BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 22 2012, pages 2891–2897
doi:10.1093/bioinformatics/bts552

Sequence analysis

Advance Access publication September 8, 2012

Comparing clustering and pre-processing in taxonomy analysis

Marc J. Bonder[1,2,*], Sanne Abeln[2], Egija Zaura[1] and Bernd W. Brandt[1,*]

[1]Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University Amsterdam and [2]Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, The Netherlands

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Massively parallel sequencing allows for rapid sequencing of large numbers of sequences in just a single run. Thus, 16S ribosomal RNA (rRNA) amplicon sequencing of complex microbial communities has become possible. The sequenced 16S rRNA fragments (reads) are clustered into operational taxonomic units and taxonomic categories are assigned. Recent reports suggest that data pre-processing should be performed before clustering. We assessed combinations of data pre-processing steps and clustering algorithms on cluster accuracy for oral microbial sequence data.

Results: The number of clusters varied up to two orders of magnitude depending on pre-processing. Pre-processing using both denoising and chimera checking resulted in a number of clusters that was closest to the number of species in the mock dataset (25 versus 15). Based on run time, purity and normalized mutual information, we could not identify a single best clustering algorithm. The differences in clustering accuracy among the algorithms after the same pre-processing were minor compared with the differences in accuracy among different pre-processing steps.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: bonder.m.j@gmail.com or b.brandt@acta.nl

Received on July 6, 2012; revised on August 27, 2012; accepted on September 1, 2012

1 INTRODUCTION

Recent developments in massively parallel sequencing allow for rapid characterization of a large number of sequences in a single run. 16S ribosomal RNA (rRNA) amplicon sequencing can reveal the composition of complex microbial communities (Schuster, 2008). The unprecedented sequencing depth means that even less abundant species can be detected via their sequences. The taxonomic composition of such samples may be characterized by grouping the sequence fragments (reads) into clusters of operational taxonomic units (OTUs).

For many, often uncultivable, microbes, the 16S rRNA sequences are unknown. To avoid the dependency on incomplete databases (Cole *et al.*, 2005; Sun *et al.*, 2012), microbial community composition is typically determined by using a taxonomy-independent analysis (TIA). In this approach, the 16S rRNA sequences are compared against each other without directly assigning a taxonomy to the sequence; the number and size of the clusters (OTUs) indicate the diversity of the sample. Recent benchmark studies have revealed several clustering programs that adequately cluster into OTUs (Huse *et al.*, 2010; Schloss and Westcott, 2011; Sun *et al.*, 2012). For this study, we selected all commonly used clustering approaches implemented in Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso *et al.*, 2010), and we also included ESPRIT-Tree as it performed well in a recent comparison of clustering methods (Sun *et al.*, 2012) and as well as methods that include rRNA secondary structure information (Wang *et al.*, 2012).

Recent reports suggest that before the clustering step several data cleaning, or pre-processing, steps should be performed to correct errors introduced during (pyro)sequencing (Edgar *et al.*, 2011; Haas *et al.*, 2011; Kunin *et al.*, 2010; Quince *et al.*, 2011; Reeder and Knight, 2010; Schloss *et al.*, 2011). The two most important pre-processing steps are denoising, which corrects errors within the reads based on raw sequencing output (flowgrams), and chimera checking, which removes chimeric sequences. The aim of this study is to identify suitable combinations of pre-processing steps and clustering methods for oral microbial sequence data.

We used two sets of pyrosequenced reads of 16S rRNA to evaluate cluster accuracies. The first set, a mock dataset, had a known species composition (15 species; Supplementary Table 1). The second set, a clinical dataset, has reads from samples of human saliva collected in a study on the influence of *Candida* on the composition of the oral microbiome (Kraneveld *et al.*, 2012).

Assessing the performance of the pre-processing and clustering combinations is non-trivial, since no gold standard is available for the taxonomic composition of the samples. In this work, we use two strategies to assess this performance. First, we take a simple approach and check whether the number of clusters agrees with the expected number of species in the mock sample. The taxonomic composition of the clinical sample may be approximated by assigning a taxonomic label from a reference rRNA database to each read.

