BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1530–1538
doi:10.1093/bioinformatics/btu085

Sequence analysis

Advance Access publication February 10, 2014

Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations

Ali May[1,2,*], Sanne Abeln[2,3], Wim Crielaard[1], Jaap Heringa[2,3,4] and Bernd W. Brandt[1,*]

[1]Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University Amsterdam, Amsterdam, The Netherlands, [2]Centre for Integrative Bioinformatics VU and [3]AIMMS Amsterdam Institute for Molecules Medicines and Systems, VU University Amsterdam, Amsterdam, The Netherlands and [4]NBIC Netherlands Bioinformatics Centre, Nijmegen, The Netherlands

Associate Editor: Janet Kelso

ABSTRACT

Motivation: 16S rDNA pyrosequencing is a powerful approach that requires extensive usage of computational methods for delineating microbial compositions. Previously, it was shown that outcomes of studies relying on this approach vastly depend on the choice of pre-processing and clustering algorithms used. However, obtaining insights into the effects and accuracy of these algorithms is challenging due to difficulties in generating samples of known composition with high enough diversity. Here, we use *in silico* microbial datasets to better understand how the experimental data are transformed into taxonomic clusters by computational methods.

Results: We were able to qualitatively replicate the raw experimental pyrosequencing data after rigorously adjusting existing simulation software. This allowed us to simulate datasets of real-life complexity, which we used to assess the influence and performance of two widely used pre-processing methods along with 11 clustering algorithms. We show that the choice, order and mode of the pre-processing methods have a larger impact on the accuracy of the clustering pipeline than the clustering methods themselves. Without pre-processing, the difference between the performances of clustering methods is large. Depending on the clustering algorithm, the most optimal analysis pipeline resulted in significant underestimations of the expected number of clusters (minimum: 3.4%; maximum: 13.6%), allowing us to make quantitative estimations of the bacterial complexity of real microbiome samples.

Contact: a.may@vu.nl or b.brandt@acta.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online. The simulated datasets are available via http://www.ibi.vu.nl/downloads.

Received on November 29, 2013; revised on January 24, 2014; accepted on February 4, 2014

1 INTRODUCTION

Rapidly improving next-generation sequencing technologies have rendered 16S rDNA amplicon sequencing a standard approach to explore the composition and dynamics of complex microbial communities (Woo *et al.*, 2008). The accurate identification of bacteria and discovery of new species with this approach help describe, in unprecedented detail, the hitherto unknown biodiversity in environments such as ocean water (Kirchman *et al.*, 2010) and soil (Roesch *et al.*, 2007). On a broader scale, simultaneous sequencing of internal transcribed spacer regions, 16S and 18S rDNA amplicons has led to the successful characterization of bacterial, archaeal and eukaryotic microorganisms (Kittelmann *et al.*, 2013; Somboonna *et al.*, 2012). Within a clinical context, this technique provides novel insights into the relationship between the health status of the host and associated microbiome, for example, the human gut (Le Chatelier *et al.*, 2013) or oral cavity (Crielaard *et al.*, 2011).

A typical 16S rDNA-based profiling study involves PCR amplification of one or more hypervariable regions (HVRs) of microbial 16S rRNA genes, followed by massively parallel sequencing. Because incomplete 16S rDNA databases limit reference-based taxonomic classification (Sun *et al.*, 2012), a common approach is to perform a taxonomy-independent analysis (TIA), where the sequences (reads) are compared with each other and are clustered into operational taxonomic units (OTUs; conventionally 97% similarity for species-level clustering). The number and sizes of the resulting OTUs describe the diversity in the sample.

Among next-generation sequencing platforms that are suitable for amplicon sequencing, 454 pyrosequencing has possibly been the most popular technique due to its relatively long reads and sufficient sampling depth (Siqueira *et al.*, 2012; Tamaki *et al.*, 2011). Nevertheless, sample diversity and structure estimates based on this procedure are affected by a number of factors, including the DNA extraction method (Sergeant *et al.*, 2012), choice of primers (Klindworth *et al.*, 2013), copy number variation (Kembel *et al.*, 2012), presence of chimeric (Haas *et al.*, 2011) or non-target sequences (Hartmann *et al.*, 2010), sequencing errors (Kunin *et al.*, 2010), and OTU clustering (Huse *et al.*, 2010). Among these, detecting chimeric sequences (Edgar *et al.*, 2011) and correcting sequencing errors (Quince *et al.*, 2011; Reeder and Knight, 2010) have perhaps been the most successfully addressed issues by computational means, resulting in a number of algorithms for data pre-processing that incorporate these corrective steps. The removal of such artifacts is key to avoiding accumulation of spurious records in databases (Hugenholtz and Huber, 2003). More importantly, recent studies stress the necessity of data pre-treatment by pointing out that diversity estimates differ orders of magnitude depending on the pre-processing procedure used (Bonder *et al.*, 2012; Schloss *et al.*, 2011). However, conclusive reports on how these methods work

