Supplementary material for "PERFect: PERmutation Filtering test for microbiome data"

Ekaterina Smirnova *[1], S. Huzurbazar[2] and Farhad Jafari[3]

[1]Department of Mathematical Sciences, University of Montana
[2]Department of Statistics, University of Wyoming
[3]Department of Mathematics and Statistics, University of Wyoming

S1. Illustration of PERFect permutation algorithm

Filtering loss

To illustrate the intuition behind the choice of filtering loss $FL(J)$ and the PERFect permutation approach, we use a small subset of the mock community data 2: bias experiment data with 7 signal taxa. The number of taxa identified by the sequencing and bioinformatics pipeline was 46. From this data set, we have selected a total of $p = 20$ taxa and $n = 10$ samples. We have ordered columns by taxa abundance, such that the first 13 rare taxa are noise and the rest are signal. We have labeled taxa according to noise and signal abundance ordering, such that N_1 stands for the least abundant noise taxon, N_2 the second least abundant noise taxon, S_1 the first least abundant signal taxon, and so on. The corresponding counts data matrix is given in Table 1.

Table 1: Toy data with 10 samples and 20 taxa we use to illustrate PERFect permutation algorithm steps.

	Noise taxa													Signal taxa						
	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9	N_{10}	N_{11}	N_{12}	N_{13}	S_1	S_2	S_3	S_4	S_5	S_6	S_7
Sample 1	0	0	0	0	6	0	0	0	0	1	0	0	0	2422	0	2	4971	5493	0	1
Sample 2	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	12725	663	0	4926	3
Sample 3	0	0	0	0	0	0	1	0	0	0	0	3	1	3307	0	0	3252	2	3	2
Sample 4	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3199	0	0	0	1854	6501
Sample 5	0	9	0	0	0	2	0	0	3	24	3	36	4	1	19	4	51332	2	1	14
Sample 6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	14212	0	883	7	11
Sample 7	0	0	0	0	0	1	0	0	0	4	3	1	0	1020	1	1	0	2	0	23306
Sample 8	1	0	1	1	0	0	0	1	0	0	0	0	0	2094	45	12	1	4	47	14188
Sample 9	0	0	0	0	0	0	0	0	1	0	1	1	1	0	16037	0	1557	2217	5	30
Sample 10	0	0	0	0	0	0	2	4	0	0	0	0	0	14	17	4559	3770	2	25	4836

An underlying assumption for our filtering loss is that if a noise taxon is not important, then removing it will not dramatically affect the magnitude of taxa covariance matrix (up to a scaling factor) X^TX. We define the filtering loss statistic (2.2) due to removing a group of taxa, J, as

$$FL(J) = 1 - \frac{\|X_{-J}^T X_{-J}\|_F^2}{\|X^T X\|_F^2},$$

[*] Corresponding author: ekaterina.smirnova@mso.umt.edu

where X_{-J} is the $n \times (p - |J|)$ dimensional matrix obtained by removing the columns indexed by the set J from the data matrix X. To illustrate the ability of this filtering loss to distinguish between noise and signal taxa, we calculate filtering loss values for the toy data in Table 1. In this example, the magnitude of the full OTU table, denominator in the filtering loss function value, is 8.1258×10^{18}. Table 2 compares the magnitudes of reduced OTU tables and corresponding filtering loss values due to removing: 1) the 10th least abundant noise taxon, $J_1 = \{N_{10}\}$; 2) the 1st least abundant signal taxon $J_2 = \{S_1\}$; 3) all noise taxa, $J_3 = \{N_1, \dots, N_{13}\}$; and 4) all noise taxa plus the 1st least abundant signal, $J_4 = \{N_1, \dots, N_{13}, S_1\}$.

The filtered OTU table magnitude $\|X_{-J}^T X_{-J}\|_F^2$ for the 10th least abundant taxon is almost the same as that of the full OTU table $\|X^T X\|_F^2$ (displayed as the same value due to rounding), thus filtering loss is minimal, $FL(J) = 3.7876 \times 10^{-07}$ (-14.786 on the log scale). However, the loss due to removing the signal taxon is significantly larger $FL(J) = 9.3960 \times 10^{-04}$ (-6.970 on the log scale). Log filtering loss due to removing all noise taxa is minimal -13.545, thus removing an additional signal taxon S_1 increases overall log filtering loss for the set J_4 to -6.969. These differences reflect that all noise taxa cumulatively have less contribution than a single signal taxon.

Table 2: Filtering loss due to removing noise and signal taxa from the toy data set in Table 1.

	$\|X_{-J}^T X_{-J}\|_F^2$	$\|X^T X\|_F^2$	$FL(J)$	$\log[FL(J)]$
J_1	8.1258×10^{18}	8.1258×10^{18}	3.7876×10^{-07}	-14.786
J_2	8.1181×10^{18}	8.1258×10^{18}	9.3960×10^{-04}	-6.970
J_3	8.1257×10^{18}	8.1258×10^{18}	1.3109×10^{-06}	-13.545
J_4	8.1181×10^{18}	8.1258×10^{18}	9.4091×10^{-04}	-6.969

Permutation PERFect algorithm

```
Input:  OTU table 1, test critical value $\alpha = 0.10$
   1.  Run simultaneous PERFect algorithm to obtain taxa p-values $p_j, j = 1, \dots, p$

   2.  Order columns of $X$ such that $p_1 \geq p_2 \geq p_p$.
```

In the first row of counts table below, we list taxa simultaneous PERFect p-values. Notice that some taxa that were initially less important according to the abundance ordering, gained higher ranking according to simultaneous PERFect p-values. For example, N_2, the 2nd least important taxon in the abundance ordering, is now the 7th least important taxon in the simultaneous PERFect p-values ordering.

	Noise taxa													Signal taxa						
	N_1	N_3	N_4	N_5	N_6	N_7	N_2	N_8	N_9	N_{11}	N_{10}	N_{12}	N_{13}	S_6	S_4	S_5	S_1	S_3	S_2	S_7
Simultaneous PERFect p-values	NA	0.87	0.8	0.8	0.8	0.79	0.77	0.61	0.55	0.52	0.45	0.34	0.24	0.13	0.09	0.09	0.06	0.06	0.03	0.03

```
   3. For taxon $j = 1, \dots, p - 1$
```

```
    Let $J_j = \{1, \ldots, j\}$
    Calculate $DFL(j+1) = FL(J+1) - FL(J)$
end
```

The test statistic DFL and its corresponding values on the log scale are displayed below. The values of $\log(DFL)$ range between -23.71 and -13.98 for the noise taxa and increase dramatically $\log(DFL) = -6.79$ for the value of the first signal taxon S_6. The $\log(DFL)$ values for signal taxa range between -7.5 and -0.12, which is much larger compared to corresponding statistic values for the noise taxa. In the next step of the algorithm, we construct the distribution for each taxon N_1 through N_{13} and S_1 through S_7 to evaluate significance of corresponding taxa $\log(DFL)$ values.