Second, we assess the clusters based on labelling of sequences. Therefore, we calculate the purity and normalized mutual information (NMI) scores (Press *et al.*, 2007). Purity scores reflect the homogeneity of a cluster. If the majority of the labels in each cluster are the same, the purity score is high. However, the purity score does not penalize for the creation of multiple clusters that contain the same label(s). The NMI score, in addition, reflects the mutual information between the clusters and the taxonomic labels; a high NMI score is achieved when high purity is reached with a minimum number of clusters.

*To whom correspondence should be addressed.

2 METHODS

2.1 Data

2.1.1 Datasets We used a mock and clinical dataset. Both sets were pyrosequenced (Margulies et al., 2005). Amplification was performed using 16S rRNA PCR primers targeting the V5–V7 hypervariable region: forward primer 785F (GGATTAGATACCCBRGTAGTC) and reverse primer 1175R (ACGTCRTCCCCDCCTTCCTC). The primers included the 454 Life Sciences (Branford, CT, USA) Adapter A (for forward primers) and B (for reverse primers) fused to the 5′ end of the 16S rDNA bacterial primer sequence and a unique 10 nucleotide molecular identification code.

From the clinical dataset, all samples with at least 1250 reads and at most 20 000 reads were selected. These thresholds were chosen to exclude samples with possible errors in sample preparation and over- or under amplification before sequencing.

2.1.2 Reference database A 16S rRNA reference is necessary for chimera checking (UCHIME), clustering (QIIME BLAST, UCLUST reference) and for evaluating clustering results. This reference database was produced by an *in silico* amplification of the SILVA SSU 16S rRNA reference database, version 108 (Pruesse et al., 2007). A single nucleotide mismatch was allowed for each primer and the longest amplicon was taken in case multiple amplicons were formed from one sequence. The resulting amplicon sequences were made non-redundant. If identical sequences had different taxonomic lineages, the longest shared lineage was used. A suffix was added if multiple different sequences had exactly the same lineage.

2.2 Read preparation and pre-processing

QIIME (Caporaso *et al.*, 2010) was used to pre-process the reads. Default settings were used, except for the following: minimum sequence length: 150; sliding window size: 50; reverse primer removal: truncate_only. The reverse primer was removed when it was found. Reads with one or more mismatch to the forward primer were discarded.

Next, the two sets were pre-processed in four different ways: (i) no cleaning (NC), (ii) chimera checked (CC), (iii) denoised (D) and (iv) denoised and chimera checked (DCC). This resulted in eight different test sets for further analysis.

The reads were denoised with denoiser with default settings for denoising 454 GS FLX Titanium (Reeder and Knight, 2010), which is now part of QIIME (Caporaso *et al.*, 2010). For chimera checking, we used UCHIME (Edgar *et al.*, 2011) in both the reference and the *de novo* mode. If a read was marked as chimeric in either of the two methods, the read was filtered out. We used our non-redundant SILVA database (see above) as the reference for UCHIME.

2.3 Clustering algorithms

The datasets, pre-processed in the four different ways, were all clustered into OTUs with five different clustering approaches: UCLUST version 4.2.40 (Edgar, 2010), mothur clustering version 1.6.0 (Schloss *et al.*, 2009), ESPRIT-Tree beta release September 2011 (Cai and Sun, 2011), CD-HIT version 3.1.1 (Li and Godzik, 2006) and QIIME BLAST clustering (QIIME 1.3.0 and BLAST 2.2.22) (Altschul *et al.*, 1990). The QIIME BLAST clustering approach matches the reads to the closest sequence in the database and groups the reads based on the BLAST label. For UCLUST, we used both the *de novo* mode and the reference mode. For the 'denoised and chimera checked' set, the UCLUST reference optimal mode was also tested. This method often gives the same output as UCLUST reference but has a much longer run time (Edgar, 2010). For mothur, both average and complete linkages were tested. Furthest linkage is the default, but in a recent study (Huse *et al.*, 2010), it was concluded that average linkage outperforms furthest linkage. For both BLAST and UCLUST in reference mode, we used our primer-specific non-redundant SILVA rRNA database as reference. Except for the distance, which we set to 97% sequence identity, we used default settings for the clustering algorithms as these were found to be optimal by the authors of the different (respective) clustering methods.