*To whom correspondence should be addressed.

on a representative scale are limited due to the absence of ground truth in complex environmental datasets.

Following pre-processing, accurate clustering of reads into OTUs facilitates the recovery of the community composition in the sample. However, numerous clustering methods used for TIA, such as single linkage or furthest linkage hierarchical clustering (Huse *et al.*, 2010), have been reported to yield considerably different results, and even small changes in algorithm parameters within individual methods can lead to vast differences (White *et al.*, 2010). In addition, differences in clustering algorithms, such as different criteria for the ordering of input sequences or the calculation of distance matrices, are likely to yield different clustering outcomes.

The evaluation of the various data treatment and clustering methods and their use on real datasets have hitherto been limited to *in vitro* mock studies, where sample diversities are typically (at least) an order of magnitude less compared with those in real-life situations. Such limitations of *in vitro* studies have led to the development of simulation software, such as MetaSim (Richter *et al.*, 2008) and Grinder (Angly *et al.*, 2012), which are capable of generating complex sequence datasets. These packages, which do not produce the 454-specific standard flowgram format, can be complemented by FlowSim (Balzer *et al.*, 2010) or 454Sim (Lysholm *et al.*, 2011) to simulate flowgrams (the pyrosequencing output). Nevertheless, some aspects of 16S rDNA pyrosequencing are not amenable to implementation in simulations or cannot be incorporated at all. These include PCR efficiency, primer bias, 16S rDNA copy number variation and realistic read quality scores.

Here, using *in silico* datasets, we aimed to provide an independent assessment of the accuracy and effects of computational techniques on delineating compositions of complex microbial communities using 16S rDNA pyrosequencing. We analyzed two low-complexity experimental datasets and derived data characteristics, which were then used to simulate datasets that are likely to mimic real-life complexity. We applied two widely used data pre-processing tools followed by 11 clustering algorithms on these datasets, and evaluated the accuracy of the resulting clusters using the normalized mutual information (NMI) score and the number of OTUs formed. The overall experimental design is shown in Figure 1. We show that pre-processing methods have a larger impact on the results of the TIA pipeline than the clustering methods themselves. Without pre-processing, clustering methods were found to yield considerably different outcomes. The workflow that yielded the best agreement with the expected number of clusters resulted in significant underestimations of the expected number of OTUs, depending on the clustering algorithm used. We conclude with a discussion on how these new findings can be projected to earlier studies that describe the microbial complexity in human saliva samples.

2 METHODS

2.1 *In vitro* datasets

We studied two GS-FLX Titanium datasets, hereafter referred to as Mock-I and Mock-II, from a mock community of 15 oral bacteria. For both sequencing runs, the V5–V7 HVR of the 16S rDNA was amplified using the forward primer 785 F (GGATTAGATACCCBRGTAGTC) and the reverse primer 1175R (ACGTCRTCCCCDCCTTCCTC)

