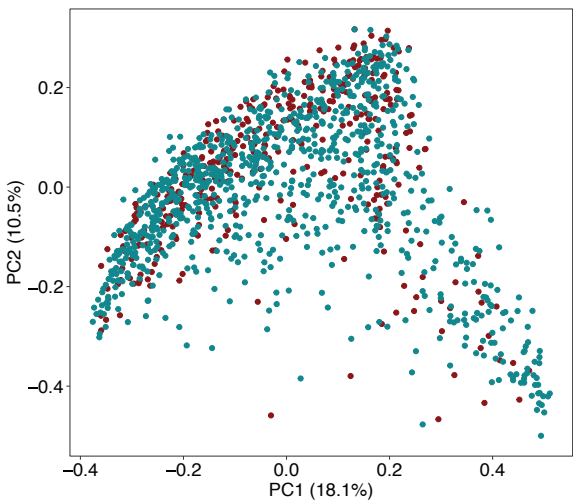
b



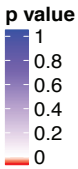




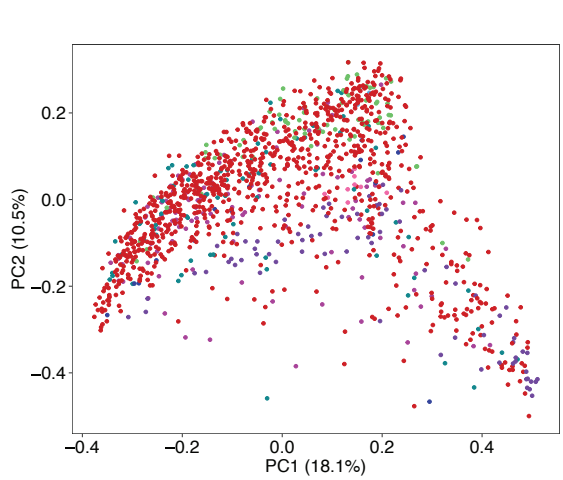
