Effects of microbiome rare taxa filtering on statistical analysis

Quy Xuan Cao (quy.cao@pennmedicine.upenn.edu)
University of Montana Missoula College https://orcid.org/0000-0002-6204-1305
Xinxin Sun
Virginia Commonwealth University
Karun Rajesh
Virginia Commonwealth University
Naga Chalasani
Indiana University
Kayla Gelow
Indiana University
Barry Katz
Indiana University
Vijay H. Shah
Mayo Clinic Minnesota
Arun J. Sanyal
Virginia Commonwealth University
Ekaterina Smirnova
Virginia Commonwealth University

Research

Keywords: Filtering, Fast Permutation Test, Quality Control, Microbiome, Contaminants

Posted Date: July 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34781/v1

Version of Record: A version of this preprint was published at Frontiers in Microbiology on January 12th, 2021. See the published version at https://doi.org/10.3389/fmicb.2020.607325.

Effects of microbiome rare taxa filtering on statistical analysis

Cao, Quy*
quy.cao@pennmedicine.upenn.edu
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine, University of Pennsylvania

Sun, XinXin
sunx4@mymail.vcu.edu
Biostatistics Department, Virginia Commonwealth University

Rajesh, Karun
rajeshk@mymail.vcu.edu
Biostatistics Department, Virginia Commonwealth University

Chalasani, Naga
nchalasa@iupui.edu
Division of Gastroenterology
Department of Internal Medicine, Indiana University

Gelow, Kayla
peterkay@iu.edu
Department of Biostatistics, Indiana University

Katz, Barry
bkatz@iu.edu
Department of Biostatistics, Indiana University

Shah, Vijay H.
Shah.Vijay@mayo.edu
Division of Gastroenterology
Department. of Internal Medicine, Mayo Clinic

Sanyal, Arun J.
arun.sanyal@vcuhealth.org
Division of Gastroenterology, Hepatology and Nutrition
Department of Internal Medicine, Virginia Commonwealth University

Smirnova, Ekaterina
ekaterina.smirnova@vcuhealth.org
Biostatistics Department, Virginia Commonwealth University

June 25, 2020

*Corresponding author

RESEARCH

Effects of microbiome rare taxa filtering on statistical analysis

Quy Cao[1*], Xinxin Sun[2], Karun Rajesh[3], Naga Chalasani[4], Kayla Gelow[5], Barry Katz[5], Vijay H. Shah[6], Arun J. Sanyal[7] and Ekaterina Smirnova[2]

[*]Correspondence:
quy.cao@pennmedicine.upenn.edu
[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
Full list of author information is available at the end of the article

Abstract

Background: Accuracy of microbial community detection in 16S rRNA marker-gene and metagenomic studies suffers from contamination and sequencing errors that lead to either falsely identifying microbial taxa that were not in the sample or misclassifying the taxa of DNA fragment reads. Filtering is defined as removing taxa that are present in a small number of samples and have small counts in the samples where they are observed. This approach reduces extreme sparsity of microbiome data and has been shown to correctly remove contaminant taxa in cultured "mock" datasets, where the true taxa compositions are known. Although filtering is frequently used, careful evaluation of its effect on the data analysis and scientific conclusions remains unreported. Here, we assess the effect of filtering on the alpha and beta diversity estimation, as well as its impact on identifying taxa that discriminate between disease states.

Results:
The effect of filtering on microbiome data analysis is illustrated on four datasets: two mock quality control datasets where same cultured samples with known microbial composition are processed at different labs and two disease study datasets. Results show that in microbiome quality control datasets, filtering reduces the magnitude of differences in alpha diversity and alleviates technical variability between labs, while preserving between samples similarity (beta diversity). In the disease study datasets, DESeq2 and linear discriminant analysis Effect Size (LEfSe) methods were used to identify taxa that are differentially expressed across groups of samples, and random forest models to rank features with largest contribution towards disease classification. Results reveal that filtering retains significant taxa and preserves the model classification ability measured by the area under the receiver operating characteristic curve (AUC). The comparison between filtering and contaminant removal method shows that they have complementary effects and are advised to be used in conjunction.

Conclusions:
Filtering reduces the complexity of microbiome data, while preserving their integrity in downstream analysis. This leads to mitigation of the classification methods' sensitivity and reduction of technical variability, allowing researchers to generate more reproducible and comparable results in microbiome data analysis.

Keywords: Filtering; Fast Permutation Test; Quality Control; Microbiome; Contaminants

Background

Studies of microbiota association and human disease states have received increasing attention over the last decade [1]. It was shown that microbiota composition plays an important role in the development of multiple diseases including inflammatory bowel disease [2], diabetes [3, 4], preterm birth [5, 6], and liver diseases [7, 8]. Next generation sequencing (NGS) of the 16S rRNA marker is currently among the most widely used methods for microbial organisms identification. In these studies, samples collected at different body sites (e.g., vaginal swab, stool or blood) give counts of DNA fragments which are then grouped into similar microbial organisms, usually referred to as taxa. Hence, the resulting data is usually referred to as the "taxa table", or "derived feature data". In contrast to other -omics measurements, microbiome data are very sparse as many taxa are rare and often have zero counts in most samples.

The extreme levels of sparsity in microbiome datasets is one of the major challenges in data analysis. Indeed, it is not unusual to have over 90% of 0s in this data as it contains a large number of rare taxa observed in as few as 1 to 5% of samples. Recent microbiome quality control studies indicate that many rare taxa are caused by sequencing artifacts [9], contamination and/or sequencing errors [10, 11, 12, 13]. The most common approach to address this problem in the derived feature data is filtering, or removing spurious taxa from the 16S data set. Most filtering approaches are based on the rules of thumb, which vary from lab-to-lab. Recently, a filtering loss measure and a principled filtering test, namely PERFect [14], is introduced for deciding which taxa to remove. These methods are implemented in `Bioconductor` package `PERFect` [15], which includes a novel fast implementation of the permutation PERFect method. The implemented approach successfully reduces the original algorithm running time by almost four times.

While some techniques have been proposed to detect and remove contaminant and/or rare taxa, the literature in this research area is scarce. Davis et al [16] addressed this problem by introducing `decontam` `R` package that identifies contaminants by: (1) inversely correlating taxa frequencies with sample DNA concentration; and (2) using the prevalence of sequenced negative controls [17]. This method requires the auxiliary data from DNA quantitation, which is in most cases intrinsic to sample preparation, or negative controls data that is intrinsic to sequencing protocol. This approach is closely related to, but is not identical to filtering.

Traditional filtering methods were previously compared to the PERFect approach proposed by Smirnova et al [14] and tested on two datasets acquired from mock community experiments carried out at Virginia Commonwealth University (VCU) [12, 18] and a reagent and laboratory contamination dataset [17]. The authors used the number of contaminant taxa removed from the mock datasets as the method evaluation criteria. However, in practice, filtering is used as an intermediary step applied to the derived taxonomic feature table prior to data analysis. While filtering is a commonly used and recommended approach [19, 20], its benefits on data analysis and the effects on the scientific conclusions drawn from filtered and unfiltered data have not been reported.

The objectives of the current study are to evaluate: (1) the effects of filtering on technical variability for identical mock samples processed under different conditions;

(2) the advantages and disadvantages of using filtering for detecting significant taxa discriminating two groups of medical conditions. To address the first goal, we analyze the recent MicroBiome Quality Control (MBQC) project [13] that includes 1,016 oral mock samples sequenced at 15 laboratories and the results were then randomly distributed to 9 bioinformatics facilities for taxonomic classification; and the previously studied laboratory contamination dataset (denoted as Salter data) [17]. To address the second goal, we analyze two novel datasets on the gut microbiome studies on the TREAT consortium alcoholic hepatitis study [8] and the Human Microbiome Project inflammatory bowel disease [21]. To evaluate the effects of filtering, we concentrate on: (1) alpha (within) and beta (between) samples diversity analysis; (2) identification of significant taxa using random forest classification, LEfse and DESeq2 methods. Finally, we discuss the filtering and contaminant removal methodologies, and show that these approaches have complementary effects.

The rest of the paper is organized as follows. In the section "Motivating datasets" we introduce four datasets used in this study. In the section "Methods" we review the filtering and contaminant removal methodology details, as well as the statistical analysis methods used in the paper. We discuss analysis results in section "Results and discussion". Finally, in the section "Conclusions" we summarize major findings of the current study.

Motivating datasets

Mock data

The Microbiome Quality Control data

Consider the dataset from the MBQC project, a collaborative effort designed to comprehensively evaluate sample processing and computational methods for human microbiome data analysis [13]. There are four types of samples in this dataset: (1) 11 unique fresh stool samples; (2) seven unique freeze-dried stool samples; (3) two unique chemostat samples generated from a Robogut; and (4) two artificial colonies representing the gut and oral cavity. The aliquot of these samples were first randomly sequenced at 15 laboratories and the results were then randomly distributed to 9 bioinformatics facilities for taxonomic classification. Here, we consider the oral artificial communities data comprised of 22 true taxa. The MBQC project identified a total of 27,140 taxa across the four types of samples. For the purposes of this analysis, 14,861 taxa that have a 0 count across all oral artificial community samples are excluded; 1,277 taxa that matched names at the species level were combined; finally, 10,210 taxa that appeared in less than 5% of the samples were removed. The final dataset considered for this analysis contains 1,016 samples and 792 taxa. A limitation of this dataset is that the samples were created from the species in prescribed proportions; however, after the samples were processed many taxa were only identified up to the genus level (higher order phylogenetic hierarchy). As a consequence, only two signal taxa, *Veillonellaceae Veillonella Parvula* and *Coriobacteriaceae Eggerthella Lenta*, were correctly detected while the other 20 signal species are among the 184 taxa identified at the genus level.