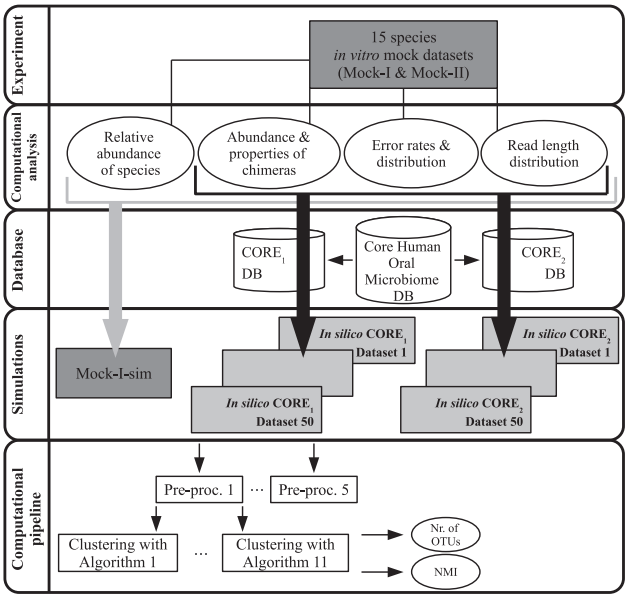

Fig. 1. Flow diagram of the study design. Two *in vitro* datasets were analyzed and the resulting statistics were used in simulations to replicate one of these datasets (gray), as well as to generate datasets of real-life complexity (black). Different pre-processing approaches and clustering algorithms were evaluated in a combinatorial way by calculating the NMI score and the number of clusters formed. The computational pipeline applies to $CORE_1$ and $CORE_2$ datasets

(Kraneveld *et al.*, 2012). The reference 16S rDNA sequences were downloaded from the SILVA SSU 16S rRNA database v111 (Quast *et al.*, 2013). The presence of three different 16S rDNA copies in the database for one of the 15 species led to the retrieval of 17 non-redundant reference sequences.

2.2 Analysis of *in vitro* mock datasets

2.2.1 Quality filtering The reads were demultiplexed using QIIME v1.5.0 (Caporaso *et al.*, 2010) with the 'truncate_remove' option for reverse primer removal. Two mismatches to both primers and zero mismatches to the barcode were allowed. No filtering was performed based on quality scores, read length, homopolymer region length or the number of ambiguous bases. The read length distribution was calculated by demultiplexing without primer removal.

2.2.2 Chimera detection After quality filtering, the set of true chimeras was approximated by identifying a set of high-confidence chimeras, as previously described by Schloss *et al.* (2011), using the seq.error (hereafter referred to as SEQERR) routine in mothur v1.25.0 (Schloss *et al.*, 2009). The SEQERR method is applicable only to datasets where the reference sequences from which the reads originated are known. We used USEARCH UCHIME v6.1.544 (Edgar *et al.*, 2011), an algorithm that is widely used for datasets of unknown composition, to identify putative chimeras. The union set of chimeras detected by the *de novo* and reference-based chimera detection modes of UCHIME was taken as the final list of putative chimeras. Reads were first dereplicated (derep_fulllength), otherwise default parameters were used. The reference database (primers removed) used for the reference-based mode was produced by *in silico* amplification of the SILVA SSU 16S rRNA database v111 using the 785F and 1175R primers (allowing for one mismatch). The sensitivity and specificity of UCHIME chimera detection in *in vitro* sets

were evaluated based on the set of high-confidence chimeras identified by SEQERR.

2.2.3 Error analysis (in vitro reads) Error rates were calculated using the non-chimeric reads only. To determine the reference sequences of these reads and their relative abundance, chimeras detected by SEQERR were removed from the datasets and the remaining reads (primers removed) were queried against the set of reference sequences using BLASTN (Altschul *et al.*, 1990, 1997) with a minimum similarity threshold of 80%. Reads that were mapped to a reference with a unique top score were then globally aligned to their parents using the needle program in EMBOSS v6.5.7 (Rice *et al.*, 2000). Global alignments were truncated to discard the end gaps. The error rate per base for insertions, deletions, substitutions and ambiguous base calls was normalized by dividing the number of occurrences of each error type by the sum of truncated (pairwise) alignment lengths of all reads. The overall error rate per base for mismatches was calculated as the sum of error rate per base for each error type.

2.3 Simulations

First, the Mock-I-sim dataset was generated to replicate the *in vitro* Mock-I dataset as closely as possible. After establishing the *in silico* reproducibility of the *in vitro* data properties, two groups of compositionally more complex datasets, collectively referred to as CORE-sim, were simulated. Below, we first detail the methods shared by Mock-I-sim and CORE-sim simulations and then describe these individually.