c

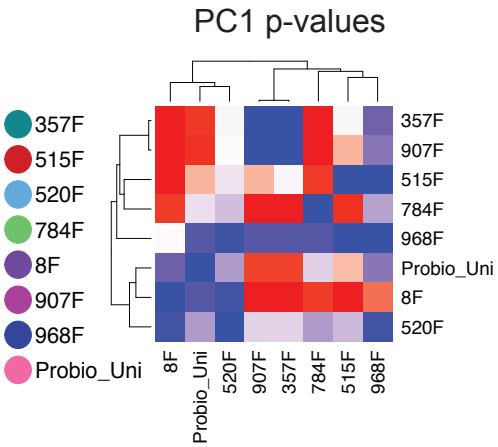


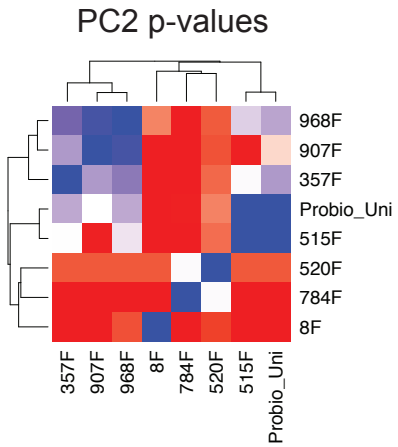
PC1 p-value = 0.86
PC2 p-value < 0.05