Figure 1 (A) displays the log-counts heat map for the 100 most abundant taxa for the first five labs, arranged in decreasing order of abundance. Here, we rank taxa abundance by the number of samples a taxon is present in, where most abundant

taxon is ranked as 1, second abundant as 2, and so on. The white areas of the heatmap in the lower right corner indicate unobserved taxa, showing the decrease of signal strength with different processing institutes/labs. Figure 1 (B) displays the Principal Coordinate Analysis (PCoA) Bray-Curtis distance [22] multidimensional scaling (MDS) plots for $1,016$ samples from the heat map on the left. The first two principal components (PCs) that explain 32.2% of variability are shown on the plot. The samples cluster by bioinformatic labs, indicating differences across samples processed at different institutes even though they contain exactly the same signal species. These observations highlight the strong effects of the sequencing and bioinformatics protocol choice on taxa identification. Filtering, which removes rare taxa displayed in columns on the right-hand-side of the heatmap in Figure 1 (A) is one approach that could mitigate these differences. Left unresolved, this problem may cause a number of practical issues including: (1) falsely inflating within sample diversity, called α-diversity [23]; (2) obscuring true distances between samples, called β-diversity [24]; and (3) interpreting rare taxa as disease biomarkers (especially in low sample biomass environments).

The reagent and laboratory contamination data

The reagent and laboratory contamination study was designed to determine the effects of DNA extraction kits and other laboratory reagent contamination on sequencing output [17]. These data contain mock samples of a pure Salmonella bongori culture that had been processed at three different institutes: (1) Imperial College London (ICL); (2) University of Birmingham (UB); and (3) Wellcome Trust Sanger Institute (WTSI). Each mock sample underwent five rounds of serial ten-fold dilutions to generate a series of high (dilution = 0) to low (dilution = 5) biomass samples. Data visualization heatmap in Figure 6 (A) top panel displays the log-counts heat map for 635 observed taxa, generated using 40 Polymerase Chain Reaction (PCR) cycles. The taxa on the horizontal axis are arranged in decreasing order of abundance and the 18 samples on the vertical axis arranged by low to high (0 to 5) degrees of dilution. Results indicate that as the dilution number increases, true taxa contain less signal and are observed in lower counts, which makes it difficult to separate the signal from the noise.

Disease study data

In many microbiome studies, it is of interest to identify specific bacterial taxa that discriminate between two or more disease groups. We investigate the effects of filtering on identifying specific taxa that contribute to these differences using two recently reported microbiome studies.

Alcoholic hepatitis data

The study to characterize changes in fecal microbiome due to alcohol consumption and alcoholic hepatitis was performed by the sites involved in the TREAT consortium from 2014-2018 [8]. A total of 78 participants (healthy control (HC), n=24; heavy drinking control (HDC), n=20; moderate alcoholic hepatitis (MAH), n=10; severe alcoholic hepatitis (SAH), n=24) were studied. Results indicated that in random forest classification models, alcoholic hepatitis (moderate and severe alcoholic

hepatitis groups combined; n = 34) was associated with a distinct microbiome signature compared to heavy drinking controls (AUC=0.826), and multiple microbial genera were identified as the key contributors to these differences.

Inflammatory bowel disease data

The inflammatory bowel disease (IBD) data was generated as a part of the NIH Common Fund's Integrative Human Microbiome Project (iHMP/HMP2). The initial findings and multi-omic datasets from these studies were published in the Nature family of journals in May and June of 2019 [21], and data are publicly available through the HMP Data Coordination Center (HMP-DACC) and `HMP2Data` Bioconductor package [25] in `R`. The subset of 132 patients (control (nonIBD), n=46; Crohn's disease (CD), n= 86) with the open-source 16S data available through the `HMP2Data` package was selected for the analyses presented in this manuscript.

Methods

Filtering methods

Currently, there are only a few filtering methods used to alleviate the issue of contaminant and rare taxa. The majority of these methods are based on a heuristic non-statistical rule, with two methods where a threshold is derived statistically from the data. Here, we give a brief overview of these methods.

Rule of thumb approaches

In practice, filtering is a variation of an ad-hoc, albeit simple, procedure. One of the most widely used techniques for filtering in microbiome studies selects taxa that have a number of counts above $m = 0$ in at least k samples. This approach is borrowed from the RNA-seq gene expression literature and is implemented in the R package `genefilter` [26] and in QIIME bioinformatics pipeline function `filter_otus_from_otu_table.py` [27]. The choice of the threshold k comes from the count that is 0.1% of the minimal library size (the total number of count reads in the sample). For example, often the minimal library size is set to 5000 reads, thus popular filtering rules keep taxa present in at least $k = 5$ samples. Another popular approach is to remove taxa that are observed in fewer than k% of the samples. The advantage of these methods is that they are simple, intuitive, and easy to communicate with collaborators. However, they do not have an explicit loss function and objective criteria for choosing the tuning parameters m and k.

Statistical threshold selection

In contrast to the rule of thumb approaches where thresholds for filtering taxa are determined heuristically, the statistical approach selects an empirical filtering threshold based on the information given by the data. It extends traditional rule of thumb filtering approaches to find the best subset of retained taxa for further analysis by implementing statistical data-driven significance cut-off thresholds. The current method for such approach is `PERFect`, a principled filtering test that removes taxa with insignificant contribution to the total covariance [14]. Specifically, this method ranks taxa importance, measures their contribution to the total covariance, and quantifies the chance that the loss increases for a set of filtered taxa

is due to randomness using permutation tests. One drawback of the permutation filtering method is that it might be computationally expensive. Indeed, given that k permutations are performed for each taxon $j = 2, 3, \ldots, p$, the algorithm requires a total of $k(p-1)$ permutations, where k and p are large. Thus, the newer version of this package (see supplementary information), employs parallel processing and an unbalanced binary search algorithm [28] that optimally finds the cut-off taxon j to remove the set of taxa without building the permutation distribution and computing the p-values for all $p-1$ taxa.

Contaminant removal method

Contaminants in microbiome studies may arise from external sources such as the body of the study participant or sample collector [29, 30], sample collection instruments and laboratory reagents [17, 31, 32] or from internal sources (cross-contamination) when samples were mixed with each other during sample processing [31] or sequencing [33]. The recently developed contaminants removal method `Decontam` [16] identifies external contaminants by either: 1) inversely correlating taxa frequencies with sample DNA concentration; or 2) using the prevalence of sequenced negative controls. Our results suggest that `Decontam` removes abundant taxa that are likely contaminants but does not address the issue of rare taxa. A practical limitation of this method is that it requires auxiliary information from DNA quantitation or negative controls that is intrinsic to the sequencing protocol and might not always be available.

Statistical analysis methods

Within sample (alpha) [23] and between samples (beta) [24] diversity are used to evaluate the effects of filtering on reducing technical variability in mock datasets. Differences in estimated alpha diversity between processing labs were evaluated using Dunn's test with Benjamini-Hochberg controlling the false discovery rate [34] multiple comparisons adjustment implemented in R.

A number of methods for disease state prediction and differentially expressed taxa identification commonly used in metagenomic data analysis are considered in studying the effect of filtering. These methods are first performed on unfiltered and filtered data, then features importance for prediction and differentially expressed taxa selected by each method are compared. For predictive modeling, random forest [35], which is extensively applied in computational biology and genomics [36], is used to identify the set of most predictive taxa based on their Mean Decrease Gini measures. The classification model diagnostic ability in filtered and unfiltered models is compared using area under receiver operating characteristic curve (AUC). To identify differentially expressed taxa, DESeq2 [37] and linear discriminant analysis effect size (LEfSe) [38], are used. DESeq2 fits a negative binomial generalized linear model for each taxon count to obtain maximum likelihood estimates log-fold change between two classes, uses Bayesian shrinkage to shrink the log-fold change towards zero for taxa with low counts and/or high dispersion, and performs the Wald test on these shrunken log-fold changes for significance testing. LEfSe determines differentially expressed features by coupling non parametric standard tests for statistical significance with linear discriminant analysis (LDA), allowing researchers to further identify features that are consistent with biologically meaningful categories

[38]. However, there are some limitations for these methods when applied to microbiome data. Our results suggest that in many instances each of these three methods tends to flag a taxon as significant when the difference between two classes is driven by outlier counts. For example, a rare taxon that is absent for most samples and present in a few samples of one class can potentially be classified as differentially expressed taxon.

Results and discussion

PERFect simultaneous and permutation filtering approaches were previously validated [14] on three mock community data sets ([10], [11], and [12]) using the number of contaminant taxa correctly removed as an efficiency criterion. Here, we concentrate on the effects of filtering on downstream analyses, using the two major exploratory analyses used in microbiome research, alpha and beta diversity, as well as its impact on identifying taxa that discriminate between disease states.

The MicroBiome Quality Control Data Analysis

One of the main goals of the MBQC project was to understand major differences in technology and methods for analyzing human microbiome data. This was achieved by analyzing the observed taxa variation between: (1) the labs that sequenced samples according to the internal protocol; and (2) bioinformatics pipelines used to perform taxomonic classification. Here, we concentrate on the effect of bioinformatics processing laboratories on the observed oral mock community data measured by alpha and beta diversity, two of the most commonly used summaries in microbiome research. Figure 2 (A) shows the Shannon index, the most widely-used diversity metric which accounts for both abundance and evenness of the species present [39] in the unfiltered and filtered data. The Shannon index, H is defined as $H = -\sum_{i=1}^{S} p_i \ln(p_i)$, where p_i is the proportion of total sample represented by species i and S is the total number of species in the community. This plot and the summary statistics in Table 1 indicate a decrease in the Shannon index between the unfiltered and filtered data, implying a reduction in the diversity and evenness of taxa. Specifically, the diversity decreases from 792 taxa to 175 and 222 taxa using the simultaneous and permutation method respectively, while retaining 22 true taxa. Since both filtering methods remove more than 70% of taxa, the distribution of the remaining taxa shifts in favor of the true taxa by increasing their proportions in the samples, resulting in less even communities. As the result, this reduction of alpha diversity of samples tends towards the true alpha diversity as indicated by the red dashed line.