2.3.1 In silico dataset generation Grinder v0.5.3 (Angly *et al.*, 2012) was used to simulate the PCR amplification of full-length (V5–V7) non-chimeric amplicons based on a given total read count and a relative abundance profile for the reference sequences. The percentage of non-erroneous reads in Mock-I after quality filtering, which was 66% of non-chimeric reads, was approximated by merging two Grinder-produced datasets: a dataset of amplicons with uniformly distributed substitution errors and another dataset with non-erroneous amplicons (6.2 and 93.8% of Grinder reads, respectively). Further errors were subsequently introduced by 454Sim as described below. Chimeras that constitute 14.5% of all reads in the datasets, in the form of bimeras, trimeras and quadmeras, were generated using CHSIM (Edgar *et al.*, 2011), based on a slightly modified n-mera state distribution (bi:81%, tri:18.7% and quad: 0.3%) reported in Quince *et al.* (2011). The relative abundance of chimeras for each top-parent similarity interval between 89 and 99% was modeled to have a linear increase from 0.05 to 0.15 (e.g. 5% of all chimeras were 89–90% similar to their top-parents). The relative abundance of a given parent (reference) sequence in the CHSIM input was assigned the same as that in the Grinder-generated amplicons. Next, Grinder and CHSIM reads were used as input for 454Sim v1.04 (Lysholm *et al.*, 2011) to generate the corresponding flowgrams, which included further insertion and deletion errors. The 454Sim parameters that control the read lengths, homopolymer model for positive flows, and degeneration of positive and negative flows, were manually tuned to obtain a comparable length distribution and error rates to those of the Mock-I dataset.

2.3.2 Mock-I-sim: simulation of 15-species mock dataset The reference sequences of *in vitro* mock species and their relative abundances found in Mock-I were used in Grinder to simulate 34 172 full-length and non-chimeric amplicons, of which 32 047 (93.8%) did not contain any errors, while the remainder contained 0.72% substitution errors on average. In accordance with the abundance of chimeric sequences detected by SEQERR in Mock-I, 5786 chimeras were simulated using CHSIM. Finally, a FASTA file of 39 958 reads was processed with 454Sim to obtain flowgrams. To analyze the errors, reads were quality-filtered/demultiplexed in QIIME by removing primers ('truncate_remove' option) and true chimeras, known by their read labels, were discarded. The remaining non-chimeric reads were aligned to the reference sequences from which they were generated by the needle program, and error rates were calculated as described above for *in vitro* datasets (Section 2.2). In addition, to investigate the influence of the pre-processing order, the putative chimeras were identified by UCHIME (Section 2.2), without or with denoising the reads using denoiser (Reeder and Knight, 2010). For both treatments, the performance of UCHIME was calculated based on the set of simulated chimeras that remained after quality filtering.

2.3.3 Complex mock dataset simulations (CORE-sim) CORE v2012-02-09 (Griffen *et al.*, 2011), an oral microbiome database of 1159 phylogenetically curated 16S rDNA sequences, was amplified *in silico* using TaxMan (Brandt *et al.*, 2012) with the 785 F and 1175R primers (no mismatches allowed). Sequences containing ambiguous bases or more than one unique taxonomic lineage in their headers were discarded. The thus filtered CORE database ($CORE_1$) contained 672 amplicons from 447 taxonomic lineages (447 species; 139 genera). Next, we constructed a second database in which each taxonomic lineage contained only one amplicon, and then used alignments constructed by the needle program (Rice *et al.*, 2000) to remove all amplicons that had (based on the shorter sequence) ≥97% similarity to any other sequence. This second database ($CORE_2$) contained 292 sequences from 292 taxonomic lineages (292 species; 126 genera). When present, the subspecies part of a lineage label (starting with 'subsp.') was removed (for $CORE_1$ and $CORE_2$).

Simulations For each database ($CORE_1$ and $CORE_2$), 50 compositionally different datasets were generated. To obtain ~200 OTUs for each dataset at a 97% similarity threshold, 250 ($CORE_1$) or 200 ($CORE_2$) reference sequences were randomly drawn from their respective databases. For each dataset, ~40 000 reads were simulated, as for Mock-I-sim (see Section 2.3.2). However, in this case, the relative abundances of the sequences were randomly assigned using a powerlaw distribution with parameter 0.4 (Grinder). Reads were quality filtered (QIIME) using the following options: reverse primer removal: 'truncate_only'; minimum sequence length: 150; ambiguous bases: 0; maximum homopolymer run: 6 nt; one forward and two reverse primer mismatches; a minimum quality score of 25; sliding window: 50; and zero barcode errors. Next, the reads were pre-processed in five different ways: no cleaning (NC), denoised (D), chimera checked (CC), denoised and chimera checked (DCC) and chimera checked and denoised (CCD).