2.4 Clustering accuracy

To determine the accuracy of clustering, we labelled the reads with a taxonomic lineage by BLASTing (Altschul *et al.*, 1997) (BLAST 2.2.25+) all reads to our (trimmed) SILVA reference database (95% identity threshold). It is not trivial to assign a unique taxonomic label to a read, since a single read sometimes can be mapped to two labels with equal BLAST scores (this can also occur at 100% identity and 100% coverage). To solve this problem, we use either a consensus or a unique labelling approach. In the consensus approach, the labels were merged to the highest shared taxonomic category when a query had two or more hits with equal score. The suffix, added during reference database - creation, was removed. For instance, when a read is labelled with: '*Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae_21*' and '*Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus_2*', then the label of the sequence in the consensus treatment becomes: '*Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae*'.

In the unique approach, all the labels remain unique: if a query had two or more hits with equal score, they were combined into one header by adding a separation mark. The headers were always written in alphabetical order. The appended number, which made the header unique, was not removed in the unique treatment. For instance, if a read is assigned the same labels as the previous example, the unique treatment would result in this label: '*Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus_2_/_Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae_21*'.

Clustering accuracy was evaluated with cluster purity (Press *et al.*, 2007) and the NMI (Press *et al.*, 2007). For both scoring functions, we discarded the reads that did not have a label assigned after BLASTing as there is no obvious way to evaluate clusters, including reads with 'unknown' labels. Cluster purity is calculated as:

$$\text{purity}(\Omega, C) = \frac{1}{N}\sum_{k}\max_{j}|\omega_k \cap c_j| \quad (1)$$

where $\Omega = \{w_1, w_2, \ldots, w_k\}$ is the set of clusters, $C = \{c_1, c_2, \ldots, c_j\}$ is the set of classes as defined by the taxonomic labels and N is the total number of sequences. As a second scoring function, the NMI was used. The NMI is calculated as:

$$\text{NMI}(\Omega, C) = \frac{I(\Omega, C)}{[H(\Omega) + H(C)]/2} \quad (2)$$

where $I(\Omega,C)$ is the mutual information gain, $H(X)$ is the entropy.

The NMI and purity were calculated using the lineage returned by BLAST for both unique and consensus treatments. Depending on the BLAST results, the lineages can be up to any taxonomic level.

3 RESULTS

The accuracy of several clustering algorithms was assessed using two test sets of sequence reads: a mock dataset and a clinical dataset. The datasets were pre-processed in four ways: (i) no cleaning (NC), (ii) chimera checked (CC), (iii) denoised (D) and (iv) denoised and chimera checked (DCC), for further details see Methods section. The resulting sets were all clustered using five different clustering algorithms: UCLUST, mothur clustering, ESPRIT-Tree, CD-HIT and QIIME BLAST.

The clustering results were evaluated in two ways: a consensus and a unique method (see Methods section). The consensus method merges the assigned lineages to a highest shared taxonomic level, while the unique method only merges the labels. Here, the number of unique labels in the unique set is larger than in the consensus set (Table 1). The clustering accuracy was measured with the purity and the NMI. Below, the results of the clustering evaluation on the mock and clinical datasets are presented.

3.1 Mock dataset

The mock community test set contained 15 oral microbial species (Supplementary Table 1). All pre-processing steps affected the number of reads and unique reads (Table 1).

First, we compared the total run time of the pre-processing and clustering algorithms combined (Table 2). The QIIME BLAST clustering method required the most time. The fastest QIIME BLAST clustering, on 'denoised and chimera checked' data, was around 118 h slower than the other clustering methods on the same dataset.

Next, the number of clusters formed was compared (Table 3). Ideally, in a TIA, the number of species is identical to the number of clusters. However, this was not the case. Even the lowest number of clusters (25) was 66% higher than expected (15), based on the number of species in the mock. Furthermore, the effect of denoising is much larger than chimera checking on both the number of clusters and the number of unique reads (Tables 1 and 3).