In order to study the effect of filtering on differences across bioinformatics processing labs, we applied Dunn's test with a Benjamini-Hochberg correction for multiple testing to all possible pairwise Shannon alpha diversity comparisons between processing labs. Results are summarized in Table 2. Since all samples contained the same mock communities, in the absence of technical variability, none of the differences should be significant. For the unfiltered data, 21 out of 28 possible pairs have significant differences in alpha diversity at the 0.05 significance level. Applying simultaneous and permutation filtering decreases differences in alpha diversity for most pairs. Moreover, *there are a total of* 4 *and* 8 *pairwise comparisons that are no*

longer significant at the 0.05 *level after simultaneous and permutation filtering was applied respectively.* While filtering does not remove all differences due to processing labs, these results indicate that it dramatically alleviates differences in alpha diversity estimates caused by the lab-to-lab variability.

To assess the effect of filtering on beta diversity, we calculated the pairwise Bray-Curtis distances between samples using a combined taxa matrix which consists of the unfiltered taxa matrix, and the taxa filtered matrices of PERFect simultaneous and PERFect permutation each at the p-value threshold of 0.1. The multidimensional scaling ordination plot for the first two principal components which explain 30.5% of the variability in the data is shown in Figure 2 (B). Three filtering methods (unfiltered, simultaneous and permutation PERFect) are arranged in columns and samples are colored according to 8 processing institutes. Figure 2 (B) shows that while data clusters by laboratory in each dataset, the proximity between clusters decreases when simultaneous or permutation filtering is applied. This observation indicates that filtering decreases dissimilarity between samples that contain the same mock communities and slightly alleviates the effects of lab-to-lab variability. Thus, filtering achieves dimension reduction (reduces the number of taxa) while preserving beta diversity.

The reagent and laboratory contamination data

Figure 2 (C) displays the difference in the Shannon index of the filtered outputs, corresponding to the p-value threshold 0.1, using simultaneous and permutation filtering among 6 dilution levels and 3 processing institutes. It is expected that as the dilution levels increase, more uncertainty in true taxa identification is introduced into the biological system, thus the proportion of signal taxa decreases whereas that of noise taxa increases. This phenomenon is displayed on the top heatmap in Figure 6 (A), where taxa are arranged from left to right in decreasing abundance order (noise taxa to the bottom right of the heatmap). As the dilution levels increase (rows of the heatmap), each dilution 'band' becomes denser due to the increase in noise taxa counts. Therefore, dilution causes the true signal to become more even with noise, and thus leading to a higher Shannon index. This effect may cause problems comparing alpha diversity for different groups of samples with variable biomass because it will be more difficult to differentiate between signal and noise taxa in low biomass samples. The filtering methods address this issue by removing noise taxa in highly diluted samples (dilutions $3, 4$ and 5), where the simultaneous filtering removes more taxa than the permutation algorithm and has more impact on reducing the alpha diversity.

To compare the beta diversity for filtered outputs, the pairwise between-sample Bray-Curtis distances were calculated using the taxa matrices' combination with a similar set up to the analysis with the MBQC data. The multidimensional scaling ordination plot for the first two principal components that explain 81.3% of the variability in the data is shown in Figure 2 (D). The six dilution levels and three filtering methods (none, simultaneous and permutation PERFect) are arranged in columns and rows respectively; samples are colored according to the three processing institutes. Ideally, the samples from all three processing institutes should have the same composition of taxa regardless of the dilution levels. However, contaminants

that went into the samples during the DNA extraction and PCR process lead to the differences between the three processing institutes. Figure 2 (D) shows that filtering does not dramatically change samples' pairwise distances in ordination plots. This is due to the fact that PERFect, like many other filtering methods, removes taxa with low abundance which do not contribute to the signal, and thus do not dramatically affect samples' pairwise distances. These observations lead to the important conclusion that filtering reduces the number of taxa considered in the analysis, and thus reduces dimensionality of the taxa table, without affecting beta diversity.

Alcoholic Hepatitis Data Analysis

Random forest results

The ROC curves of the random forest models between unfiltered (AUC = 0.826) and filtered (AUC = 0.816) data are shown in Figure 3 (A). The predictive abilities of the filtered and unfiltered models as measured by AUC values are similar, although there is a small decrease of 0.01 in the AUC for the model built on the filtered data. This implies that removing rare taxa has little effect on the classification ability of the random forest model. Further, the most predictive taxa, as measured by the mean Gini decrease criteria, also tend to be abundant (see Supplementary Table 1). Specifically, the ranks of the top 35 predictive taxa in the unfiltered data range between the first and 81st most abundant taxa out of the total 345 taxa in the unfiltered data set.

To compare the discrepancy of the taxa importance rank between these two random forest models, we use the elbow method on taxa mean Gini decrease in unfiltered data to choose the top 60 most predictive taxa. The taxa are chosen so that the differences of consecutive taxa mean Gini decrease are no less than 0.001. Then, their importance ranks are compared to those from the filtered data. Results indicate that in general, while there is minor variation, ranks are consistent and strong predictive taxa keep their classification ability after filtering (see Supplementary Table 1).

LEfSe results

The LDA score for all significant taxa using LEfSe from unfiltered and filtered data are shown in Figure 4 (A). For each taxon, the log fold change values from unfiltered and filtered data are similar, although the values from unfiltered data tend to be slightly higher (range between 0.01 and 0.75). *This indicates that filtering retains the differential expression for almost all taxa*, thus taxa that are significant in unfiltered data tend to be significant in filtered data.

There are four taxa that are present in the unfiltered but absent in the filtered data results, where two are identified at the family level (*Lactobacillaceae* and *Coriobacteriaceae*) and two are identified at the genus level (*Ruminococcaceae Butyricicoccus* and *Clostridiaceae Proteiniclasticum*). At the family level, filtering removes rare taxa from each family (3 out of 6 taxa from *Lactobacillaceae* and 6 out of 12 taxa from *Coriobacteriaceae*). The remaining taxa aggregated to each the two families do not discriminate between the heavy drinking control (HDC, n = 20) and the Alcohol Hepatitis (AH, n = 34) group. At the genus level, *Ruminococcaceae Butyricicoccus* and *Clostridiaceae Proteiniclasticum* are flagged as significant in the

unfiltered but as non-significant in the filtered data. Both taxa are overall more rare (42nd and 179th most abundant taxa out of 345 taxa), with relative abundance between 0 and 0.02 (max without outlier) (see Supplementary Figure 1), one outlier for *Ruminococcaceae Butyricicoccus* (relative abundance = 0.08), and only a few low relative abundance observations in the HDC group for *Ruminococcaceae Butyricicoccus*. This suggests that in the presence of taxa with outliers, the difference between groups appears to be stronger when tested in the unfiltered dataset with a large number of rare taxa. However, the strength of the outliers' effect is reduced when testing is performed in the filtered data, where extremely rare taxa are removed.

DESeq2 results

The DESeq2 method based on log(Count + 1) transformed data was used to identify differentially expressed taxa in filtered and unfiltered taxa tables; results are shown in Figure 4 (B). Taxa colored in black were identified as significant in both filtered and unfiltered datasets, while the taxon in blue (*Lachnospiraceae Anaerostipes*) was present in filtered data but was not significant. This discrepancy was not surprisingly due to small differences in significance level of this taxon for filtered and unfiltered data. At the alpha level of 0.1, this taxon was significant for both unfiltered and filtered data with raw p-value of 0.028 and 0.035, respectively. After the p-value adjustment step using Benjamini-Hochberg procedure, it remained significant in unfiltered data (p-value = 0.093) but became non-significant in filtered data (p-value = 0.113). Since the change between raw p-values is relatively small and the adjusted p-values are close to the alpha level, we may conclude that for DESeq2 method, there are no major differences due to filtering in this dataset.

Summary of discrimination results

Common significant taxa between random forest models, LEfSe and DESeq2 results on unfiltered data are shown in Figure 4 (C). There are two taxa that are significant in the unfiltered but not significant in the filtered: *Lachnospiraceae Anaerostipes* (not significant if DESeq2) and *Fusobacteriaceae Fusobacterium* (low importance in random forest). Results indicate that these discrepancies occur for the borderline significant taxa.

IBD Data Analysis

Filtering analysis on the IBD data is performed using the workflow of Alcoholic Hepatitis data and the results are shown in Figures 5. We observe similar results' patterns with the Alcoholic Hepatitis analyses: 1) significant taxa tend to be highly abundant; 2) predictive abilities of the filtered and unfiltered random forest models as measured by AUC values are similar (unfiltered AUC = 0.852; filtered AUC = 0.853); 3) in the discriminant analysis, the differences between LDA scores for all significant taxa from unfiltered and filtered data are small; 4) the significance effect of more rare taxa with outliers is stronger in the unfiltered compared to filtered dataset.

Compared to Alcoholic Hepatitis data, DESeq2 identified three additional significant taxa in the filtered data: *Lachnospiraceae Eubacterium* (p-value = 0.080),

Fusobacteriaceae Fusobacterium (p-value = 0.0999) and *Enterobacteriaceae EscherichiaShigella* (p-value = 0.070). Since these adjusted p-values are borderline significant in the filtered data that includes less taxa, these genera are non significant when DESeq2 is run on the unfiltered data due to a larger number of taxa used in multiple comparison adjustment.

Comparison with contaminant removal method

Contaminant removal and filtering methods have a common goal of identifying potential features derived due to technical limitations that occur with sequencing and taxonomic classification. However, that main goal of each method is different, which led to complementary effects in our comparison studies. Filtering concentrates on removing rare taxa relying mostly on sparsity assumptions and using no auxiliary information about the derived feature data. Contaminant removal methods implemented in `R` package `decontam` use additional information from the sequencing process to apply statistical threshold rules marginally to one taxon at a time. The `decontam` package implements two methods, each using a specific auxiliary information about derived feature data: (1) frequency method uses DNA quantitation data recording the concentration of DNA in each sample; and (2) prevalence method employs a set of "negative control" samples in which sequencing was performed on blanks without any biological sample added.