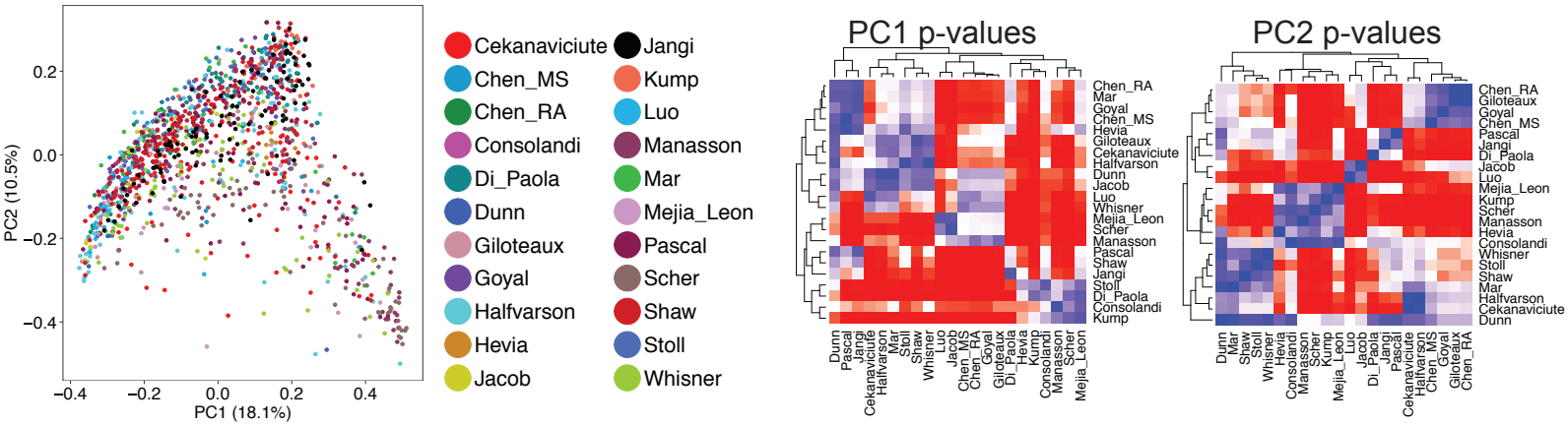
d



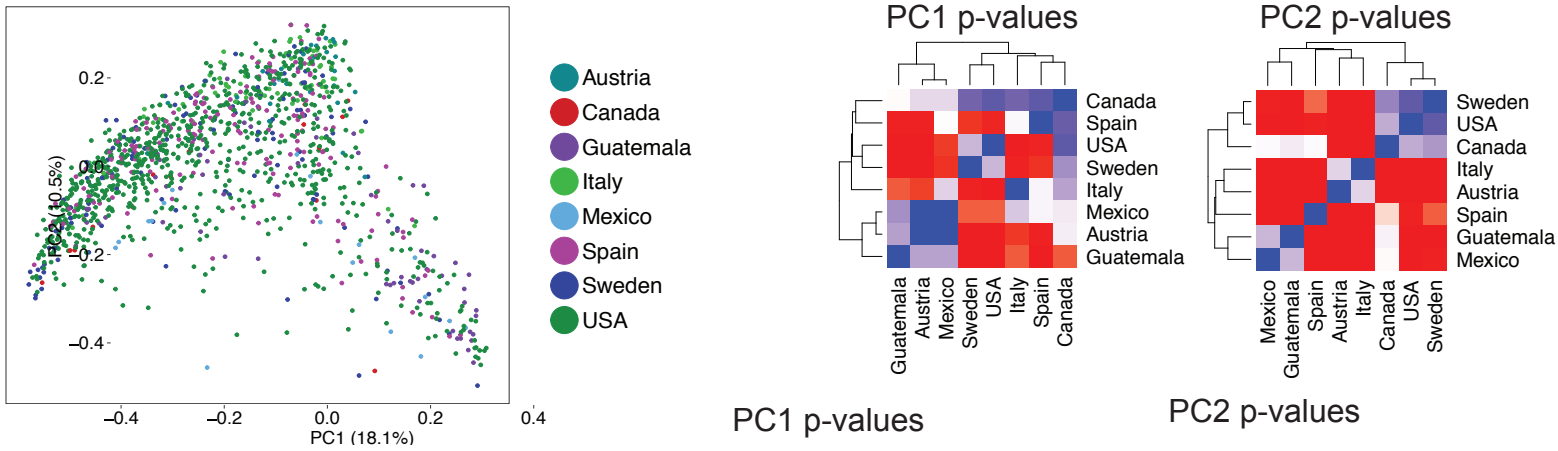




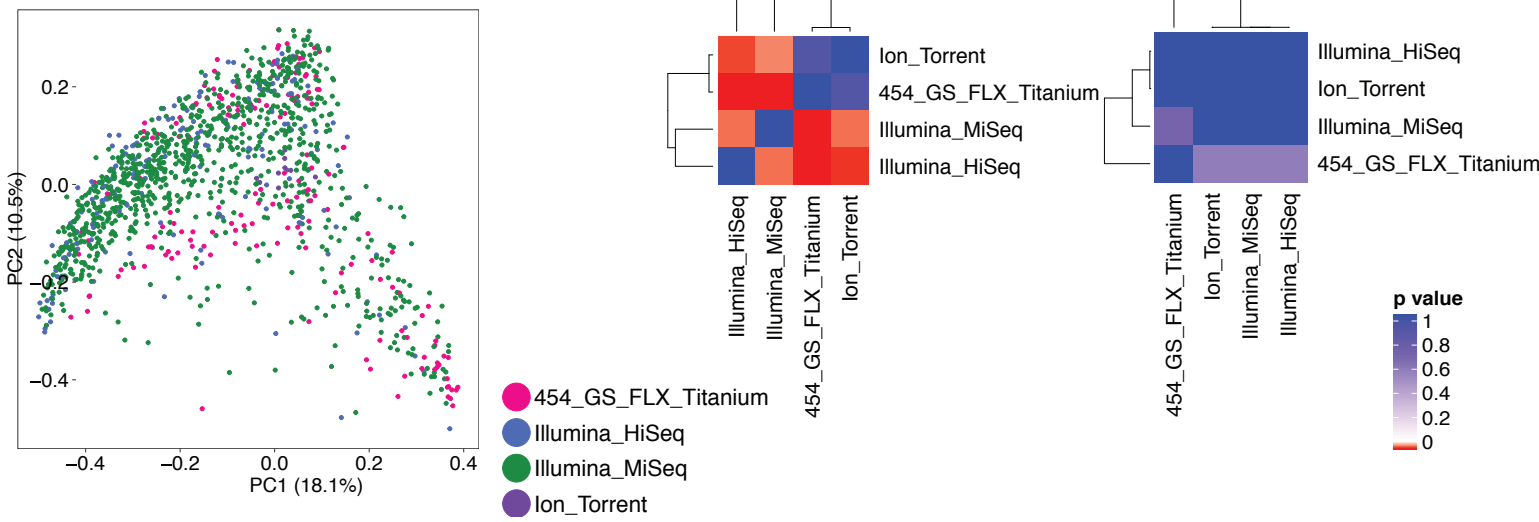