Since the composition of the mock set is known, we also tested whether all taxa present in the mock could be identified (using the consensus taxonomic labels). It was not always possible to unambiguously assign a taxonomic lineage, including the species level, by BLASTing the reads against our reference database. However, this was mainly caused by the extensive reference database and not by the reads themselves. For most species, even *in silico* produced amplicons match more than two species in our reference database (with the same score and at 100% identity). For example, the *Streptococcus mutans* amplicon is 100% identical to *Streptococcus uncultured Streptococcus* sp. and *Streptococcus uncultured bacterium*, allowing only the assignment of the genus name. Most importantly, all 10 genera present in the mock could be found after any pre-processing of the data. Next, we will evaluate the different clustering methods and pre-processing steps based on consensus and unique labelling of the reads.

3.1.1 Consensus labelling of reads The cluster purities depend on the specific clustering algorithm and on the pre-processing approach (Fig. 1A). The clusters based on 'denoised and chimera checked' data had the highest purity. The clustering methods QIIME BLAST, 'UCLUST reference' and 'UCLUST reference optimal' resulted in a cluster purity of one while the others were very close to one. The purity showed a larger variation for the other pre-processing steps, especially when the data was not denoised (Fig. 1A). It should be noted that high purity is easier to acquire when the number of clusters is high. Indeed, a positive correlation between the number of clusters and the purity was observed. However, for mothur clustering with average linkage, this was not the case. The NMI shows a similar trend (Fig. 1B): the 'denoised and chimera checked' data again showed superior scores. CD-HIT, ESPRIT-Tree and UCLUST were the best performing clustering algorithms on this dataset with a score

Table 1. Comparison of the number of total reads, non-redundant (unique) reads, removed reads, number of reads without a label and number of labels for each pre-processing method on the mock set

Pre-processing	No. of reads	No. of unique reads	No. of removed reads	No. of reads labelled as unknown	No. of unique labels	No. of consensus labels
NC	38612	11618	0	371	637	106
CC	33375	8431	5237	15	390	61
D	38612	190	0	192	85	42
DCC	34885	45	3727	18	29	23

NC stands for no cleaning, CC stands for chimera checked, D stands for denoised and DCC stands for denoised and chimera checked.

Table 2. Comparison of the total run time of the clustering algorithms and pre-processing methods on the mock dataset

Pre-processing	Clustering algorithms							
	QIIME BLAST	CD-HIT	ESPRIT-Tree	Mothur furthest	Mothur average	UCLUST	UCLUST ref	UCLUST ref optimal
NC	141:49:27	0:18:00	0:15:48	2:44:30	3:19:13	0:12:59	0:14:53	
CC	138:53:15	14:56:23	14:57:23	15:54:17	16:06:13	14:54:51	14:56:06	
D	140:28:04	1:44:09	1:42:55	1:42:54	1:42:55	1:42:55	1:43:21	
DCC	120:09:04	1:56:46	1:56:17	1:56:17	1:56:18	1:56:18	1:56:41	1:59:03

Run time is given in hours:minutes:seconds.

Table 3. Comparison of the number of clusters formed for each clustering algorithm and pre-processing method

Pre-processing	Clustering algorithms							
	QIIME BLAST	CD-HIT	ESPRIT-Tree	Mothur furthest	Mothur average	UCLUST	UCLUST ref	UCLUST ref optimal
NC	562	485	328	3286	2236	379	651	
CC	288	131	66	2107	1373	77	323	
D	95	112	104	155	154	104	124	
DCC	30	25	25	38	38	25	31	31

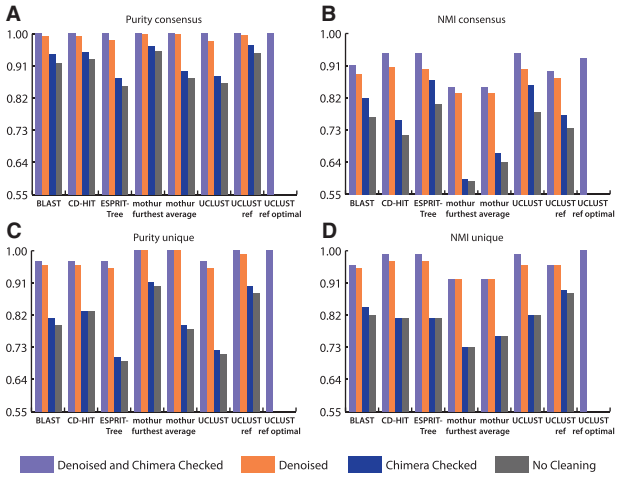

Fig. 1. The purity and NMI scores for the consensus and unique treatments on the mock dataset. (**A**) Consensus purity, (**B**) consensus NMI, (**C**) unique purity and (**D**) unique NMI

of 0.94. All results for the consensus read labelling are shown in Supplementary Table 2.