Following the methods discussed by Davis et al, 2018 [16], we applied `decontam` frequency method to the reagent and laboratory contamination data to compare filtering and contaminant identification methods in terms of the type of taxa they remove and their effects on diversity. Results are illustrated in Figure 6, which compares the heatmaps, alpha and beta diversity for the derived feature data without filtering to the data where taxa are removed using `decontam` frequency and `PERFect` simultaneous filtering methods. Heatmaps in Figure 6 (A) indicate that `decontam` frequency identifies abundant taxa as contaminants (left oval in the middle panel heatmap) leaving rare taxa to the right of the plot in the data set. In contrast, `PERFect` removes rare taxa (right oval in the bottom panel heatmap) that `decontam` was not able to detect, while leaving abundant taxa in the data set. These observations highlight important methodological differences between two methods. Specifically, `decontam` frequency fits a regression model to compare a contaminant model, in which expected frequency varies inversely with total DNA concentration, and a non-contaminant model, in which expected frequency is independent of total DNA concentration [16]. For rare taxa regression model fit is unstable due to the small number of observations (a few samples where rare taxa appear), and thus `decontam` returns missing values for the taxa significance. Specifically, out of a total of 635 taxa, `decontam` identified 61 taxa as contaminants, and was not able to evaluate statistical significance for 221 taxa. Filtering approach has a major limitation of being skewed toward retaining more dominant features, as a result, a persistent contaminant feature might appear in a large number of samples, have a high contribution towards covariance and would not be removed from the data set.

Comparison of alpha diversity in Figure 6 (B) reveals that both `decontam` and `PERFect` reduce Shannon diversity. Based on this data, `decontam` leads to greater alpha diversity reduction, which is expected since it removes more abundant taxa.

However, it should be noted that this is a small sample size study with only 18 observations (six observations per each of the three institutes) and results may not be conclusive. Figure 6 (C) compares beta diversity Bray-Curtis distance plots for each method (rows) by dilution level (columns) with samples colored by processing institutes. All samples contain the same biological material, thus under no technical variability scenario, the points should overlap on the plot. This is the case for undiluted samples (first column dilution = 0), however the observed dissimilarity between samples increases with dilution. Davis et al, 2018 [16] showed that removing abundant contaminants (second row in Figure 6 (C)) reduces technical beta diversity. Comparing these results with filtering output confirms our previous observations based on the Microbiome Quality Control data set that removing rare taxa via filtering does not significantly effect beta diversity.

Conclusions

It is generally believed that filtering rare taxa is an effective quality control approach to remove possible contaminants, sequencing and taxomomic assignment artifacts. The current study supports this paradigm and demonstrates that filtering has a strong potential to reduce lab-to-lab variability between samples that contain similar microbial species and processed according to different protocols. Moreover, filtering removes rare taxa that have low contribution to the signal, thus reducing dimensionality of the data with minimal information loss. The ability of the methods to detect taxa significantly different across two disease groups is almost unaffected by filtering. Except for a small number of taxa detected as significant in unfiltered but not filtered (or visa versa) data, each method produces the same results. Major discrepancies in taxa that are identified as significant come from the data analysis method choice (Random Forest, LEfSe or DESeq2) but not from filtering. To the best of our knowledge, *this is the first report on the effects of filtering on statistical analyses of microbiome data.*

The statistical methodology literature on quality control for the derived feature data is scarce. Most previous studies either recommended filtering without thorough evaluation of its effects [19, 20], or focused on the number of taxa removed from the mock artificial community studies [14] and on contaminant identification [16, 10]. We have previously demonstrated that filtering methods were effective in identifying true species in mock data [14]. An underlying assumption of filtering is that most rare taxa are not informative in the analysis; however presence of rare taxa in the derived feature data increases sparsity and affects performance of statistical methods. The current study supports earlier hypotheses and validates that removing rare taxa does not impact the scientific conclusions.

It has also been established that the contaminant removal method implemented in `decontam` package [16] was effective in reducing technical variability across processing institutes. Comparison of filtering and contaminant removal methods on Salter data [17] reveals that the two methods have complementary effects: `decontam` removes persistent contaminant features that appear in a large number of samples while filtering removes rare taxa that appear in a small number of samples. This is not surprising because this is exactly the assumptions of these two methods, nevertheless this is a significant finding which suggests that in practice both methods may be used to remove sequencing artifacts from the derived feature data.

Another noteworthy finding is that most significant taxa in unfiltered data were abundant. The random forest variable importance ranks of the top 35 predictive taxa in the unfiltered data ranged between: (1) the 1st and 81st most abundant taxa out of the total 345 taxa in the alcoholic hepatitis; and (2) the 1st and 93rd most abundant taxa out of the total 409 taxa in the inflammatory bowel disease data set. Furthermore, in LEfSe and DESeq2 discrimination models, taxa that were found significant in filtered but not unfiltered data (or similarly in unfiltered but not filtered data) were overall more rare (present in small number of samples) and with low relative abundance. This is an important observation that may guide researchers' decision regarding how aggressive filtering should be.

A limitation of filtering is that the reduction of type I errors (probability of removing important taxa) will inevitably increase type II errors (probability of keeping unimportant taxa). Indeed, if we want to be cautious in removing rare taxa to ensure that important taxa will still remain in the data, we will not remove many taxa and will likely have a lot of unimportant taxa remained; if we remove taxa aggressively, there is a high chance of filtering important rare taxa. In particular, in studies that aim to discover rare taxa, filtering would not be advisable since it will likely remove the rare but important taxa. This issue can be moderated by having a good understanding of the data (where the data are sampled and how they are generated) and using auxiliary study information that allows us to filter with confidence. In particular for predictive modeling, for example using random forest approach in predicting alcoholic hepatitis, building a model with more abundant taxa may lead to higher reproducibility across studies as rare taxa may not be observed in another cohort sampled at different conditions.

We would like to stress that the goal of the current study is evaluation of filtering methods on commonly used microbiome analyses. As a part of this study, filtering was compared to a closely related contaminant removal method implemented in `R` package `decontam` using one of the datasets that was previously illustrated by the package developers [16]. It would be of interest to perform thorough comparison of these methods on other datasets used in this study, however this is outside of the scope of this paper.

In summary, the current study provides information on the effects of removing rare taxa on technical variability and scientific conclusions drawn from statistical analyses. We provide insights into the role of filtering in microbiome studies, and highlight the importance of derived feature data quality control prior to scientific analysis.

Abbreviations

AUC: area under the receiver operating characteristic curve; CD: Crohn's disease; HC: healthy control; HDC: heavy drinking control; HMP: Human Microbiome Project; IBD: imflammatory bowel disease; ICL: Imperial College London; LEfSe: linear discriminant analysis effect size; MAH: moderate alcoholic hepatitis; MBQC: microbiome quality control; MDS: multidimensional scaling; NGS: next generation sequencing; PC: principle component; PCoA: principal coordinate analysis; PCR: polymerase chain reaction; rRNA: ribosomal RNA; SAH: severe alcoholic hepatitis; UB: University of Birmingham; WTSI: Wellcome Trust Sanger Institute.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. R version 3.6.1 was used with Bioconductor packages phyloseq version 1.30.0, HMP2Data version 1.0.0, DESeq2 version 1.26.0, PERFect version 1.0.0. Analyses using LEfSe are done on the website https://huttenhower.sph.harvard.edu/galaxy/.

Author's contributions

QC and ES designed the study; NC, KG, BK, VS, AS collected and processed alcoholic hepatitis data; QC, XS, KR, AS and ES analyzed the data; QC and ES drafted the manuscript; QC, XS, KR, NC, KG, BK, VS, AS, and ES reviewed the data and the manuscript.

Acknowledgements

This work was funded by the National Institute of Alcohol Abuse and Alcoholism (UO1 AA021891-01 and T32 DK07150-40 to A.J.S.); National Center for Advancing Translational Sciences (CTSA UL1TR002649 to E.S.); and Virginia Commonwealth University for Clinical and Translational Research (National Institutes of Health Clinical and Translational Award).

Author details

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA. [2]Biostatistics Department, Virginia Commonwealth University, Richmond, USA. [3]Bioinformatics Department, Virginia Commonwealth University, Richmond, USA. [4]Div. of Gastroenterology, Dept. of Internal Medicine, Indiana University, Indianapolis, USA. [5]Dept. of Biostatistics, Indiana University, Indianapolis, USA. [6]Div. of Gastroenterology, Dept. of Internal Medicine, Mayo Clinic, Rochester, USA. [7]Div. of Gastroenterology, Hepatology and Nutrition, Dept. of Internal Medicine, Virginia Commonwealth University, Richmond, USA.