e



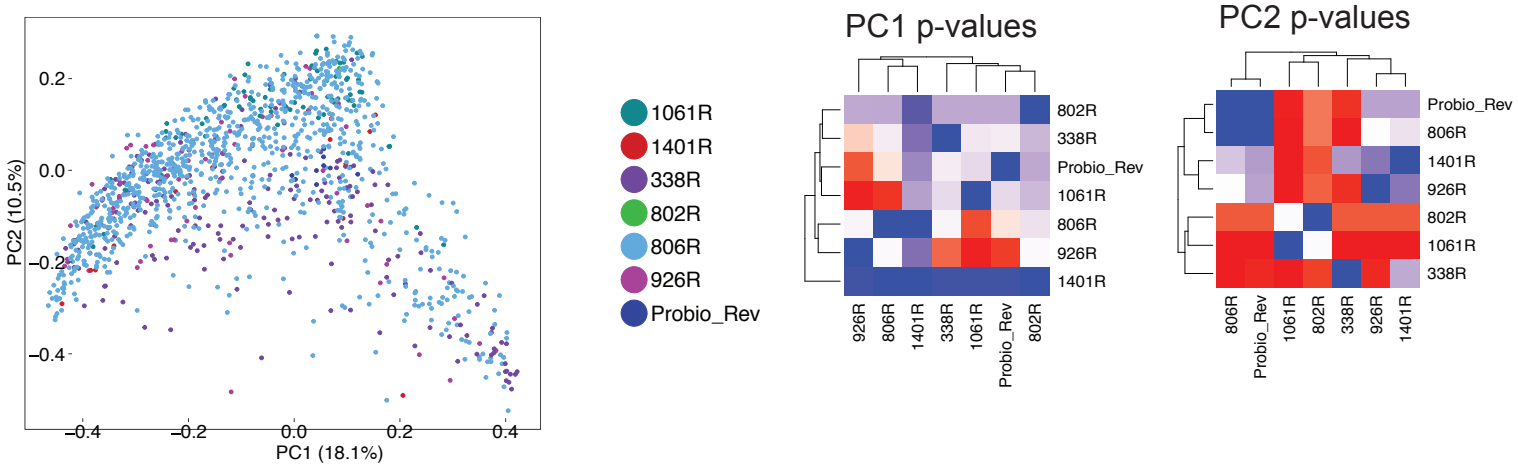
f

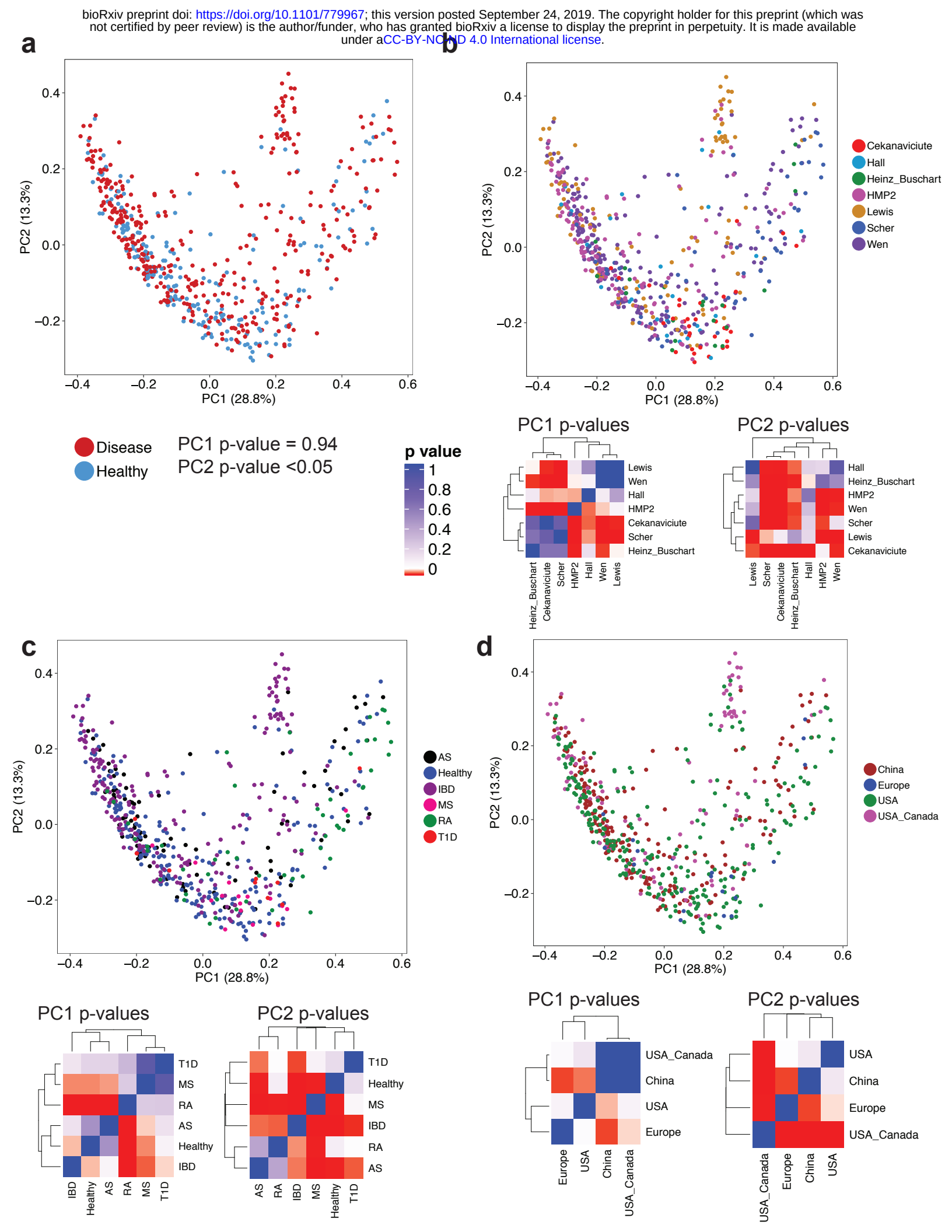


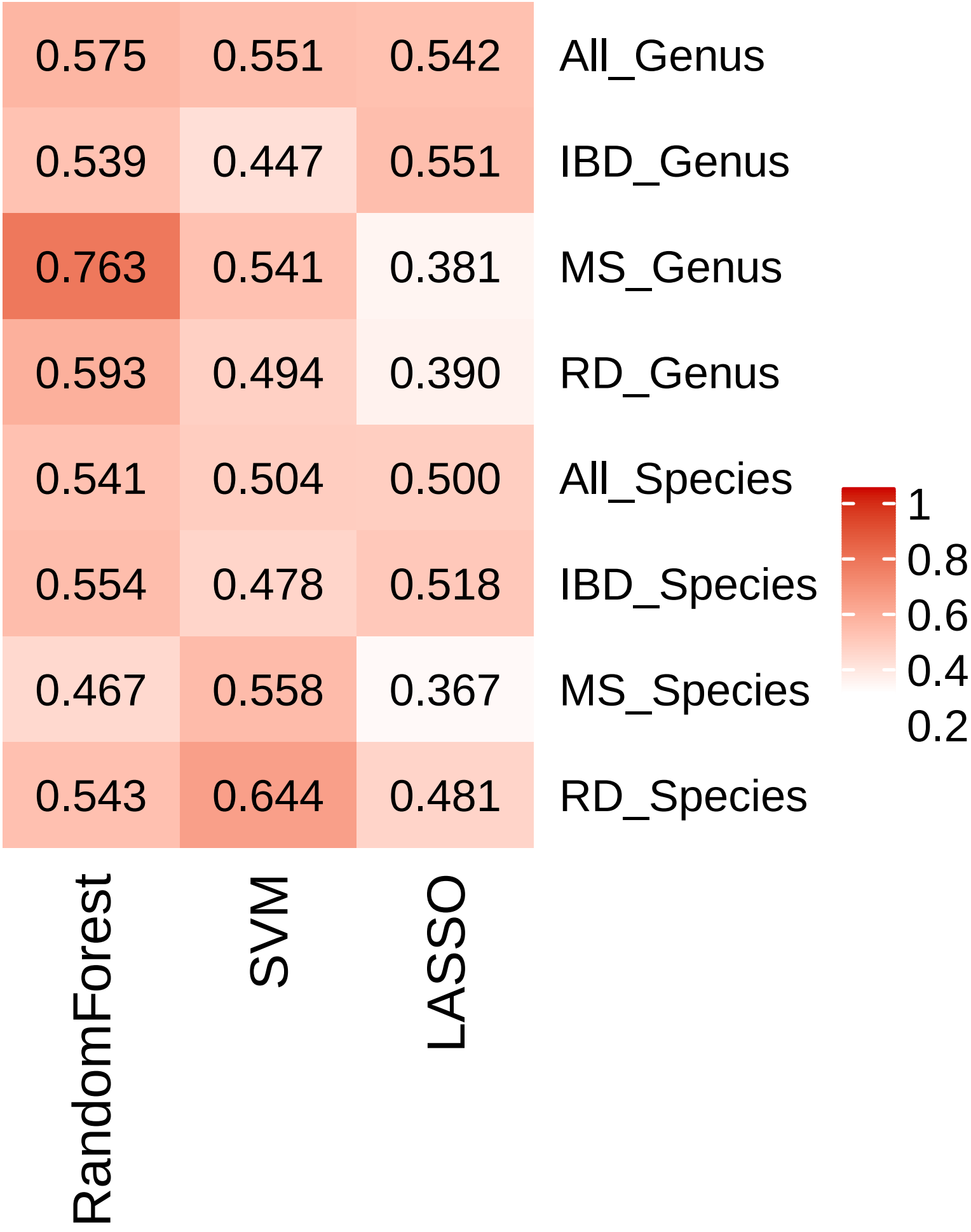