Among the different pre-processing steps, the'denoised and chimera checked' pre-processing showed the best results. Based on purity, NMI scores and run time, CD-HIT, ESPRIT-Tree and UCLUST performed equally well after this pre-processing (NMI of 0.94, purity of almost 1 and 25 clusters).

3.1.2 Unique labelling of reads As in the consensus read labelling, the average purity and NMI was highest for the 'denoised and chimera checked' pre-processing (Fig. 1). UCLUST in reference modes and mothur resulted in the best purity scores (about one). However, both the NMI scores and the number of clusters showed that mothur formed too many clusters. Out of the non-reference based cluster methods, CD-HIT, ESPRIT-Tree and UCLUST perform similarly well based on purity, NMI and number of clusters.

The average cluster purity was generally lower in the unique than in the consensus read labelling. The average NMI scores, for all pre-processing steps, were higher in the unique than in the consensus read labelling. Both are consequences of the higher number of unique labels in the unique set, which is due to merging the labels differently. The complete results of the unique read labelling are shown in Supplementary Table 3.

3.2 Clinical dataset

The same evaluations of clustering as above were performed on the clinical dataset. For the two most time-consuming methods, namely mothur and QIIME BLAST, we made the reads non-redundant before clustering; after clustering, the reads were reinflated to the input sample size. The time it took to make the reads unique and to inflate was included in the total run time of mothur and QIIME BLAST OTU picking methods. In addition, a distance cutoff (5% for furthest linkage and 15% for average linkage) was used during the distance matrix calculation in mothur, such that this matrix still fits in memory. In both cases, clustering was performed at 97% sequence identity. For mothur average, the matrix was still too large for the NC and CC set.

The four different pre-processing steps resulted in different numbers of (unique) reads (Table 4). Denoising was the most time-consuming step (~552 CPU hours). The QIIME BLAST clustering method required the most time while only clustering the unique reads. The run times of the pre-processing procedures and clustering methods are shown in Table 5.

Since the species composition of clinical data is unknown, the correct number of clusters is unknown. However, there was a large difference in the number of clusters formed, especially among the pre-processing methods (Table 6). The number of unique reads decreased enormously after denoising or chimera checking, but especially after doing both. When one compares the decrease in number of unique reads per clusters between the mock and the clinical set after each cleaning step, a similar trend can be observed. Roughly the same decrease in number of clusters per unique read is observed.

3.2.1 Consensus labelling of reads The cluster purity for all the clustering algorithms and the different pre-processing approaches are shown in Figure 2A and B. The 'denoised and chimera checked' data had overall the highest purity and NMI scores, as was the case for the mock set. The clustering methods 'UCLUST reference' and 'UCLUST reference optimal' had the highest purity (about one). Out of the non-reference based cluster methods, ESPRIT-Tree and UCLUST perform similarly well when combining purity, NMI scores and number of clusters. The complete consensus results are in Supplementary Table 4.

3.2.2 Unique labelling of reads As with the consensus treatment, generally, the purity and NMI scores were highest for the 'denoised and chimera checked' pre-processing

Table 4. Comparison of the number of total reads, non-redundant (unique) reads, removed reads, number of reads without a label and number of labels for each pre-processing method on the clinical set

Pre-processing	No. of reads	No. of unique reads	No. of removed reads	No. of reads labelled as unknown	No. of unique labels	No. of consensus labels
NC	496 149	110 508	0	11 449	2754	854
CC	391 136	61 571	105 013	731	1850	62
D	496 148	2933	1	7702	1109	33
DCC	418 469	591	77 680	517	356	200

Table 5. Comparison of the total run time of the clustering algorithms and pre-processing methods on the clinical dataset

Pre-procesing	Clustering algorithms							
	QIIME BLAST	CD-HIT	ESPRIT-Tree	Mothur furthest	Mothur average	UCLUST	UCLUST ref	UCLUST ref optimal
NC	426:49:07	13:41:22	4:51:37	19:01:07	*	2:51:17	2:46:40	
CC	407:55:49	179:14:54	176:09:02	181:12:24	*	175:23:37	175:25:18	
D	566:02:45	555:59:20	555:06:41	555:05:36	555:05:34	555:05:15	555:06:39	
DCC	560:17:08	558:37:20	558:18:45	558:18:08	558:18:04	558:18:04	558:18:35	560:22:45

Run time is given in hours:minutes:seconds. *Mothur average could not be used for the NC and CC set.