						Noise taxa						
	N_1	N_3	N_4	N_5	N_6	N_7	N_2	N_8	N_9	N_{11}	N_{10}	N_{12}	N_{13}
DFL	NA	5.06×10^{-11}	5.06×10^{-11}	5.94×10^{-10}	2.73×10^{-09}	7.49×10^{-11}	5.25×10^{-08}	4.16×10^{-10}	6.02×10^{-09}	7.28×10^{-09}	3.79×10^{-07}	8.51×10^{-07}	1.13×10^{-08}
$\log(DFL)$	NA	-23.71	-23.71	-21.24	-19.72	-23.31	-16.76	-21.6	-18.93	-18.74	-14.79	-13.98	-18.3
							Signal taxa						
	S_6	S_4	S_5	S_1	S_3	S_2	S_7						
DFL	1.12×10^{-03}	8.90×10^{-01}	5.52×10^{-04}	7.67×10^{-04}	1.83×10^{-02}	8.92×10^{-03}	8.08×10^{-02}						
$\log(DFL)$	-6.79	-0.12	-7.5	-7.17	-4	-4.72	-2.52						

```
4. For taxon $j = 1, \ldots, p-1$
       For permutation $1, \ldots, k$
          Randomly select $J^*_{j+1} \subset \{1, \ldots, p\}$ with $|J^*_{j+1}| = j+1$
          Calculate $DFL^*(j+1) = FL(J^*_{j+1}) - FL(J^*)$
       end
   end
```

In this step, to build the distribution for j taxa filtering loss differences using permutations, we randomly draw k sets J^*_{j+1} taxa labels and calculate a sample of corresponding $DFL^*(j+1)$ values. For example, to obtain $k = 2$ permutations for 10 taxa, we draw sets of size $|J^*_{j+1}| = 10$. The filtering loss differences $DFL^*(j+1)$ are calculated according to the ordering given by permutation.

In particular, for $k = 1$ permutation, $J^*_{j+1} = \{N_6, N_{10}, N_2, S_5, N_7, S_7, N_1, S_3, S_1, N_3\}$ and thus $DFL^*(j+1)$ is calculated as

$$DFL^*(10) = FL^*(\{N_6, N_{10}, N_2, S_5, N_7, S_7, N_1, S_3, S_1, N_3\}) - FL^*(\{N_6, N_{10}, N_2, S_5, N_7, S_7, N_1, S_3, S_1\}).$$

For $k = 2$ permutation,

$$DFL^*(10) = FL^*(\{N_{12}, N_2, N_9, S_2, S_5, N_7, S_7, N_4, N_{13}, N_1\}) - FL^*(\{N_{12}, N_2, N_9, S_2, S_5, N_7, S_7, N_4, N_{13}\}).$$

```
5. For taxon $j = 1, \ldots, p-1$
      Using quantile matching fit the Skew Normal distribution to the
       logarithm of the sample $DFL^*(j+1), j = 1, \ldots, p-1$ to obtain
       the null distribution $X_{j+1} \sim \mathrm{SN}(\widehat{\xi}_j, \widehat{\omega}^2_j, \widehat{\alpha}_j)$
   end
```

For example, to estimate the parameters of 10 taxa distribution with k permutations, we would use $\log(DFL^*(10))$ values $\{-34.49, -27.55, \ldots, k\}$ to fit Skew Normal distribution using quantile matching method. We increase the number of permutations to $k = 1000$, which is necessary to get a reasonable distribution for the values of $DFL^*(j+1)$. Figure 1 illustrates the histogram of $\log(DFL^*(10))$ sample, where with the blue line indicates Skew Normal fit. Because we have only 20 taxa in this example, the distribution fit is not as accurate as for the larger number of taxa, but nevertheless, it illustrates the idea. The estimated distribution parameters are $\xi = -18.3, \omega = 7.11, \alpha = -0.03$.

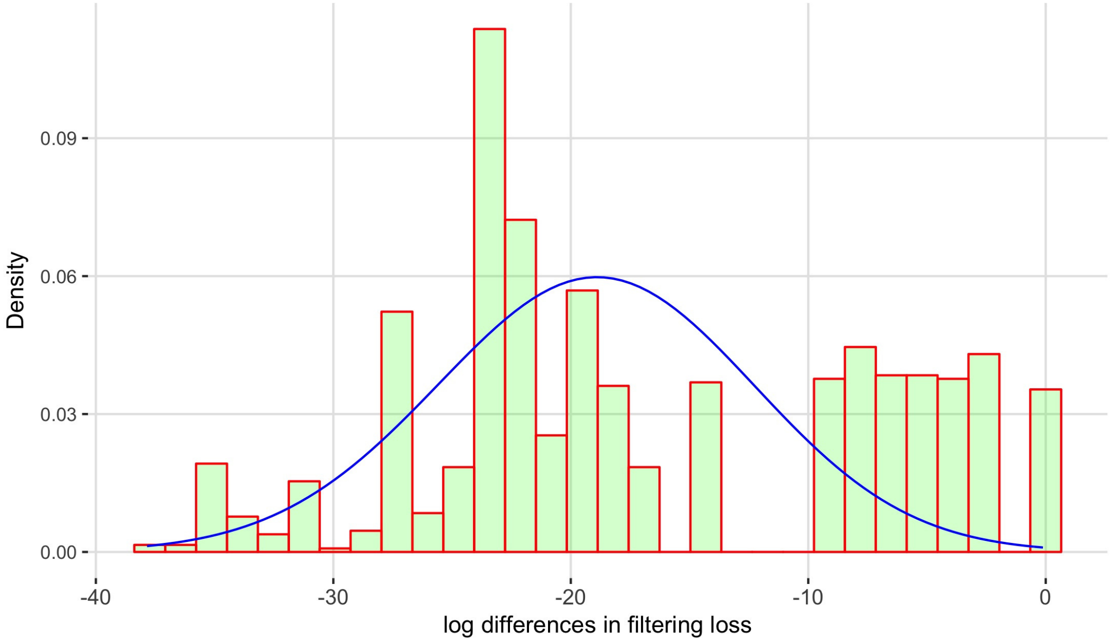


Figure 1: Histogram of log-transformed DFL values for the Fettweis et al., 2012 data. The blue line indicates $\mathrm{SN}(\xi = -18.3, \omega^2 = 7.11^2, \alpha = -0.03)$ density fitted to the log-transformed data using quantile matching method.

```
6. For taxon $j = 1, \ldots, p-1$
      Calculate the p-value $p_{j+1}$ for $DFL(j+1), j = 1, \ldots, p-1$ as
        $p_{j+1} := P[X_{j+1} > \log\{DFL(j+1)\}]$
    end
```

The $log(DFL) = -18.74$ value for the 10th taxon in simultaneous p-values ordering was calculated in step 3. Therefore, we calculate the 10th taxon p-value as

$$p_{10} = P[X_{10} > 18.74], \quad \text{where } X_{10} \sim \mathrm{SN}(\widehat{\xi}_{10} = -18.30, \widehat{\omega}^2_{10} = 7.11^2, \widehat{\alpha}_{10} = -0.03).$$

```
6. Average 3 subsequent p-values
```

The 4 rows of the example OTU table below combines taxa PERFect simultaneous p-values, their corresponding $log(DFL)$ values, raw PERFect permutation p-values, and their corresponding averaged values. As expected, the p-values of noise taxa are large and the p-values for the signal taxa are small.