g

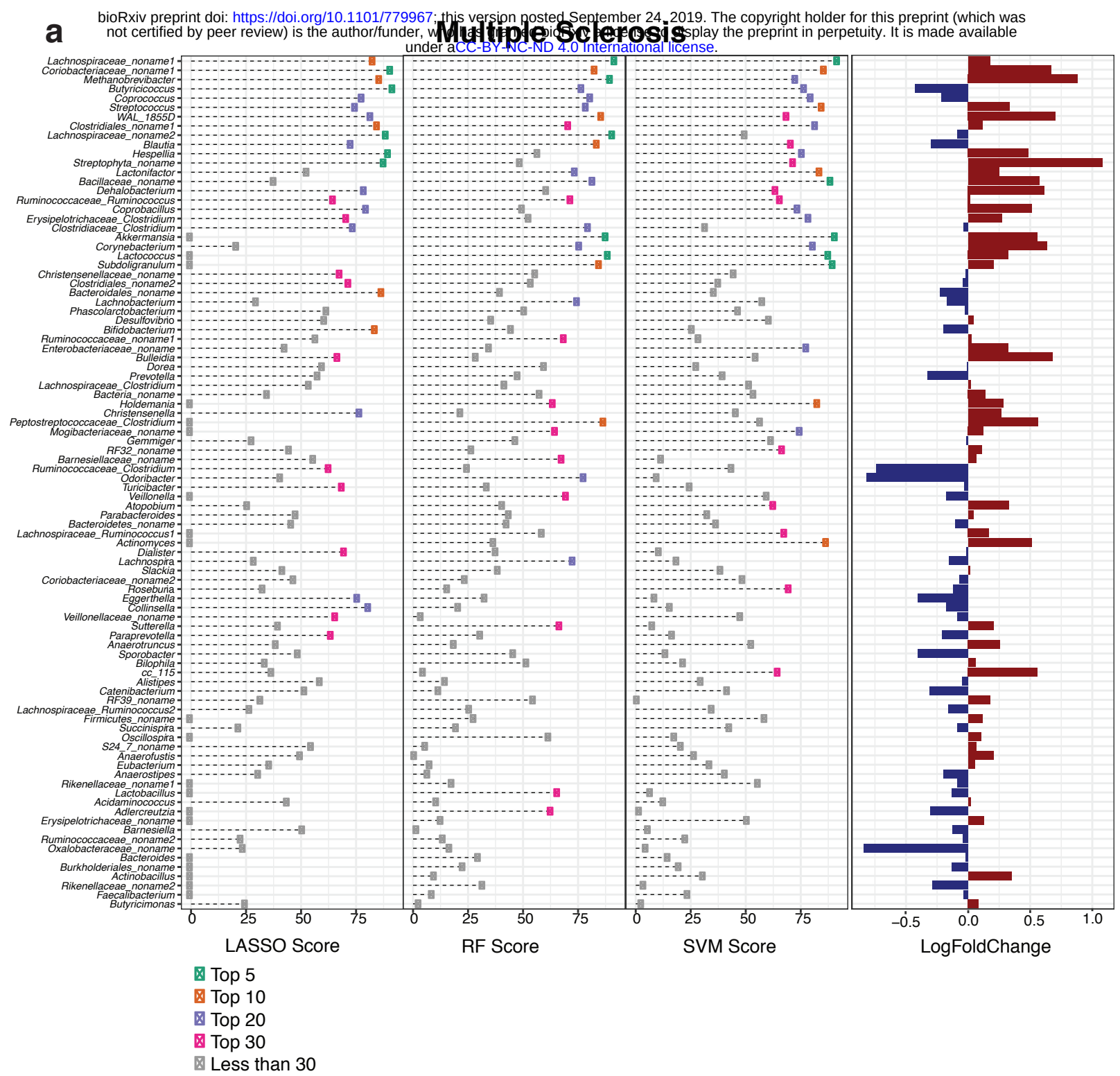


h



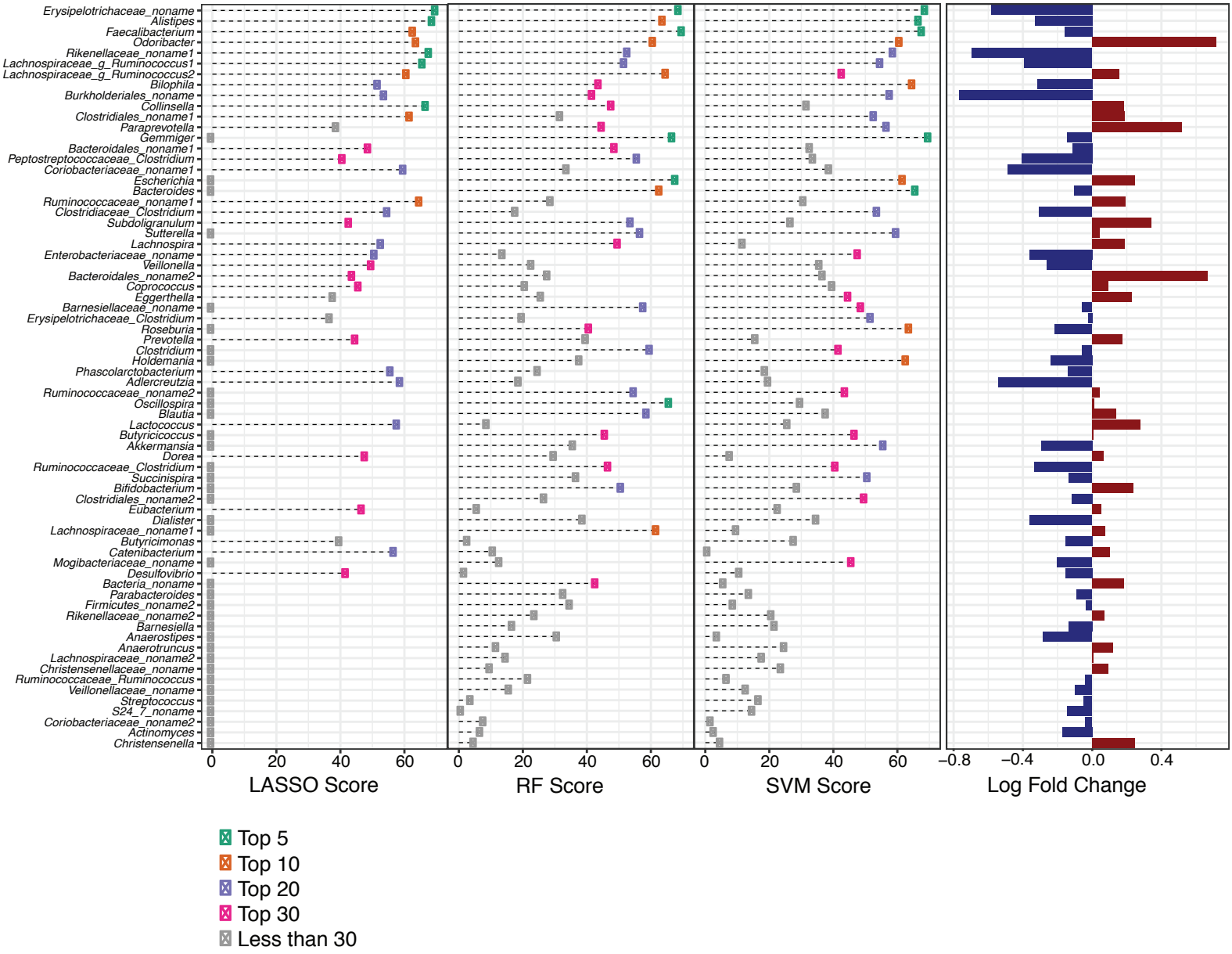



0.575
0.551
0.542
All_Genus
0.539
0.447
0.551
IBD_Genus
0.763
0.541
0.381
MS_Genus
0.593
0.494
0.390
RD_Genus
0.541
0.504
0.500
All_Species
0.554
0.478
0.518
IBD_Species
0.467
0.558
0.367
MS_Species
0.543
0.644
0.481
RD_Species
1
0.8
0.6
0.4
0.2
RandomForest
SVM
LASSO