Table 6. Comparison of the number of clusters formed for each clustering algorithm and pre-processing method

Pre-processing	Clustering algorithms							
	QIIME BLAST	CD-HIT	ESPRIT-Tree	Mothur furthest	Mothur average	UCLUST	UCLUST ref	UCLUST ref optimal
NC	3915	8985	7299	28711	*	7707	9517	
CC	1874	1585	1018	11932	*	1091	2784	
D	1048	2031	1941	2364	2314	1971	2112	
DCC	309	332	306	458	453	318	421	425

*Mothur average could not be used for the NC and CC set.

(Figure 2C and D). Interestingly, for CD-HIT, the NMI was (marginally) higher for the 'denoised' data than for the 'denoised and chimera checked' data. 'UCLUST reference optimal' and 'UCLUST reference' performed the best with the purity very close to one. Based on the NMI, the highest scoring algorithms were 'UCLUST reference optimal' and 'UCLUST reference' (NMI very close to one). The complete results of the unique read labelling are shown in Supplementary Table 5.

4 DISCUSSION

We compared different pre-processing and clustering methods with respect to the clustering of sequences into OTUs. The different pre-processing methods showed clear differences in the number of clusters that were formed.

The 'denoised and chimera checked' data formed the number of clusters that was closest to the number of species in the mock. The cluster algorithms formed minimally 25 clusters (Table 3), whereas 15 species were present in the mock community.

In the clinical dataset, we observed around 300–450 clusters after denoising and chimera checking. When using only one of the two pre-processing methods, 1000–3000 clusters were formed. Without any pre-processing, the number of clusters reached beyond 7000. As with the mock community data, in the clinical dataset clustering the 'denoised and chimera checked' data performed best if the NMI and purity scores were used as the quality criterion. The actual number of valid clusters in the clinical dataset is unknown. Based on cloning and traditional Sanger sequencing, currently, there are 600–1200 species reported that are associated with the human oral microbiome (Dewhirst *et al.*, 2010; Paster *et al.*, 2001). However, the oral microbiome diversity is estimated to be at least an order of magnitude higher based on data from high-throughput sequencing approaches (Keijser *et al.*, 2008; Yang *et al.*, 2012). Considering the decreasing number of clusters in the mock set in relation to increasing

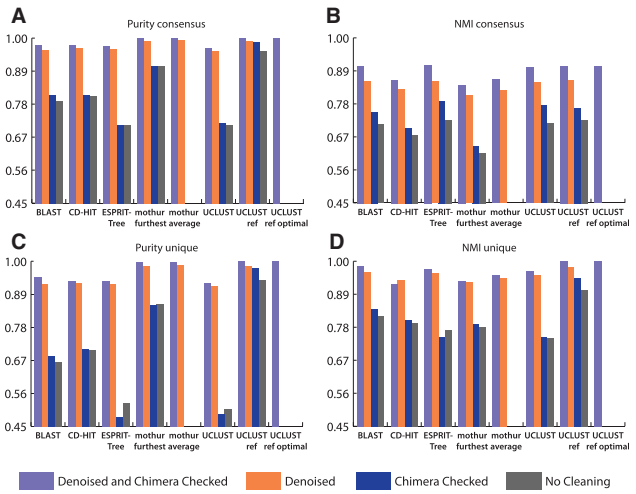

Fig. 2. The purity and NMI scores for the consensus and unique treatments on the clinical dataset. Mothur average could not be used for the NC and CC set. (**A**) Consensus purity, (**B**) consensus NMI, (**C**) unique purity and (**D**) unique NMI

read pre-processing and a similar trend in the reduction of number of clusters in the clinical set, it is likely that the previous estimates are at the higher end of the actual diversity. We observe the same overestimation in microbial diversity as the authors of pre-clustering (Huse *et al*., 2010) and denoising algorithms (Quince *et al*., 2011; Reeder and Knight, 2010).