	Noise taxa													Signal taxa						
	N_1	N_3	N_4	N_5	N_6	N_7	N_2	N_8	N_9	N_{11}	N_{10}	N_{12}	N_{13}	S_6	S_4	S_5	S_1	S_3	S_2	S_7
Simultaneous PERFect p-values	NA	0.87	0.8	0.8	0.8	0.79	0.77	0.61	0.55	0.52	0.45	0.34	0.24	0.13	0.09	0.09	0.06	0.06	0.03	0.03
$\log(DFL)$	NA	-23.71	-23.71	-21.24	-19.72	-23.31	-16.76	-21.6	-18.93	-18.74	-14.79	-13.98	-18.3	-6.79	-0.12	-7.5	-7.17	-4	-4.72	-2.52
Permutation PERFect p-values	NA	0.89	0.90	0.71	0.55	0.82	0.36	0.65	0.44	0.52	0.22	0.08	0.43	0.00	0.00	0.03	0.06	0.03	0.00	0.00
Averaged p-values	NA	0.83	0.72	0.69	0.58	0.61	0.48	0.54	0.39	0.27	0.24	0.17	0.14	0.01	0.03	0.04	0.03	0.01	0.01	0.00

7. `Filter the set of taxa` J_j `with the first p-value such that` $p_{j+1} \leq \alpha$

The OTU table above indicates that the first significantly small averaged p-value is $p_{14} = 0.01 \leq 0.1 =: \alpha$, which occurs at the first signal taxon S_6. Thus the filtered set of taxa $J_{13} = \{N_1, N_3, N_4, N_5, N_6, N_7, N_2, N_8, N_9, N_{11}, N_{10}, N_{12}, N_{13}\}$. Therefore, we preserve the last 7 columns in the final data set that correspond to the true signal taxa.

S2. Vaginal microbiome data set analysis

Table 3: Correlation table for the true (signal) taxa in the mock 1 Fettweis et al. [2012] data set.

	L. crispatus cluster	Staphyl. cluster47	Fusobact. cluster48	Gardner. vaginalis	Prevotella bivia
Enterococcus faecalis	0.79	0.79	-0.58	0.27	0.40
L. crispatus cluster		0.84	-0.3	0.14	0.30
Staphyl. cluster47			-0.29	-0.07	0.16
Fusobact. cluster48				-0.50	-0.22
Gardner. vaginalis					0.02

Table 4: Taxa common to the two traditional filtering rules, simultaneous and permutation PERFect approaches for the Ravel et al. [2011] vaginal data set.

L. iners	L. crispatus	L. gasseri	L. jensenii	Prevotella
Megasphaera	Sneathia	Streptococcus	Atopobium	Lachnospiraceae 8
Dialister	Anaerococcus	Eggerthella	Ruminococcaceae 3	Segniliparus
Prevotellaceae 1	Peptoniphilus	Ureaplasma	Finegoldia	Aerococcus
L. vaginalis	Staphylococcus	Parvimonas	Lactobacillales 2	Veillonella
Corynebacterium	Bacteroides	Lactobacillales 6	Gardnerella	Lactobacillales 5
Prevotellaceae 2	Peptostreptococcus	Mobiluncus	Porphyromonas	Clostridiales 17
Ruminococcaceae Incertae Sedis	Ruminococcaceae 4	Actinomyces	Anaeroglobus	Campylobacter
Gemella	Lactobacillales 7			

Effect of taxa ordering

Section 2 described an approach that relied on ranking taxa according to the number of occurrences NP (2.3) This is not the only reasonable ranking and we introduce three additional taxa ranking criteria.

1. *Simultaneous PERFect p-values*: Simultaneous PERFect (section 2.2.1) estimates the null distribution for the difference in filtering loss $DFL(j+1)$ for taxa $j = 1, \ldots, p-1$ and calculates simultaneous PERFect corresponding p-values according to equation (2.5) in the paper where smaller

Table 5: Comparison of traditional, and PERFect filtering results for the Ravel et al. [2011] vaginal microbiome data set. For PERFect filtering we use $\alpha = 0.10$ significance level to determine taxa to retain in the data set.

		# Taxa preserved	% Filtered
PERFect	Simultaneous abundance	42	83.00
	Permutation abundance	71	71.26
	Permutation p-values	63	74.49
Traditional	Rule 1	135	45.34
	Rule 2	126	48.99

values of p_j indicate higher importance of the jth taxon.
2. *Number of connected taxa*: For the jth taxon with $j = 1, \ldots, p$, the number of connected taxa is defined as the number of non-zero $\boldsymbol{x}_i^T \boldsymbol{x}_j$ with $i \neq j$,

$$NC(j) := \sum_{i=1, i\neq j}^{p} I(\boldsymbol{x}_i^T \boldsymbol{x}_j \neq 0),$$

where $I(\cdot)$ is the indicator function. This metric counts the number of jth taxon co-occurrences with other taxa.
3. *Weighted number of connected taxa*: We weight $NC(j)$ by the relative number of samples containing jth taxon to take into account taxa presence,

$$NC_w(j) := \frac{\tilde{n}_j}{n} \sum_{i=1, i\neq j}^{p} I(\boldsymbol{x}_i^T \boldsymbol{x}_j \neq 0),$$

where n is the total number of samples and $\tilde{n}_j$ is the number of samples in which the jth taxon has non-zero counts.

We analyzed the effect of the NP (2.3), simultaneous PERFect p-values (2.5), NC, and NC_w taxa orderings on the PERFect filtering. The number of preserved taxa for the two mock and the vaginal microbiome data sets are compared in Table 6. Results reveal that ordering does not affect detection of correct taxa in the mock data sets. The proposed *simultaneous PERFect p-values (2.5) ordering* reduces the mock data set 1 to a set of 19 and 6 taxa in the simultaneous and permutation PERFect respectively, which is the smallest set that includes the 6 true taxa among both traditional and PERFect filtering with alternative taxa orderings. We recorded the true taxa importance rank based on the four proposed orderings for the two mock data sets in Table 7, where the true taxa in the mock data set 1 are consistently ranked as the 6 most important taxa, indicating that the proposed orderings provide the correct ranking of taxa importance.

While we do not have the gold standard for the Ravel et al. [2011] data, we can compare the results we obtained with the analysis results provided by the authors in their paper. We are especially interested in making sure that taxa that were found to be relevant and descriptive in the Ravel et al. [2011] paper are ranked as important by PERFect and are not filtered out. This is especially important given that PERFect has proven to be much more aggressive at removing taxa than standard techniques. Ravel et al. [2011] analyzed the composition and structure of vaginal communities and identified core microbiomes by grouping samples into five community state types

(CSTs) according to taxa relative abundance and Spearman correlation profiles using hierarchical clustering techniques; results were presented in Figure 1A and Supplementary Table 5 in Ravel et al. [2011]. CST I, II, III, and V were dominated by Lactobacillus *L.crispatus, L.gasseri, L.iners* and *L.jensenii* species respectively, while CST IV was characterized as a diversity group (Table 1 in Ravel et al. [2011]). Correspondence of taxa arranged in decreasing NP order to simultaneous PERFect p-values, NC and NC_w values, and taxa average abundance in CST groups are presented in Figures 3-6 of the supplementary material. Results in Figure 3 reveal that CST I, II, III, and V dominant taxa are ranked as the 2nd, 5th, 1st and 9th significant taxa according to the simultaneous PERFect p-values ordering. Moreover, the other taxa with significant simultaneous PERFect p-values are the most abundant taxa in the diverse CST IV group. For example, Lachnospiraceae 8, Proteobacteria 12, Mycoplasmataceae 1 are ranked as the 32nd, 48th and 34th most significant PERFect taxa and have 4.51%, 2.43% and 1.79% abundance levels in CST IV, respectively. However, these taxa are the 42nd, 55th and 51th ranked NP taxa respectively. This important information, which is not revealed by the NP ordering used in traditional filtering methods, suggests that lower (and thus more significant) PERFect p-values can be used not only for filtering decision, but also as a taxa importance classification method. Such classification may be useful in studying taxa associated with risk of BV, a disease characterized by imbalance of vaginal bacterial species. Finally, taxa with low connectivity values are among the least abundant taxa, indicating that PERFect does not contradict the results of traditional filtering criteria for rare taxa.