References

1. Nguyen, L.D.N., Viscogliosi, E., Delhaes, L.: The lung mycobiome: an emerging field of the human respiratory microbiome. Frontiers in Microbiology **6**, 89 (2015). doi:10.3389/fmicb.2015.00089
2. Huttenhower, C., Kostic, A., Xavier, R.: Inflammatory bowel disease as a model for translating the microbiome. Immunity **40**(6), 843–854 (2014)
3. Proctor, L.M.: The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell host and microbe **16**(3), 276–289 (2014)
4. Pascale, A., Marchesi, N., Govoni, S., Coppola, A., G aruso, C.: The role of gut microbiota in obesity, diabetes mellitus, and effect of metformin: new insights into old diseases. Current Opinion in Pharmacology **49**, 1–5 (2019)
5. Callahan, B.J., DiGiulio, D.B., Goltsman, D.S.A., Sun, C.L., Costello, E.K., Jeganathan, P., Biggio, J.R., Wong, R.J., Druzin, M.L., Shaw, G.M., Stevenson, D.K., Holmes, S.P., Relman, D.A.: Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. Proceedings of the National Academy of Sciences **114**(37), 9966–9971 (2017). doi:10.1073/pnas.1705899114. https://www.pnas.org/content/114/37/9966.full.pdf
6. DiGiulio, D.B., Callahan, B.J., McMurdie, P.J., Costello, E.K., Lyell, D.J., Robaczewska, A., Sun, C.L., Goltsman, D.S.A., Wong, R.J., Shaw, G., Stevenson, D.K., Holmes, S.P., Relman, D.A.: Temporal and spatial variation of the human microbiota during pregnancy. Proceedings of the National Academy of Sciences **112**(35), 11060–11065 (2015). doi:10.1073/pnas.1502875112. https://www.pnas.org/content/112/35/11060.full.pdf
7. Puri, P., Liangpunsakul, S., Christensen, J.E., Shah, V.H., Kamath, P.S., Gores, G.J., Walker, S., Comerford, M., Katz, B., Borst, A., Yu, Q., Kumar, D.P., Mirshahi, F., Radaeva, S., Chalasani, N.P., Crabb, D.W., Sanyal, A.J.: The circulating microbiome signature and inferred functional metagenomics in alcoholic hepatitis. Hepatology **67**(4), 1284–1302 (2018)
8. Smirnova, E., Puri, P., Muthiah, M.D., Daitya, K., Brown, R., Chalasani, N., Liangpunsakul, S., Shah, V.H., Gelow, K., Siddiqui, M.S., et al.: Fecal microbiome distinguishes alcohol consumption from alcoholic hepatitis but does not discriminate disease severity. Hepatology (2020)
9. Lahr, D.J., Katz, L.A.: Reducing the impact of pcr-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity dna polymerase **47**(4), 857–866 (2009)
10. Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., Kelley, S.T.: Bayesian community-wide culture-independent microbial source tracking. Nature methods **8**(9), 761–763 (2011)
11. Ravel, J., Gajer, P., Abdo, Z., Schneider, G., Koenig, S.K., McCulle, S., Karlebach, S., Gorle, R., Russell, J., Tacket, C., Brotman, R., Davis, C., Ault, K., Peralta, L., Forney, L.: Vaginal microbiome of reproductive-age women. Proceedings of the National Academy of Sciences **108**(Supplement 1), 4680–4687 (2011)
12. Fettweis, J., Serrano, M., Sheth, N., Mayer, C., Glascock, A., Brooks, J., Jefferson, K., Vaginal Microbiome Consortium, a., Buck, G.: Species-level classification of the vaginal microbiome. BMC Genomics **13**(Supplement 18), 1–9 (2012)
13. Sinha, R., Abnet, C.C., White, O., Knight, R., Huttenhower, C.: The microbiome quality control project: baseline study design and future directions. Genome Biology **16**(1), 276 (2015). doi:10.1186/s13059-015-0841-8
14. Smirnova, E., Huzurbazar, S., Jafari, F.: Perfect: Permutation filtering test for microbiome data. Biostatistics (2018). doi:10.1093/biostatistics/kxy020
15. Smirnova, E., Cao, Q.: PERFect: Permutation Filtration for Microbiome Data. (2019). R package version 1.2.0. `https://github.com/cxquy91/PERFect`
16. Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A., Callahan, B.J.: Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome **6**(1), 226 (2018). doi:10.1186/s40168-018-0605-2
17. Salter, S., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N., Walker, A.W., et al.: Reagent contamination can critically impact sequence-based microbiome analyses. BMC Biology (2014). doi:10.1101/007187
18. Brooks, J.P., Edwards, D.J., Harwich, M.D., Rivera, M.C., Fettweis, J.M., Serrano, M.G., Reris, R.A., Sheth, N.U., Huang, B., Girerd, P., Vaginal Microbiome Consortium (additional members), Strauss, J.F., Jefferson,

K.K., Buck, G.A.: The truth about metagenomics: quantifying and counteracting bias in 16s rrna studies. BMC Microbiology **15**(1), 66 (2015). doi:10.1186/s12866-015-0351-6
19. Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., Koren, O., Walters, W.A., Caporaso, J.G., Knight, R., Ley, R.E.: Conducting a microbiome study. Cell **158**(2), 250–262 (2014)
20. Cullen, C.M., Aneja, K.K., Beyhan, S., Cho, C.E., Woloszynek, S., Convertino, M., McCoy, S.J., Zhang, Y., Anderson, M.Z., Alvarez-Ponce, D., *et al.*: Emerging priorities for microbiome research. Frontiers in Microbiology **11**, 136 (2020)
21. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., *et al.*: Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature **569**(7758), 655–662 (2019)
22. Quaak, F.C.a., Kuiper, I.: Statistical data analysis of bacterial t-rflp profiles in forensic soil comparisons. Forensic Science International **210**(1-3), 96–101 (2011). doi:10.1016/j.forsciint.2011.02.005
23. Park, C., Allaby, M.: alpha diversity. Oxford University Press (2017). `http://www.oxfordreference.com/view/10.1093/acref/9780191826320.001.0001/acref-9780191826320-e-278`
24. Park, C., Allaby, M.: beta diversity. Oxford University Press (2017). `http://www.oxfordreference.com/view/10.1093/acref/9780191826320.001.0001/acref-9780191826320-e-706`
25. John Stansfield, N.Z.J.F.L.W.M.D. Ekaterina Smirnova: HMP2Data: 16s rRNA Sequencing Data from the Human Microbiome Project 2. (2019). R package version 1.3.0. `https://github.com/jstansfield0/HMP2Data`
26. Gentleman R, ., V., C., W., H., F., H.: Genefilter: Methods for Filtering Genes from High-throughput Experiments. (2019). R package version 1.68.0
27. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al.: Qiime allows analysis of high-throughput community sequencing data. Nature Methods **7**(5), 335–336 (2010). doi:10.1038/nmeth.f.303
28. Morin, P.: Open Data Structures: An Introduction. Athabasca University Press, ??? (2013)
29. Kitchin, P.A., Szotyori, Z., Fromholc, C., Almond, N.: Avoidance of false positives. Nature **344**(6263), 201–201 (1990). doi:10.1038/344201a0
30. Meadow, J.F., Altrichter, A.E., Bateman, A.C., Stenson, J., Brown, G., Green, J.L., Bohannan, B.J.m.: Humans differ in their personal microbial cloud. PeerJ **3** (2015). doi:10.7717/peerj.1258
31. Jousselin, E., Clamens, A.-L., Galan, M., Bernard, M., Maman, S., Gschloessl, B., Duport, G., Meseguer, A.S., Calevro, F., Dacier, A.C., et al.: Assessment of a 16s rrna amplicon illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus. Molecular Ecology Resources **16**(3), 628–640 (2015). doi:10.1111/1755-0998.12478
32. Glassing, A., Dowd, S.E., Galandiuk, S., Davis, B., Chiodini, R.J.: Inherent bacterial dna contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathogens **8**(1) (2016). doi:10.1186/s13099-016-0103-7
33. Larsson, A.J.M., Stanley, G., Sinha, R., Weissman, I.L., Sandberg, R.: Computational correction of index switching in multiplexed sequencing libraries. Nature Methods **15**(5), 305–307 (2018). doi:10.1038/nmeth.4666
34. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (1995). doi:10.1111/j.2517-6161.1995.tb02031.x
35. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001). doi:10.1023/a:1010933404324
36. Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M.J., Aliferis, C.F., Alekseyenko, A.V., et al.: A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome **1**(1) (2013). doi:10.1186/2049-2618-1-11
37. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology **15**(12) (2014). doi:10.1186/s13059-014-0550-8
38. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C.: Metagenomic biomarker discovery and explanation. Genome Biology **12**(6) (2011). doi:10.1186/gb-2011-12-6-r60
39. Reese, A.T., Dunn, R.R.: Drivers of microbiome biodiversity: A review of general rules, feces, and ignorance. mBio **9**(4) (2018). doi:10.1128/mBio.01294-18. https://mbio.asm.org/content/9/4/e01294-18.full.pdf

Figures

Figure 1 Heat map and multidimensional scaling plot of MBQC data. Heat map and multidimensional scaling plot of MBQC data. (A) The heat map of 100 observed taxa on the log-scale, with taxa on the x-axis arranged in decreasing abundance order and samples on the y-axis arranged by processing institutes. (B) The multidimensional scaling plot of 1016 samples, colored by the processing institutes. Data source: [13].

Tables

Additional Files

Additional file 1 — Figures for the main paper
Figures for this paper are stored in a tex file named 'Figures for Filtering paper'. It will be stored as a zip file with associated figures.

Additional file 2 — Supplementary Materials
The supplementary material is a tex file that includes the analysis of the reagent and laboratory contamination data, additional figures and tables for the alcoholic hepatitis and IBD data.

Figure 2 Diversity comparison Diversity comparison on MBQC and Salter data. (A) Shannon alpha diversity index for the original data and two filtered data, colored by the bioinformatics labs. The horizontal dashed line represents the true Shannon index. (B) Beta diversity multidimensional scaling plots of the unfiltered, simultaneous and permutation PERFect filtered data colored by bioinformatics processing institutes. Data source: [13]. (C) Shannon index for the original data and two filtered data, colored by the dilution levels. (D) Multidimensional scaling plots of the unfiltered and filtered data at different dilution levels, colored by the processing institutes. Data source: [17].

Figure 3 ROC curves of random forest models. ROC curves of the random forest models from unfiltered and filtered data that are differentiated by colors. (A) ROC curves from the Alcoholic Hepatitis data [8]. (B) ROC curves from the IBD data [21].

Figure 4 Alcoholic Hepatitis results Alcoholic Hepatitis analysis results for Random Forest, LEfSe and DESeq2. (A) Log fold changes for all significant taxa from LEfSe results from unfiltered and filtered data that are differentiated by colors. Taxa that are present in filtered data but are not significant are colored in dark blue. (B) Barchart of log(count+1) for significant taxa from DESeq2 results, colored by the disease states. Taxa that are present in filtered data but are not significant are colored in dark blue. (C) Barchart of log(count+1) for common significant taxa between random forest models, LEfSe and DESeq2 results on unfiltered data, colored by the disease states. From filtered data, while black taxa are common results with those from unfiltered data, blue taxa are non-significant in DESeq2 results and green taxa are not in the top 60 predictive taxa in the random forest model. Data source: [8].

Figure 5 IBD results IBD analysis results for Random Forest, LEfSe and DESeq2. (A) Log fold changes for all significant taxa from LEfSe results from unfiltered and filtered data that are differentiated by colors. Taxa that are present in filtered data but are not significant are colored in dark blue. Taxa that are present in unfiltered data but are not significant are colored in dark red. (B) Barchart of log(count+1) for significant taxa from DESeq2 results, colored by the disease states. Taxa that are present in unfiltered data but are not significant are colored in dark red. Data source: [21].