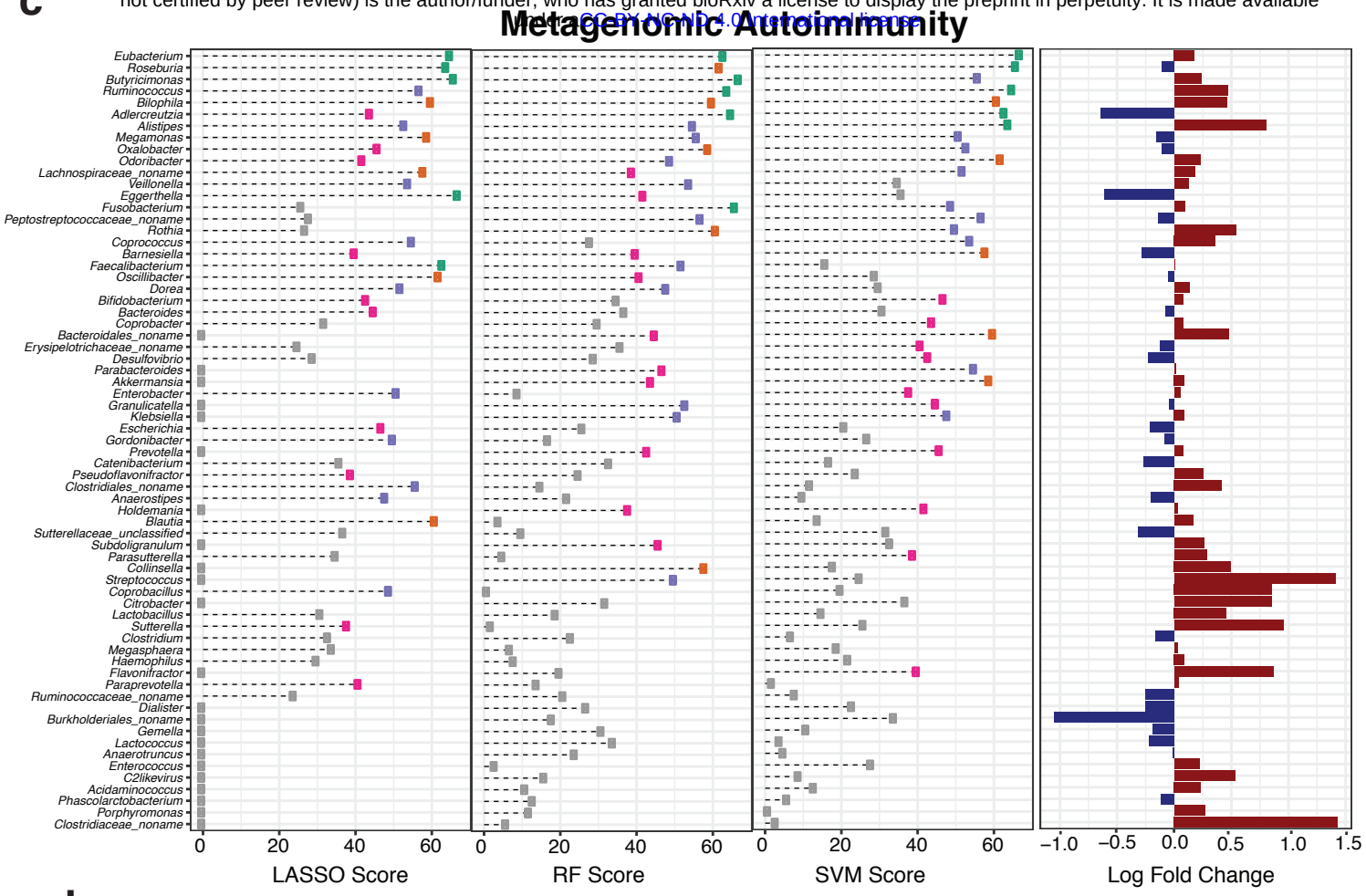
b

Rheumatic Diseases



- Top 5
- Top 10
- Top 20
- Top 30
- Less than 30

c



d