Table 6: Ordering effect comparison for simultaneous and permutation PERFect filtering results using: 1) mock data set 1 (Fettweis et al. [2012]); 2) mock data set 2 (Brooks et al. [2015]); and 3) vaginal microbiome data set (Ravel et al. [2011]). Significance level $\alpha = 0.10$ was used to determine filtering cut off.

		Mock Data Set 1		Mock Data Set 2		Vaginal Microbiome	
		Positive Controls Data		Bias Experiment Data		Data Set	
		# Taxa Preserved	%Filtered	# Taxa Preserved	%Filtered	# Taxa Preserved	%Filtered
Simultaneous	NP	22	77.78	10	78.26	42	83.00
	p-values	22	77.78	10	78.26	48	80.57
	NC	19	80.81	10	78.26	42	83.00
	NC_w	20	79.80	10	78.26	43	82.59
Permutation	NP	17	82.83	8	82.61	71	71.26
	p-values	17	82.83	8	82.61	63	74.49
	NC	17	82.83	8	82.61	76	69.23
	NC_w	17	82.83	8	82.61	72	70.85

Table 7: True taxa ranking in the four proposed taxa orderings for: 1) mock data set 1 (Fettweis et al. [2012]); and 2) mock data set 2 (Brooks et al. [2015]). Higher rank indicates the taxon's larger importance in given ordering.

Mock Data Set 1					Mock Data Set 2				
Positive Controls Data					Bias Experiment Data				
	NP	p-values	NC	NC_w		NP	p-values	NC	NC_w
Prevotella.bivia	1	4	2	1	Sneathia.amnii	1	1	3	1
Gardnerella.vaginalis	2	1	4	2	Lactobacillus.iners	2	2	1	2
Fusobacterium.cluster48	3	5	5	3	Streptococcus.agalactiae	3	3	5	5
Staphylococcus.cluster47	4	2	1	4	Lactobacillus.crispatus cluster	4	5	2	3
Lactobacillus.crispatus cluster	5	3	3	5	Atopobium.vaginae	5	5	4	4
Enterococcus.faecalis	6	6	6	6	Prevotella.bivia	6	6	6	6
					Gardnerella.vaginalis	7	7	7	7

Choice of test level

The PERFect filtering cut-off depends on the choice of the test α level, which we set to 0.10 to obtain the results discussed in Sections 4 of the manuscript and S2 of the supplementary materials. In practice, this choice might filter out a taxon with undesirably large filtering loss or other importance measure. We therefore suggest examining the p-values plots to confirm the choice of α. Figure 2 illustrates the plot of permutation PERFect (Algorithm 2) p-values for the Ravel et al., 2011 data. For each taxon on the x-axis, we color its p-value on the $y-$axis according to the quantiles of individual filtering loss FL_u (2.1). The dashed horizontal red line indicates the filtering at $\alpha = 0.10$ level, therefore taxa to the left of the dashed purple vertical line correspond to the filtered set J and to the right of this line to the set $\{-J\}$ of taxa retained for further analysis. The plot indicates that taxa retained in this data set at the 0.10 level have the 80% largest percentile of FL_u values. This plot, which can be colored by alternative taxa importance criteria discussed in Section S2 of the supplementary materials, reflects the effect of the test level on the retained taxa and corresponds to taxa importance information.