Figure 6 Decontam results Comparison of the original data (no filtering), contaminant removal (decontam frequency) and filtering (PERFect simultaneous) methods. (A) Heat map of log transformed taxa counts in decreasing abundance order on the x-axis and samples by dilution level on the y-axis for the original data (top panel), data where contaminants are removed using decontam (middle panel) and rare taxa filtered using PERFect (bottom panel). True taxa are colored in green to the left of each heatmap; ovals indicate taxa removed by decontam and PERFect methods. (B) Alpha diversity for the three comparisons colored by dilution level and processing institute. (C) Beta diversity PCoA Bray-Curtis distances plots colored by processing institute and arranged by dilution level (rows) and three taxa removal methods (columns). Data source: [17].

		BL-1	BL-2	BL-3	BL-4	BL-6	BL-8	BL-9A	BL-9B
Median	Unfiltered	5.301	5.293	5.700	5.622	5.324	5.270	5.317	5.536
	Simultaneous	4.122	4.147	4.061	3.998	4.381	3.247	4.332	4.061
	Permutation	4.287	4.301	4.261	4.300	4.552	3.519	4.459	4.269
IQR	Unfiltered	0.137	0.175	0.165	0.446	0.181	0.062	0.161	0.477
	Simultaneous	0.135	0.231	0.123	0.420	0.205	0.456	0.280	0.083
	Permutation	0.166	0.235	0.127	0.475	0.224	0.489	0.288	0.098

Table 1 Summary statistics of the Shannon index for each processing lab. Data source: [13]

	Unfiltered		Simultaneous		Permutation	
Comparison	Difference	P-values	Difference	P-values	Difference	P-values
BL-1 - BL-2	0.00	0.4990	0.20	0.4214	-0.06	0.4778
BL-1 - BL-3	-10.75	< 0.0001	2.52	0.0074	1.27	0.1307
BL-2 - BL-3	-10.83	< 0.0001	2.35	0.0115	1.33	0.1283
BL-1 - BL-4	-7.22	< 0.0001	3.90	0.0001	0.29	0.4173
BL-2 - BL-4	-7.28	< 0.0001	3.73	0.0001	0.35	0.4088
BL-3 - BL-4	3.91	0.0001	1.25	0.1243	-1.01	0.1816
BL-1 - BL-6	-1.63	0.0632	-6.35	< 0.0001	-7.10	< 0.0001
BL-2 - BL-6	-1.64	0.0646	-6.60	< 0.0001	-7.10	< 0.0001
BL-3 - BL-6	9.36	< 0.0001	-8.78	< 0.0001	-8.23	< 0.0001
BL-4 - BL-6	5.70	< 0.0001	-10.47	< 0.0001	-7.55	< 0.0001
BL-1 - BL-8	2.47	0.0090	11.30	< 0.0001	9.99	< 0.0001
BL-2 - BL-8	2.49	0.0089	11.19	¡ 0.0001	10.13	< 0.0001
BL-3 - BL-8	13.27	< 0.0001	8.51	< 0.0001	8.50	< 0.0001
BL-4 - BL-8	9.81	< 0.0001	7.60	< 0.0001	9.92	< 0.0001
BL-6 - BL-8	4.16	< 0.0001	17.93	< 0.0001	17.36	< 0.0001
BL-1 - BL-9A	-1.09	0.1535	-5.46	< 0.0001	-5.74	< 0.0001
BL-2 - BL-9A	-1.10	0.1583	-5.70	< 0.0001	-5.73	< 0.0001
BL-3 - BL-9A	9.66	< 0.0001	-7.86	< 0.0001	-6.88	< 0.0001
BL-4 - BL-9A	6.09	< 0.0001	-9.46	< 0.0001	-6.14	< 0.0001
BL-6 - BL-9A	0.51	0.3167	0.77	0.2455	1.24	0.1311
BL-8 - BL-9A	-3.57	0.0003	-16.78	< 0.0001	-15.76	< 0.0001
BL-1 - BL-9B	-6.44	< 0.0001	3.32	0.0006	1.58	0.0840
BL-2 - BL-9B	-6.49	< 0.0001	3.14	0.0011	1.65	0.0770
BL-3 - BL-9B	4.55	< 0.0001	0.70	0.2609	0.27	0.4090
BL-4 - BL-9B	0.71	0.2556	-0.56	0.2996	1.33	0.1233
BL-6 - BL-9B	-4.92	< 0.0001	9.79	< 0.0001	8.79	< 0.0001
BL-8 - BL-9B	-9.00	< 0.0001	-8.06	< 0.0001	-8.49	< 0.0001
BL-9A - BL-9B	-5.32	< 0.0001	8.82	< 0.0001	7.37	< 0.0001

Table 2 Pairwise comparisons of the Shannon index between laboratories using Dunn's test for each dataset. Data source: [13].

Figures

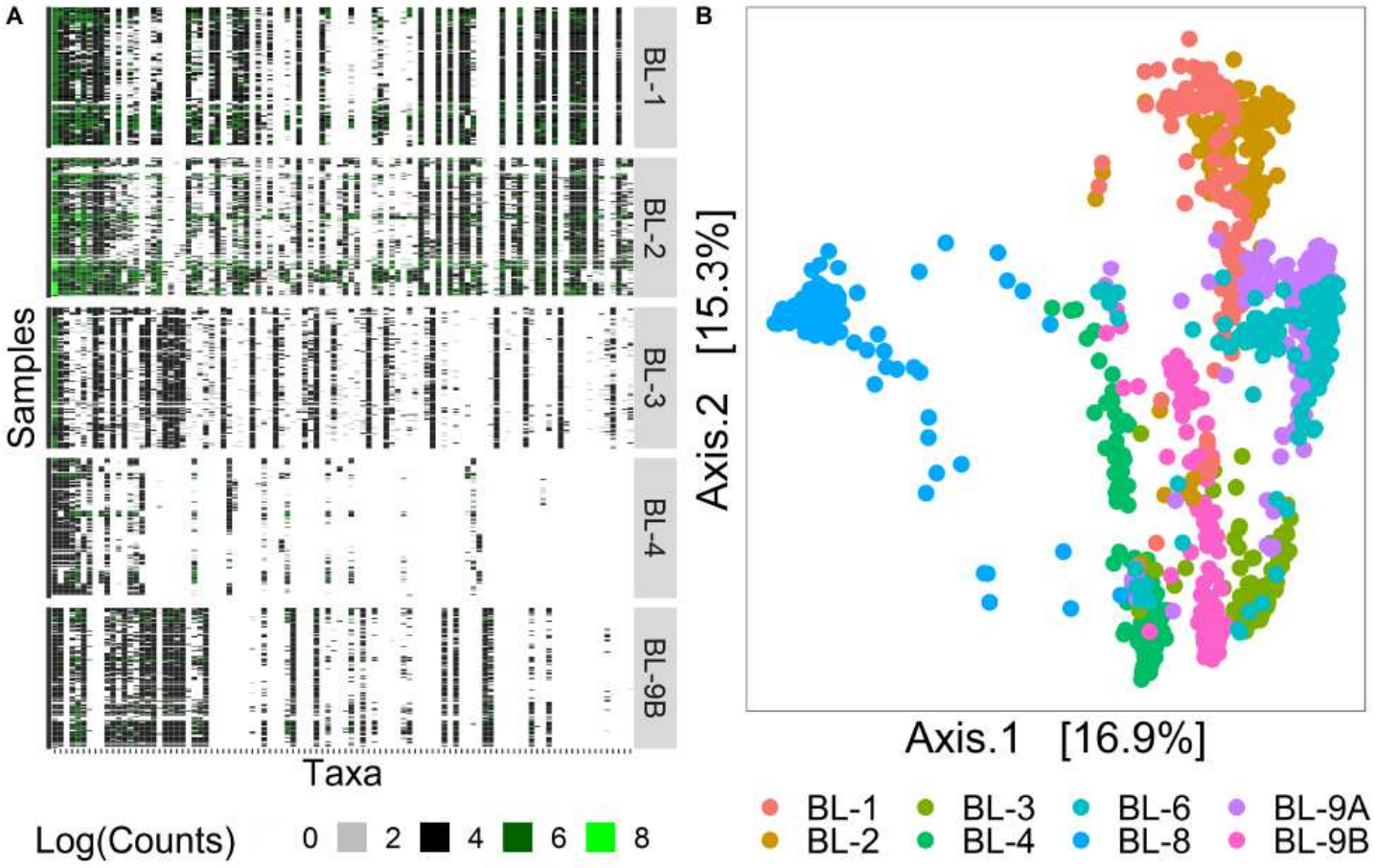

Figure 1

Heat map and multidimensional scaling plot of MBQC data. Heat map and multidimensional scaling plot of MBQC data. (A) The heat map of 100 observed taxa on the log-scale, with taxa on the x-axis arranged in decreasing abundance order and samples on the y-axis arranged by processing institutes. (B) The multidimensional scaling plot of 1016 samples, colored by the processing institutes. Data source: [13].