The differences in the NMI and purity scores for the clustering methods are minor when compared with the differences for the pre-processing methods. The UCLUST reference-based methods generally outperform the other methods based on the NMI score, while the purity scores were also high. Both suggest that the UCLUST reference methods are the methods of choice. However, it is important to note that these methods form a larger number of clusters than expected and larger than the minimal number of clusters found and use the same reference database on which the taxonomic labels are based, making such a statement too bold. In general, UCLUST and ESPRIT-Tree came out top among the non-reference based methods. They showed a similar performance based on purity and NMI scores, while giving the lowest number of clusters.

A reference database is required for chimera checking in reference mode and for UCLUST in the reference modes. Here, we used a trimmed and non-redundant reference, which greatly reduces run times (Brandt *et al*., 2012). In this study, we did not use the trimmed set for training a classifier, but it has been found that this improves taxonomic classification (Werner *et al*., 2012).

The majority of the mock reads (99.8%) has been labelled with expected lineages. Although the mock contains 15 species, we do not expect 15 clusters. The V5–V7 regions of *Streptococcus mitis* and *Streptococcus oralis* are almost identical (99 - 100%). Thus, the expected number of clusters is (at least) 14. The different clustering algorithms returned 25 clusters at best. However, 11 of these 25 clusters have very few members (four are even singletons). If we use a relaxed threshold on cluster size of only >18 members, indeed, the 14 expected clusters remain. Given the relatively large number of very small clusters, we do not expect this to generalize to a clinical sample. While a threshold on cluster size is more difficult to determine for a clinical sample, very small clusters (e.g. with less than five members) could also be disregarded here (Özok *et al*., 2012), though at the risk of underestimating diversity.

The determination of the microbiome, based on a few hypervariable regions of the 16S rRNA gene, makes it difficult to differentiate between several taxa on species level or even genus level. Furthermore, the sequencer creates errors (noise), which makes it even more difficult or impossible to differentiate between related species, such as *Streptococcus* species. In such cases, denoising, to filter out errors created by the sequencer, may regard the small differences between sequences as noise. Thus, pyrosequencing (a region of) 16S rRNA can distinguish between different genera or species, possibly in combination with denoising, only when sequences of the related taxa differ enough.

It is important to note that denoising, chimera checking and clustering methods are context dependent: the result for a single read depends on the entire input set. This is specifically relevant for denoising of data from large studies, since it is typically done on several input files, like in this study. For example, similar sequences in one file can be denoised to Sequence A, while in another be denoised to Sequence B, simply because Sequence A is more dominant in the one file and Sequence B in the other.

There was a remarkable difference in the number of reads discarded by chimera checking before and after denoising. On average, there was a 27% decrease in the number of non-unique reads when denoising was performed first. Moreover, denoising, which can also be seen as a pre-clustering step, has a huge influence on the number of clusters generated. However, denoising differs from the (other) clustering methods as it takes flowgram data into account. This potentially allows denoising to differentiate between true sequence differences and sequencing noise. The results of this study and other studies (Kunin *et al*., 2010; Schloss *et al*., 2011; Schloss and Westcott, 2011) suggest that it remains important to further investigate combinations of pre-processing steps.

We find that the number of OTUs varies between one and two orders of magnitude depending pre-processing methods and clustering algorithms. Using a single-species reference (*Escherichia coli*), it was found that stringent filtering and low clustering thresholds should be applied to prevent overestimation of diversity (Kunin *et al*., 2010): unfiltered reads overestimated the number of OTUs by two orders of magnitude (Kunin *et al*., 2010). With respect to technical variation, pre-processing and reduction of error rate, the thorough work of Schloss *et al*. (2011) on mock communities is interesting. They also report that filtering (sliding window quality filter or denoising combined with chimera removal) is required to reduce the number of OTUs closer to the number of expected OTUs. Even after a 30-fold reduction in sequencing error rate, the number of expected OTUs and genera was not obtained (Schloss *et al*., 2011). More recently, it was also recommended to remove pyrosequencing errors and chimeras before clustering (Jiang *et al*., 2012).