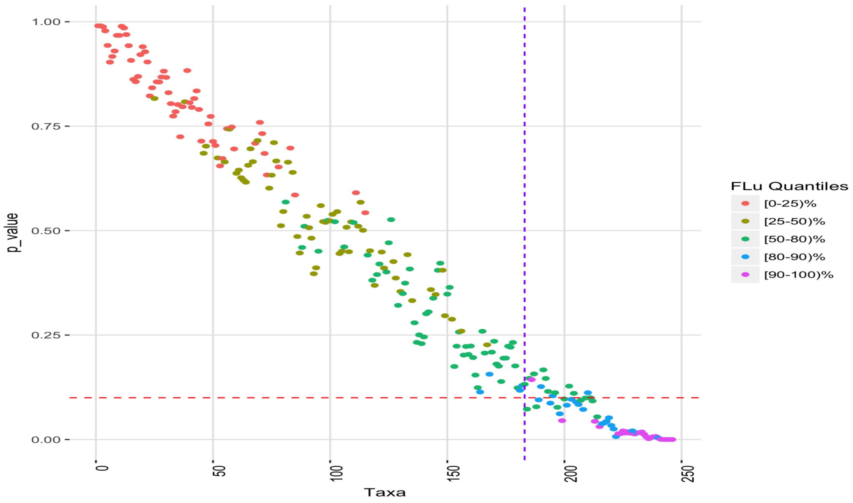
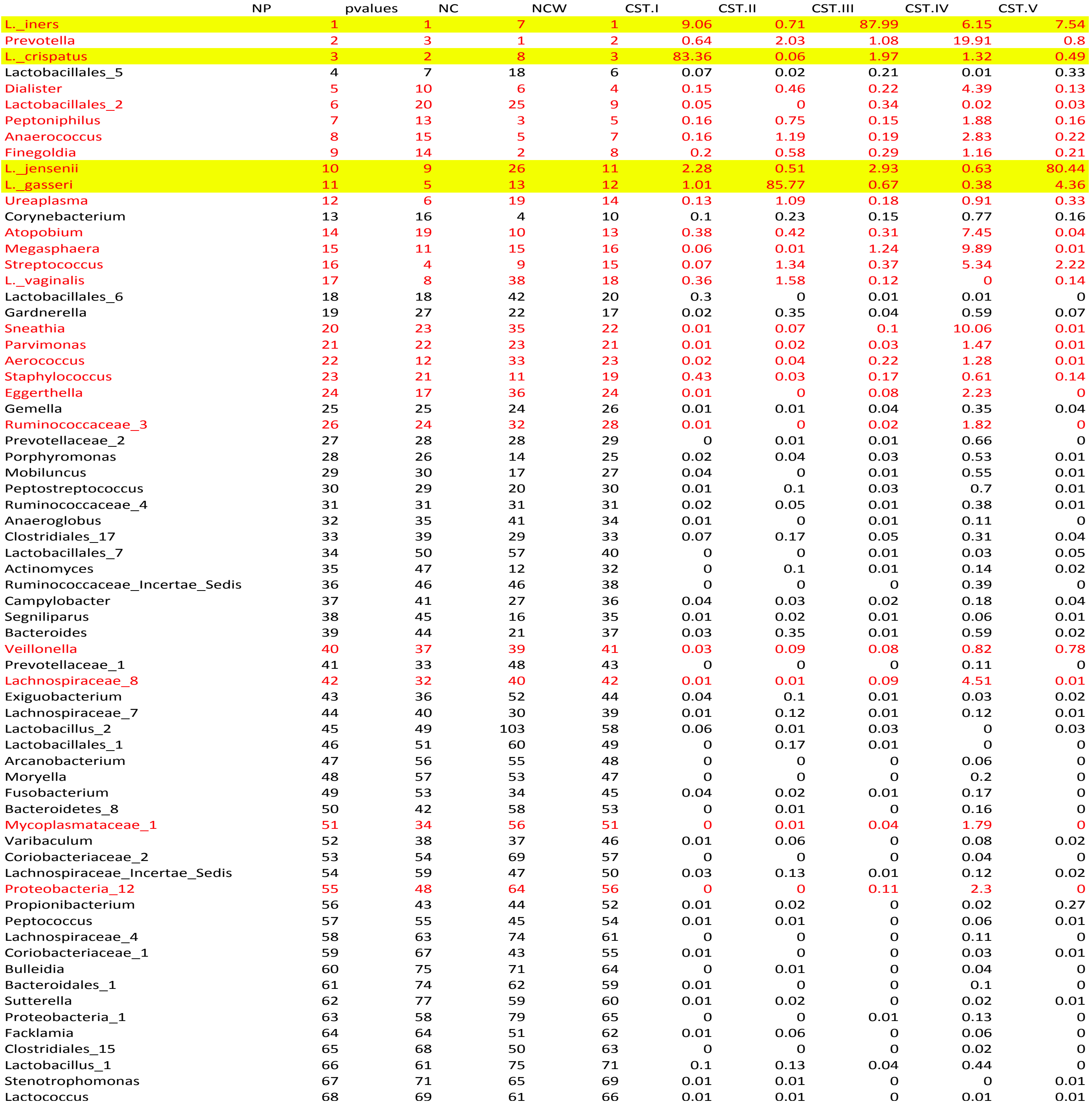
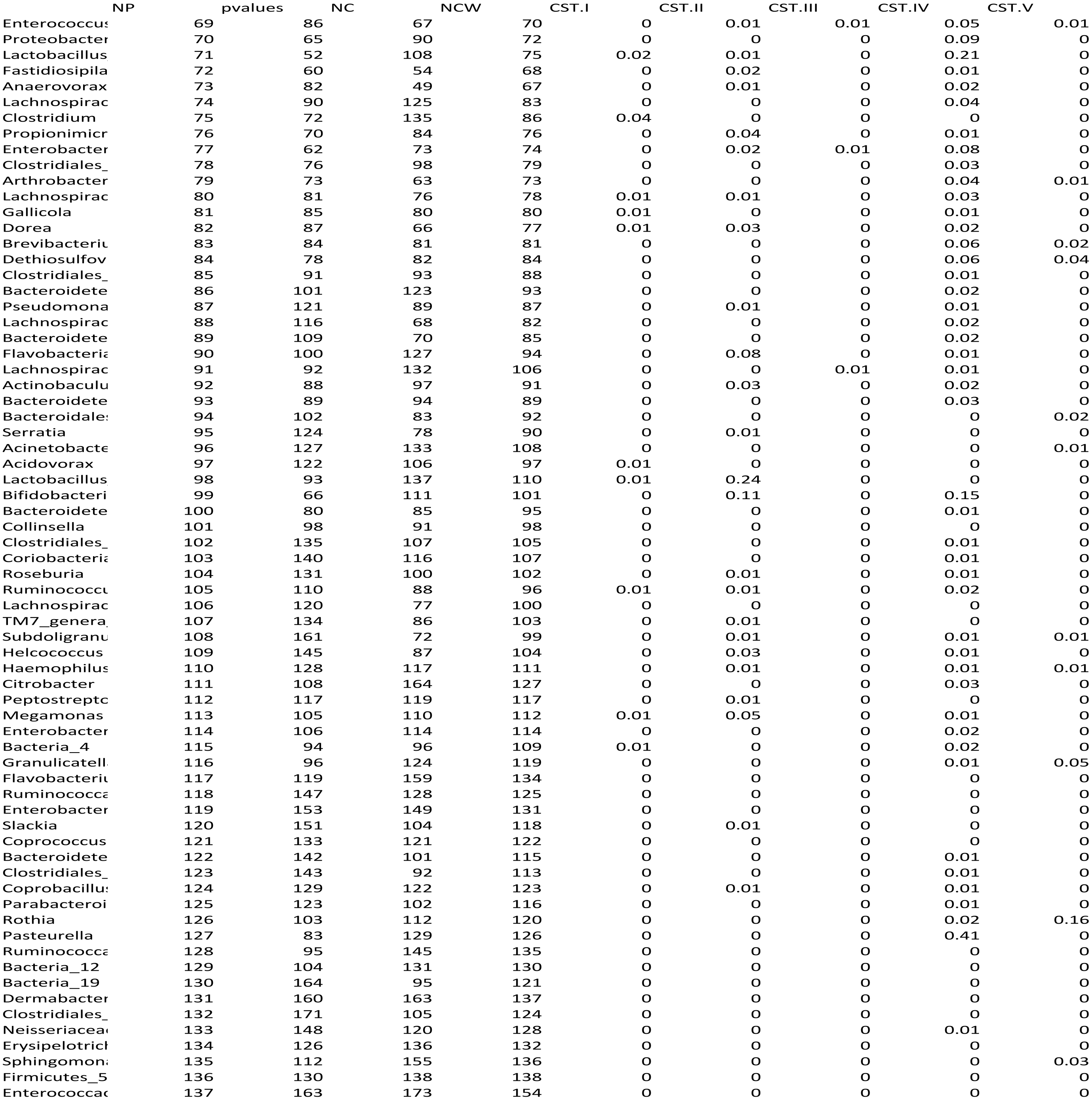


Figure 2: Permutation PERFect p-values for the Ravel et al., 2011 data. Taxa on the $x-$axis, arranged in order of simultaneous PERFect p-values (Algorithm 2), are represented by points colored according to the FL_u (2.1) quantile values. The dashed horizontal red line indicates the $\alpha = 0.10$ cutoff. Taxa to the left of the dashed purple vertical line correspond to the set of filtered out taxa J and to the right of this line to the set $\{-J\}$ of retained taxa.