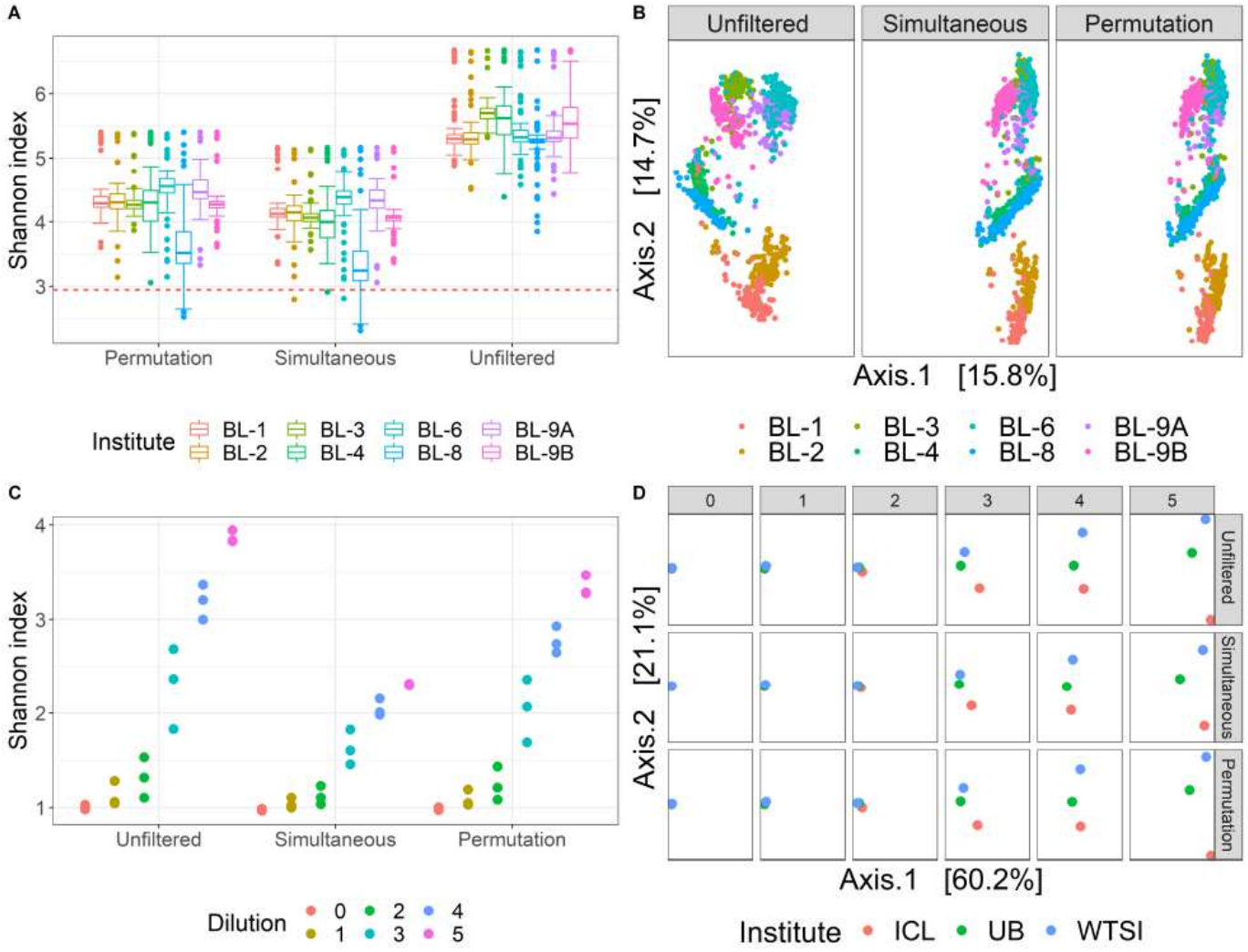

Figure 2

Diversity comparison on MBQC and Salter data. (A) Shannon alpha diversity index for the original data and two filtered data, colored by the bioinformatics labs. The horizontal red dashed line represents the true Shannon index. (B) Beta diversity multidimensional scaling plots of the unfiltered, simultaneous and permutation PERFect filtered data colored by bioinformatics processing institutes. Data source: [1]. (C) Shannon index for the original data and two filtered data, colored by the dilution levels. (D) Multidimensional scaling plots of the unfiltered and filtered data at different dilution levels, colored by the processing institutes. Data source: [2].

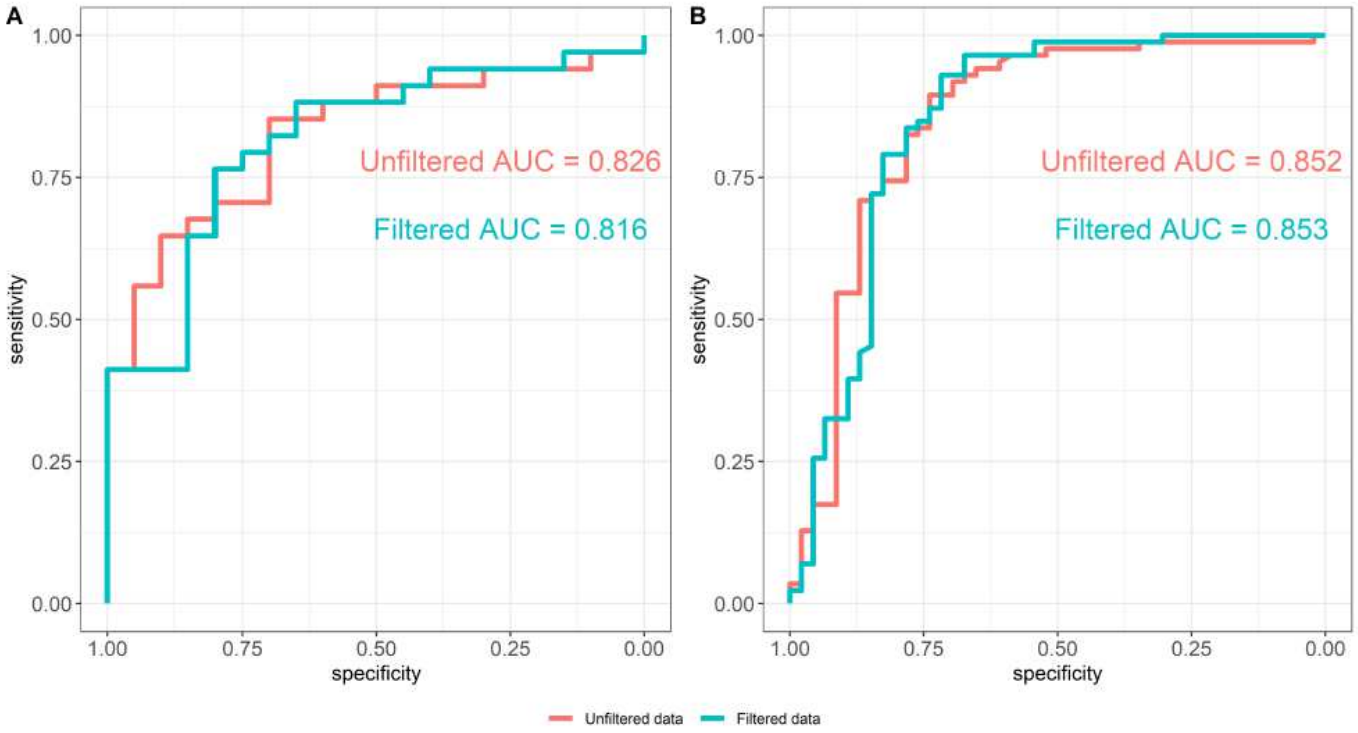

Figure 3

ROC curves of the random forest models from unfiltered and filtered data that are differentiated by colors. (A) ROC curves from the Alcoholic Hepatitis data [3]. (B) ROC curves from the IBD data [4].

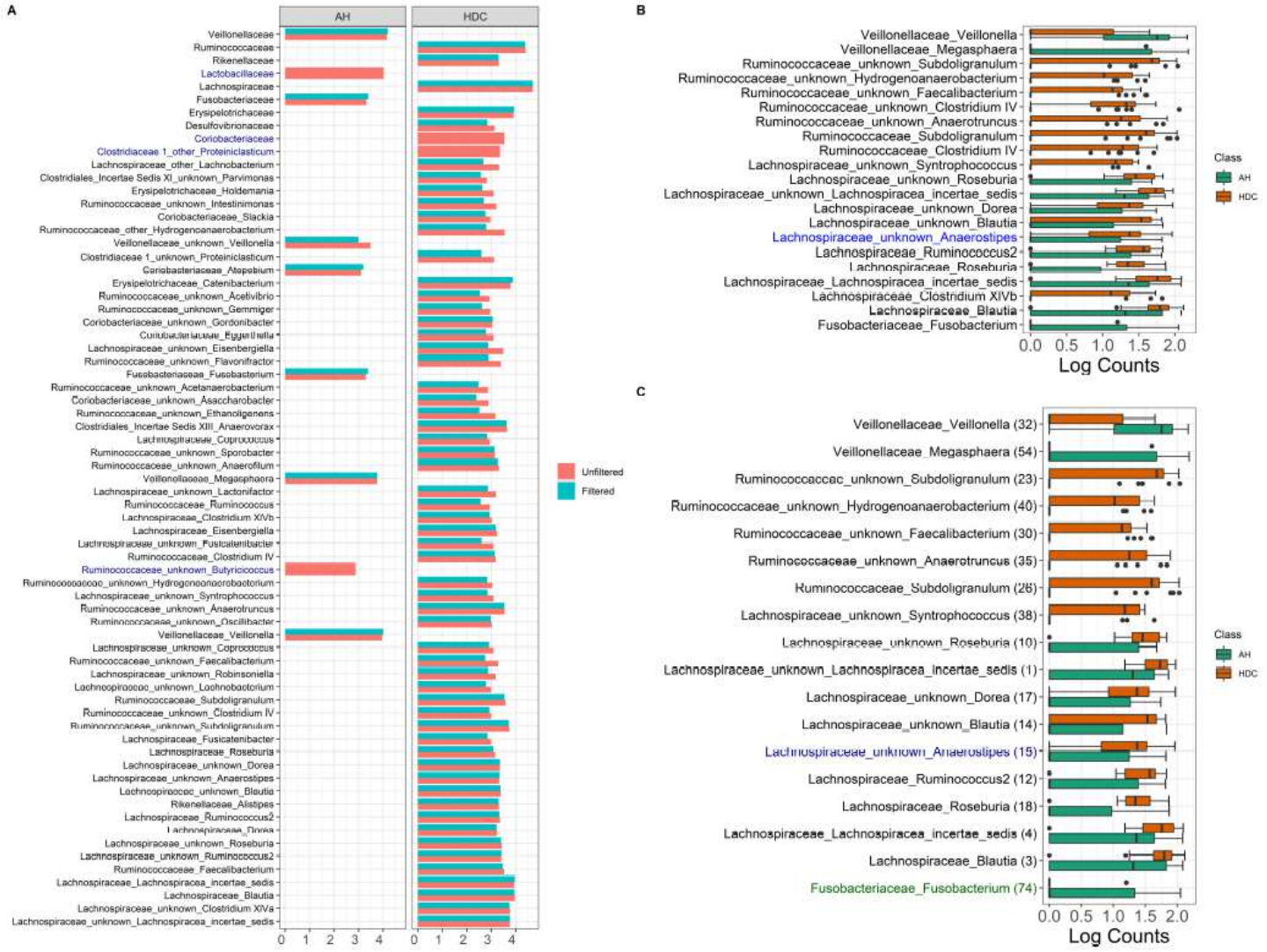

Figure 4

Alcoholic Hepatitis analysis results for Random Forest, LEfSe and DESeq2. (A) Log fold changes for all significant taxa from LEfSe results from unfiltered and filtered data that are differentiated by colors. Taxa that are present in filtered data but are not significant are colored in dark blue. (B) Barchart of log(count+1) for significant taxa from DESeq2 results, colored by the disease states. Taxa that are present in filtered data but are not significant are colored in dark blue. (C) Barchart of log(count+1) for common significant taxa between random forest models, LEfSe and DESeq2 results on unfiltered data, colored by the disease states. From filtered data, while black taxa are common results with those from unfiltered data, blue taxa are non-significant in DESeq2 results and green taxa are not in the top 60 predictive taxa in the random forest model. Data source: [3].