While we could not identify a single best clustering algorithm, CD-HIT, ESPRIT-Tree and UCLUST perform well. In another study, UCLUST also performs well and comparable to CD-HIT but is outperformed by an average neighbour algorithm (Schloss and Westcott, 2011). ESPRIT-Tree, an average linkage-based

hierarchical clustering algorithm, was found to outperform CD-HIT, UCLUST and mothur (average linkage) in another benchmark study (Sun *et al*., 2012).

In summary, we tested four pre-processing methods and multiple clustering algorithms to cluster sequences into OTUs. By combining these steps, we assessed the influence of read filtering on clustering. We showed that cleaning influences the number of OTUs much more when compared with the different clustering methods. The pre-processing method that resulted in the highest NMI, the highest purity and the number of clusters closest to the expected number of clusters was the combination of denoising with chimera checking.

Our results give strong evidence that without pre-processing steps the data contains too many errors in order for any clustering algorithm to perform well. This warrants further investigation in pre-processing techniques that allow for the correction of such errors, as more accurate pre-processing will have a larger effect than improving current clustering techniques. Alternatively, integrated approaches may be investigated.

ACKNOWLEDGEMENTS

The authors thank Eefje Kraneveld, Mark Buijs and Jessica Koopman for providing the sequence datasets.

Funding: University of Amsterdam under the research priority area 'Oral Infections and Inflammation' and the Netherlands Organisation for Scientific Research (NWO) (to S.A.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

Brandt,B.W. *et al*. (2012) TaxMan: a server to trim rRNA reference databases and inspect taxonomic coverage. *Nucleic Acids Res*., **40**, W82–W87.

Cai,Y. and Sun,Y. (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res*., **39**, e95.

Caporaso,J.G. *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Cole,J.R. *et al*. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*., **33**, D294–D296.

Dewhirst,F.E. *et al*. (2010) The human oral microbiome. *J. Bacteriol*., **192**, 5002–5017.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edgar,R.C. *et al*. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Haas,B.J. *et al*. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*., **21**, 494–504.

Huse,S.M. *et al*. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol*., **12**, 1889–1898.

Jiang,X.T. *et al*. (2012) Two-stage clustering (TSC): a pipeline for selecting operational taxonomic units for the high-throughput sequencing of PCR amplicons. *PLoS ONE*, **7**, e30230.

Keijser,B.J.F. *et al*. (2008) Pyrosequencing analysis of the oral microflora of healthy adults. *J. Dent. Res*., **87**, 1016–1020.

Kraneveld,E.A. *et al*. (2012) The relation between oral *Candida* load and bacterial microbiome profiles in Dutch older adults. *PLoS ONE*, **7**, e42770.

Kunin,V. *et al*. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol*., **12**, 118–123.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Margulies,M. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Özok,A.R. *et al*. (2012) Ecology of the microbiome of the infected root canal system: a comparison between apical and coronal root segments. *Int. Endod. J*., **45**, 530–541.

Paster,B.J. *et al*. (2001) Bacterial diversity in human subgingival plaque. *J. Bacteriol*., **183**, 3770–3783.

Press,W.H. *et al*. (2007) *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, NY, USA.

Pruesse,E. *et al*. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*., **35**, 7188–7196.

Quince,C. *et al*. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Reeder,J. and Knight,R. (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat. Methods*, **7**, 668–669.

Schloss,P.D. *et al*. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE*, **6**, e27310.

Schloss,P.D. and Westcott,S.L. (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol*., **77**, 3219–3226.

Schloss,P.D. *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol*., **75**, 7537–7541.

Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.

Sun,Y. *et al*. (2012) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform*., **13**, 107–121.

Wang,X. *et al*. (2012) Secondary structure information does not improve OTU assignment for partial 16s rRNA sequences. *ISME J*., **6**, 1277–1280.

Werner,J.J. *et al*. (2012) Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J*., **6**, 94–103.

Yang,F. *et al*. (2012) Saliva microbiomes distinguish caries-active from healthy human populations. *ISME J*., **6**, 1–10.