	NP	pvalues	NC	NCW	CST.I	CST.II	CST.III	CST.IV	CST.V
L._iners	1	1	7	1	9.06	0.71	87.99	6.15	7.54
Prevotella	2	3	1	2	0.64	2.03	1.08	19.91	0.8
L._crispatus	3	2	8	3	83.36	0.06	1.97	1.32	0.49
Lactobacillales_5	4	7	18	6	0.07	0.02	0.21	0.01	0.33
Dialister	5	10	6	4	0.15	0.46	0.22	4.39	0.13
Lactobacillales_2	6	20	25	9	0.05	0	0.34	0.02	0.03
Peptoniphilus	7	13	3	5	0.16	0.75	0.15	1.88	0.16
Anaerococcus	8	15	5	7	0.16	1.19	0.19	2.83	0.22
Finegoldia	9	14	2	8	0.2	0.58	0.29	1.16	0.21
L._jensenii	10	9	26	11	2.28	0.51	2.93	0.63	80.44
L._gasseri	11	5	13	12	1.01	85.77	0.67	0.38	4.36
Ureaplasma	12	6	19	14	0.13	1.09	0.18	0.91	0.33
Corynebacterium	13	16	4	10	0.1	0.23	0.15	0.77	0.16
Atopobium	14	19	10	13	0.38	0.42	0.31	7.45	0.04
Megasphaera	15	11	15	16	0.06	0.01	1.24	9.89	0.01
Streptococcus	16	4	9	15	0.07	1.34	0.37	5.34	2.22
L._vaginalis	17	8	38	18	0.36	1.58	0.12	0	0.14
Lactobacillales_6	18	18	42	20	0.3	0	0.01	0.01	0
Gardnerella	19	27	22	17	0.02	0.35	0.04	0.59	0.07
Sneathia	20	23	35	22	0.01	0.07	0.1	10.06	0.01
Parvimonas	21	22	23	21	0.01	0.02	0.03	1.47	0.01
Aerococcus	22	12	33	23	0.02	0.04	0.22	1.28	0.01
Staphylococcus	23	21	11	19	0.43	0.03	0.17	0.61	0.14
Eggerthella	24	17	36	24	0.01	0	0.08	2.23	0
Gemella	25	25	24	26	0.01	0.01	0.04	0.35	0.04
Ruminococcaceae_3	26	24	32	28	0.01	0	0.02	1.82	0
Prevotellaceae_2	27	28	28	29	0	0.01	0.01	0.66	0
Porphyromonas	28	26	14	25	0.02	0.04	0.03	0.53	0.01
Mobiluncus	29	30	17	27	0.04	0	0.01	0.55	0.01
Peptostreptococcus	30	29	20	30	0.01	0.1	0.03	0.7	0.01
Ruminococcaceae_4	31	31	31	31	0.02	0.05	0.01	0.38	0.01
Anaeroglobus	32	35	41	34	0.01	0	0.01	0.11	0
Clostridiales_17	33	39	29	33	0.07	0.17	0.05	0.31	0.04
Lactobacillales_7	34	50	57	40	0	0	0.01	0.03	0.05
Actinomyces	35	47	12	32	0	0.1	0.01	0.14	0.02
Ruminococcaceae_Incertae_Sedis	36	46	46	38	0	0	0	0.39	0
Campylobacter	37	41	27	36	0.04	0.03	0.02	0.18	0.04
Segniliparus	38	45	16	35	0.01	0.02	0.01	0.06	0.01
Bacteroides	39	44	21	37	0.03	0.35	0.01	0.59	0.02
Veillonella	40	37	39	41	0.03	0.09	0.08	0.82	0.78
Prevotellaceae_1	41	33	48	43	0	0	0	0.11	0
Lachnospiraceae_8	42	32	40	42	0.01	0.01	0.09	4.51	0.01
Exiguobacterium	43	36	52	44	0.04	0.1	0.01	0.03	0.02
Lachnospiraceae_7	44	40	30	39	0.01	0.12	0.01	0.12	0.01
Lactobacillus_2	45	49	103	58	0.06	0.01	0.03	0	0.03
Lactobacillales_1	46	51	60	49	0	0.17	0.01	0	0
Arcanobacterium	47	56	55	48	0	0	0	0.06	0
Moryella	48	57	53	47	0	0	0	0.2	0
Fusobacterium	49	53	34	45	0.04	0.02	0.01	0.17	0
Bacteroidetes_8	50	42	58	53	0	0.01	0	0.16	0
Mycoplasmataceae_1	51	34	56	51	0	0.01	0.04	1.79	0
Varibaculum	52	38	37	46	0.01	0.06	0	0.08	0.02
Coriobacteriaceae_2	53	54	69	57	0	0	0	0.04	0
Lachnospiraceae_Incertae_Sedis	54	59	47	50	0.03	0.13	0.01	0.12	0.02
Proteobacteria_12	55	48	64	56	0	0	0.11	2.3	0
Propionibacterium	56	43	44	52	0.01	0.02	0	0.02	0.27
Peptococcus	57	55	45	54	0.01	0.01	0	0.06	0.01
Lachnospiraceae_4	58	63	74	61	0	0	0	0.11	0
Coriobacteriaceae_1	59	67	43	55	0.01	0	0	0.03	0.01
Bulleidia	60	75	71	64	0	0.01	0	0.04	0
Bacteroidales_1	61	74	62	59	0.01	0	0	0.1	0
Sutterella	62	77	59	60	0.01	0.02	0	0.02	0.01
Proteobacteria_1	63	58	79	65	0	0	0.01	0.13	0
Facklamia	64	64	51	62	0.01	0.06	0	0.06	0
Clostridiales_15	65	68	50	63	0	0	0	0.02	0
Lactobacillus_1	66	61	75	71	0.1	0.13	0.04	0.44	0
Stenotrophomonas	67	71	65	69	0.01	0.01	0	0	0.01
Lactococcus	68	69	61	66	0.01	0.01	0	0.01	0.01

Figure 3: List of taxa arranged in order of decreasing number of occurrences NP. Taxa abundance rank is calculated using average percentage abundance of each taxon across samples. Community state types CST I-V and corresponding taxa average abundance in each CSTare taken from Supplementary Table 5 in Ravel et al. [2011]. Taxa highlighted in yellow correspond to CST I, II, III and V taxa. Red colored taxa correspond to the 25 core taxa common to all filtering methods with respect to different taxa orderings. Missing data are coded as NA.