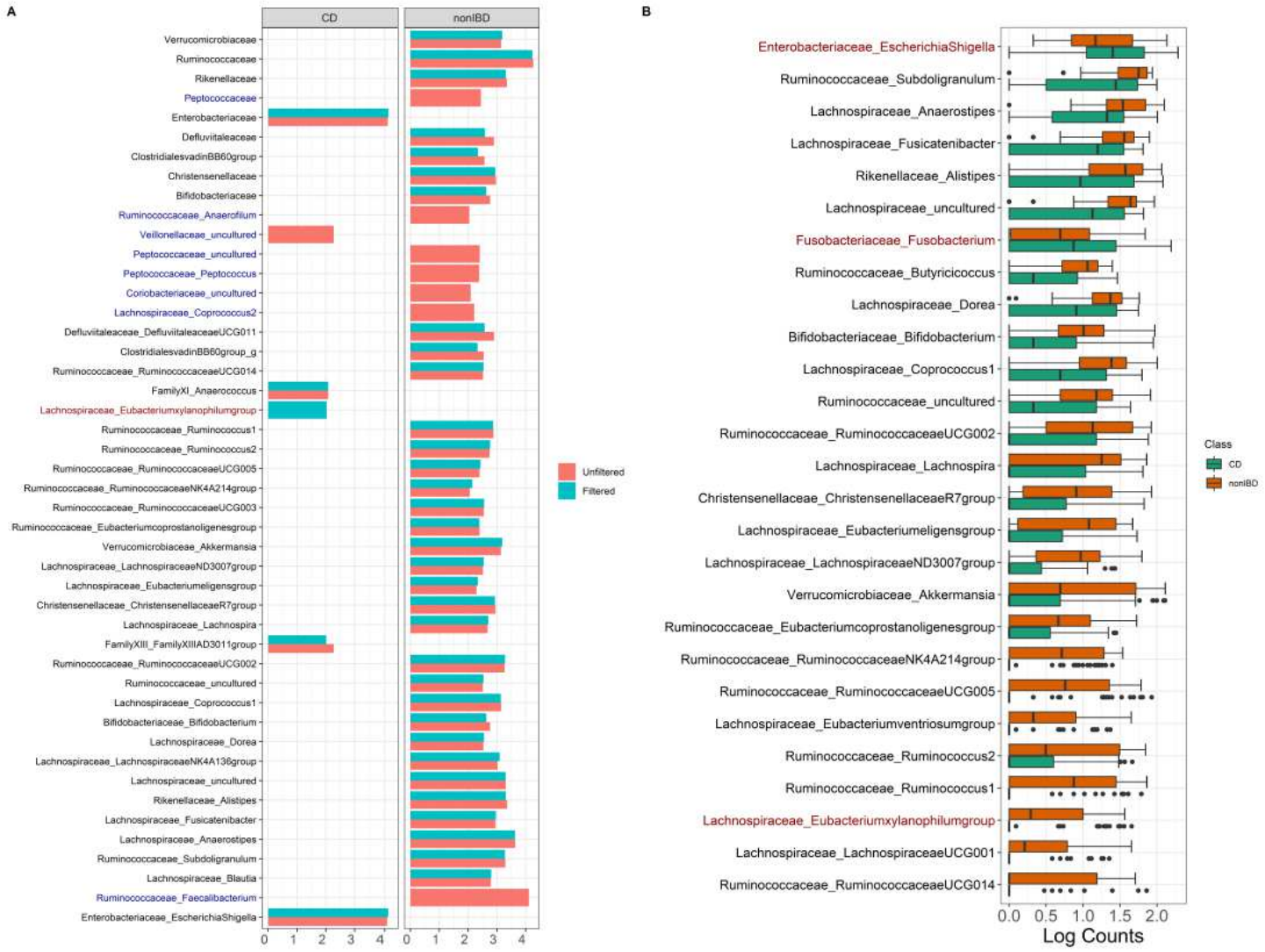

Figure 5

IBD analysis results for Random Forest, LEfSe and DESeq2. (A) Log fold changes for all significant taxa from LEfSe results from unfiltered and filtered data that are differentiated by colors. Taxa that are present in filtered data but are not significant are colored in dark blue. Taxa that are present in unfiltered data but are not significant are colored in dark red. (B) Barchart of log(count+1) for significant taxa from DESeq2 results, colored by the disease states. Taxa that are present in unfiltered data but are not significant are colored in dark red. Data source: [4].

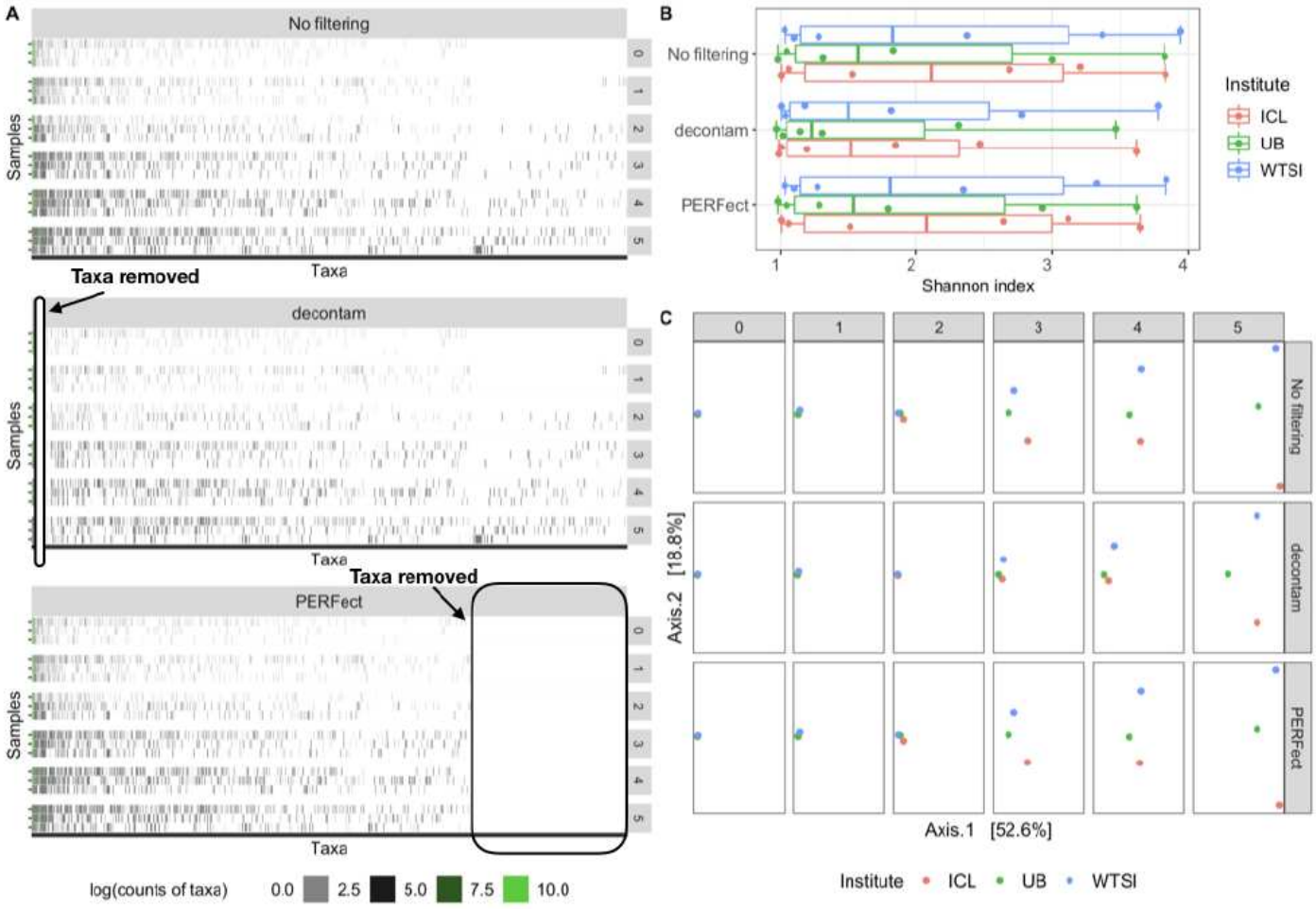

Figure 6

Comparison of the original data (no filtering), contaminant removal (decontam frequency) and filtering (PERFect simultaneous) methods. (A) Heat map of log transformed taxa counts in decreasing abundance order on the x-axis and samples by dilution level on the y-axis for the original data (top panel), data where contaminants are removed using decontam (middle panel) and rare taxa filtered using PERFect (bottom panel). True taxa are colored in green to the left of each heatmap; ovals indicate taxa removed by decontam and PERFect methods. (B) Alpha diversity for the three comparisons colored by dilution level and processing institute. (C) Beta diversity PCoA Bray-Curtis distances plots colored by processing institute and arranged by dilution level (rows) and three taxa removal methods (columns). Data source: [2].

Supplementary Files

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

- SupplementaryMaterialforFilteringPaper.pdf