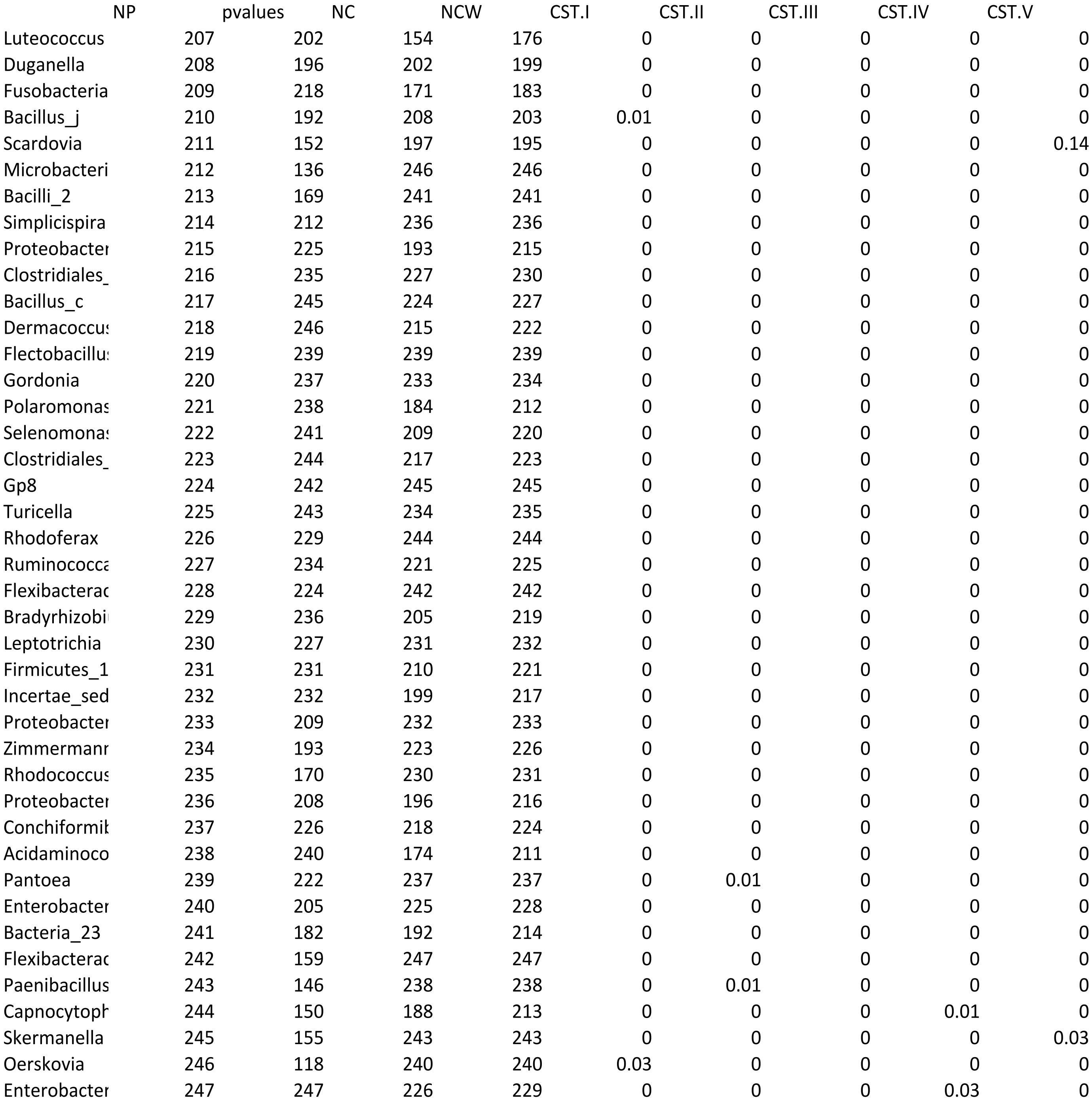


	NP	pvalues	NC	NCW	CST.I	CST.II	CST.III	CST.IV	CST.V
Enterococcus	69	86	67	70	0	0.01	0.01	0.05	0.01
Proteobacter	70	65	90	72	0	0	0	0.09	0
Lactobacillus	71	52	108	75	0.02	0.01	0	0.21	0
Fastidiosipila	72	60	54	68	0	0.02	0	0.01	0
Anaerovorax	73	82	49	67	0	0.01	0	0.02	0
Lachnospirac	74	90	125	83	0	0	0	0.04	0
Clostridium	75	72	135	86	0.04	0	0	0	0
Propionimicr	76	70	84	76	0	0.04	0	0.01	0
Enterobacter	77	62	73	74	0	0.02	0.01	0.08	0
Clostridiales_	78	76	98	79	0	0	0	0.03	0
Arthrobacter	79	73	63	73	0	0	0	0.04	0.01
Lachnospirac	80	81	76	78	0.01	0.01	0	0.03	0
Gallicola	81	85	80	80	0.01	0	0	0.01	0
Dorea	82	87	66	77	0.01	0.03	0	0.02	0
Brevibacteriu	83	84	81	81	0	0	0	0.06	0.02
Dethiosulfov	84	78	82	84	0	0	0	0.06	0.04
Clostridiales_	85	91	93	88	0	0	0	0.01	0
Bacteroidete	86	101	123	93	0	0	0	0.02	0
Pseudomona	87	121	89	87	0	0.01	0	0.01	0
Lachnospirac	88	116	68	82	0	0	0	0.02	0
Bacteroidete	89	109	70	85	0	0	0	0.02	0
Flavobacteria	90	100	127	94	0	0.08	0	0.01	0
Lachnospirac	91	92	132	106	0	0	0.01	0.01	0
Actinobaculu	92	88	97	91	0	0.03	0	0.02	0
Bacteroidete	93	89	94	89	0	0	0	0.03	0
Bacteroidales	94	102	83	92	0	0	0	0	0.02
Serratia	95	124	78	90	0	0.01	0	0	0
Acinetobacte	96	127	133	108	0	0	0	0	0.01
Acidovorax	97	122	106	97	0.01	0	0	0	0
Lactobacillus	98	93	137	110	0.01	0.24	0	0	0
Bifidobacteri	99	66	111	101	0	0.11	0	0.15	0
Bacteroidete	100	80	85	95	0	0	0	0.01	0
Collinsella	101	98	91	98	0	0	0	0	0
Clostridiales_	102	135	107	105	0	0	0	0.01	0
Coriobacteria	103	140	116	107	0	0	0	0.01	0
Roseburia	104	131	100	102	0	0.01	0	0.01	0
Ruminococcu	105	110	88	96	0.01	0.01	0	0.02	0
Lachnospirac	106	120	77	100	0	0	0	0	0
TM7_genera_	107	134	86	103	0	0.01	0	0	0
Subdoligranu	108	161	72	99	0	0.01	0	0.01	0.01
Helcococcus	109	145	87	104	0	0.03	0	0.01	0
Haemophilus	110	128	117	111	0	0.01	0	0.01	0.01
Citrobacter	111	108	164	127	0	0	0	0.03	0
Peptostrepto	112	117	119	117	0	0.01	0	0	0
Megamonas	113	105	110	112	0.01	0.05	0	0.01	0
Enterobacter	114	106	114	114	0	0	0	0.02	0
Bacteria_4	115	94	96	109	0.01	0	0	0.02	0
Granulicatell	116	96	124	119	0	0	0	0.01	0.05
Flavobacteriu	117	119	159	134	0	0	0	0	0
Ruminococca	118	147	128	125	0	0	0	0	0
Enterobacter	119	153	149	131	0	0	0	0	0
Slackia	120	151	104	118	0	0.01	0	0	0
Coprococcus	121	133	121	122	0	0	0	0	0
Bacteroidete	122	142	101	115	0	0	0	0.01	0
Clostridiales_	123	143	92	113	0	0	0	0.01	0
Coprobacillus	124	129	122	123	0	0.01	0	0.01	0
Parabacteroi	125	123	102	116	0	0	0	0.01	0
Rothia	126	103	112	120	0	0	0	0.02	0.16
Pasteurella	127	83	129	126	0	0	0	0.41	0
Ruminococca	128	95	145	135	0	0	0	0	0
Bacteria_12	129	104	131	130	0	0	0	0	0
Bacteria_19	130	164	95	121	0	0	0	0	0
Dermabacter	131	160	163	137	0	0	0	0	0
Clostridiales_	132	171	105	124	0	0	0	0	0
Neisseriacea	133	148	120	128	0	0	0	0.01	0
Erysipelotrich	134	126	136	132	0	0	0	0	0
Sphingomon	135	112	155	136	0	0	0	0	0.03
Firmicutes_5	136	130	138	138	0	0	0	0	0
Enterococcac	137	163	173	154	0	0	0	0	0

Figure 4: List of taxa (ctd.) arranged in order of decreasing number of occurrences NP. Taxa abundance rank is calculated using average percentage abundance of each taxon across samples. Community state types CST I-V and corresponding taxa average abundance in each CST are taken from Supplementary Table 5 in Ravel et al. [2011]. Missing data are coded as NA.

	NP	pvalues	NC	NCW	CST.I	CST.II	CST.III	CST.IV	CST.V
Proteobacter	139	168	143	140	0	0	0	0	0
Firmicutes_2	140	174	167	150	NA	NA	NA	NA	NA
Incertae_sed	141	167	156	148	0	0	0	0	0
Firmicutes_3	142	184	144	141	0	0	0	0	0
Janthinobact	143	172	186	159	0	0.01	0	0	0
Bacteria_17	144	165	113	133	0	0.01	0	0	0
Betaproteob	145	157	183	158	0	0	0	0	0
Ruminococca	146	156	99	129	0	0	0	0.01	0
Kocuria	147	137	150	146	0	0	0	0.01	0.01
Neisseria	148	115	147	143	0	0	0	0.03	0
Bilophila	149	114	126	147	0	0	0	0	0
Aquabacteriu	150	144	178	170	0	0	0	0	0
Chryseobacte	151	176	185	173	0	0	0	0	0
Methylobact	152	194	191	175	0	0	0	0	0
Firmicutes_4	153	185	222	189	0	0	0	0	0
Enhydrobact	154	191	151	155	0	0	0	0	0
Flexibacterac	155	197	201	178	0	0	0	0	0
Dermabacter	156	201	161	162	0	0	0	0	0
Bacteroidete	157	187	152	156	0	0	0	0	0
Lachnospirac	158	180	158	160	0	0	0	0	0
Bacteroidete	159	188	118	144	0	0	0	0	0
Proteobacter	160	166	168	164	0	0	0	0.01	0
Bacteria_6	161	154	180	171	0	0	0	0	0
Proteobacter	162	125	169	165	0	0	0	0	0
GpIX	163	138	229	196	0	0	0	0	0
Salmonella	164	139	214	185	0	0	0	0	0
Succinispira	165	149	141	152	0	0	0	0	0
Bacteria_15	166	162	142	153	0	0	0	0	0
Clostridiales_	167	179	109	139	0	0	0	0	0
Bacteria_5	168	186	115	142	0	0	0	0	0
Clostridiales_	169	173	130	149	0	0	0	0	0
Micrococcus	170	175	172	166	0	0	0	0	0
Bacteria_22	171	158	175	167	0	0	0	0.01	0
GpI	172	141	204	180	0.01	0	0	0	0
Atopobacter	173	111	165	163	0	0	0	0.01	0
Burkholderia	174	113	139	151	0	0	0	0.01	0
Enterobacter	175	107	176	168	0	0	0	0.02	0
Alistipes	176	97	160	161	0	0	0	0.03	0
Klebsiella	177	79	153	157	0	0	0	0.31	0
Patulibacter	178	99	198	197	0	0	0	0	0
Roseomonas	179	132	219	208	0	0	0	0	0
Bacillales_5	180	198	216	207	0	0	0	0	0
Microbacteri	181	199	235	218	0	0	0	0	0
Flexibacterac	182	190	213	206	0	0	0	0	0
Novosphingc	183	189	228	210	0	0	0	0	0
Pseudomona	184	204	194	193	0	0	0	0	0
Riemerella	185	217	189	191	0	0	0	0	0
Corynebacte	186	219	179	186	0	0	0	0	0
Janibacter	187	215	187	190	0	0	0	0	0
Methylovoru	188	213	220	209	0	0	0	0	0
Proteobacter	189	211	181	187	0	0	0	0	0
Rhizobium	190	206	190	192	0	0	0	0	0
Clostridiales_	191	210	206	201	0	0	0	0	0
Clostridiales_	192	221	140	172	0	0	0	0	0
OD1_genera_	193	233	146	174	0	0	0	0	0
Clostridiales_	194	230	134	169	0	0	0	0	0
Actinobacillu	195	228	203	200	0	0	0	0	0
Ruminococca	196	220	166	181	0	0	0	0	0
Flexibacterac	197	223	195	194	0	0	0	0	0
Bacteroidete	198	200	170	182	0	0	0	0	0
Bacteria_11	199	183	177	184	0	0	0	0	0
Jeotgalicoccu	200	177	157	177	0	0	0	0	0
Bacteria_8	201	178	211	204	0	0	0	0	0
Clostridia_2	202	203	200	198	0	0	0	0	0
Veillonellace	203	207	182	188	0	0	0	0	0
Crenotrichac	204	214	207	202	0	0	0	0	0
SR1_genera_	205	216	162	179	0	0	0	0	0
Catenibacter	206	195	212	205	0	0	0	0	0

Figure 5: List of taxa (ctd.) arranged in order of decreasing number of occurrences NP. Taxa abundance rank is calculated using average percentage abundance of each taxon across samples. Community state types CST I-V and corresponding taxa average abundance in each CST are taken from Supplementary Table 5 in Ravel et al. [2011]. Missing data are coded as NA.



	NP	pvalues	NC	NCW	CST.I	CST.II	CST.III	CST.IV	CST.V
Luteococcus	207	202	154	176	0	0	0	0	0
Duganella	208	196	202	199	0	0	0	0	0
Fusobacteria	209	218	171	183	0	0	0	0	0
Bacillus_j	210	192	208	203	0.01	0	0	0	0
Scardovia	211	152	197	195	0	0	0	0	0.14
Microbacteri	212	136	246	246	0	0	0	0	0
Bacilli_2	213	169	241	241	0	0	0	0	0
Simplicispira	214	212	236	236	0	0	0	0	0
Proteobacter	215	225	193	215	0	0	0	0	0
Clostridiales_	216	235	227	230	0	0	0	0	0
Bacillus_c	217	245	224	227	0	0	0	0	0
Dermacoccus	218	246	215	222	0	0	0	0	0
Flectobacillus	219	239	239	239	0	0	0	0	0
Gordonia	220	237	233	234	0	0	0	0	0
Polaromonas	221	238	184	212	0	0	0	0	0
Selenomonas	222	241	209	220	0	0	0	0	0
Clostridiales_	223	244	217	223	0	0	0	0	0
Gp8	224	242	245	245	0	0	0	0	0
Turicella	225	243	234	235	0	0	0	0	0
Rhodoferax	226	229	244	244	0	0	0	0	0
Ruminococca	227	234	221	225	0	0	0	0	0
Flexibacterac	228	224	242	242	0	0	0	0	0
Bradyrhizobi	229	236	205	219	0	0	0	0	0
Leptotrichia	230	227	231	232	0	0	0	0	0
Firmicutes_1	231	231	210	221	0	0	0	0	0
Incertae_sed	232	232	199	217	0	0	0	0	0
Proteobacter	233	209	232	233	0	0	0	0	0
Zimmermann	234	193	223	226	0	0	0	0	0
Rhodococcus	235	170	230	231	0	0	0	0	0
Proteobacter	236	208	196	216	0	0	0	0	0
Conchiformib	237	226	218	224	0	0	0	0	0
Acidaminoco	238	240	174	211	0	0	0	0	0
Pantoea	239	222	237	237	0	0.01	0	0	0
Enterobacter	240	205	225	228	0	0	0	0	0
Bacteria_23	241	182	192	214	0	0	0	0	0
Flexibacterac	242	159	247	247	0	0	0	0	0
Paenibacillus	243	146	238	238	0	0.01	0	0	0
Capnocytoph	244	150	188	213	0	0	0	0.01	0
Skermanella	245	155	243	243	0	0	0	0	0.03
Oerskovia	246	118	240	240	0.03	0	0	0	0
Enterobacter	247	247	226	229	0	0	0	0.03	0

Figure 6: List of taxa (ctd.) arranged in order of decreasing number of occurrences NP. Taxa abundance rank is calculated using average percentage abundance of each taxon across samples. Community state types CST I-V and corresponding taxa average abundance in each CST are taken from Supplementary Table 5 in Ravel et al. [2011]. Missing data are coded as NA.

References

J. P. Brooks, D. J. Edwards, M. D. Harwich, M. C. Rivera, J. M. Fettweis, M. G. Serrano, R. A. Reris, N. U. Sheth, B. Huang, P. Girerd, a. m. Vaginal Microbiome Consortium, J. F. Strauss, K. K. Jefferson, and G. A. Buck. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, 15(66):1–14, 2015.

J. M. Fettweis, M. G. Serrano, N. U. Sheth, C. M. Mayer, A. L. Glascock, J. P. Brooks, K. K. Jefferson, a. m. Vaginal Microbiome Consortium, and G. A. Buck. Species-level classification of the vaginal microbiome. *BMC Genomics*, 13(Suppl 8):1–9, 2012.

J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, R. M. Brotman, C. C. Davis, K. Ault, L. Peralta, and L. J. Forney. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4680–4687, 2011.