Clustering After each pre-processing approach, the remaining reads were clustered at a 97% similarity threshold with 11 different clustering algorithms with their default parameters: CD-HIT v3.1.1 (Li and Godzik, 2006), DNACLUST parallel release 2 (option: -l) (Ghodsi *et al.*, 2011), ESPRIT-Tree 64-bit v11152011 (Cai and Sun, 2011), UCLUST v1.2.22q (Edgar, 2010), USEARCH v6.1.544 (Edgar, 2010), CLUSTOM v0.10 (Hwang *et al.*, 2013), TBC version January 31, 2011 (Lee *et al.*, 2012), CROP v1.33 (options: -e 4400 -s -m 20 -z 400 -r 2 -b int[#uniq sequences/50 + 0.5]) (Hao *et al.*, 2011), M-pick (-f 1) (Wang *et al.*, 2013), swarm v1.2.2 (-d 11; https://github.com/torognes/swarm) and CRUNCHCLUST v43 (-strict -min 0 -max 1000 -keep_n -diff 11 -d_hl -noendgaps) (Hartmann *et al.*, 2012). USEARCH was also evaluated with options 'maxaccepts' and 'maxrejects' set to zero (referred to as USEARCH Optimal). Before clustering with USEARCH and swarm, the reads were dereplicated, sorted by abundance and sequence length. In the case of M-pick, the reads were dereplicated and randomly shuffled to avoid crashes during clustering. To improve its memory usage, some code changes that do not influence the clustering output were made in the M-pick package (see the patch file at http://www.ibi.vu.nl/downloads).

To obtain ground truths for the clustering algorithms, non-erroneous datasets (50 for $CORE_1$ and 50 for $CORE_2$) were generated, which contained only the complete V5–V7 reference sequences of non-chimeric reads (primers removed). These non-chimeric and non-erroneous reads were clustered using each algorithm to obtain the 'reference clusterings'. For evaluating the clustering results of $CORE_2$ simulations, the reference

number of OTUs for each algorithm was derived from its corresponding reference clustering, whereas the number of expected OTUs (due to 97% filtering) was 200.

The clustering accuracy was evaluated by calculating the number of OTUs and the NMI score (Bonder *et al.*, 2012). The NMI score is penalized when sequences of the same label are assigned to different clusters, as well as when sequences of different labels are assigned to the same cluster (Cai and Sun, 2011). The NMI score was calculated at both the species and genus level, where taxonomic lineage labels, which were used as the ground truth, were taken from Grinder sequence headers for non-chimeric reads. For chimeras, the lineage of the parent species with the highest similarity to the chimeric read was taken from the CHSIM read headers.

3 RESULTS

We first analyzed two *in vitro* datasets from a small mock community. Using simulation packages, we then replicated this mock community *in silico* and compared the results with the *in vitro* data. After verifying that the simulation parameters were set appropriately, we simulated two groups of more complex datasets to uncover the effects and accuracy of computational methods on a realistic scale (Fig. 1).

3.1 *In vitro* mock datasets

3.1.1 Error rate estimation To determine a set of primary data attributes required for simulations, we analyzed two *in vitro* datasets derived from pyrosequencing 16S rDNA (V5–V7 HVR) from a mock community of 15 oral microbes. First, we determined the rate of substitution, deletion and insertion errors. For pyrosequencing, one typically expects the substitution rate to be much lower than insertion and deletion rates, as the main cause of errors is over- and undercalls in homopolymer runs (Balzer *et al.*, 2010). However, error analysis of non-chimeric reads in both Mock-I and Mock-II datasets resulted in suspiciously high ratios of substitution errors to insertion and deletion errors. Moreover, the rates of substitutions were 10-fold higher than the reported rates (Gilles *et al.*, 2011).

The high substitution error rates were identified to be caused by (*i*) non-specific binding at degenerate primer positions and (*ii*) the absence of seven (unknown) variant sequences in the reference database, which contained single nucleotide mismatches at one or two positions with respect to their original database references (Fig. 2). These variants were confirmed by Sanger sequencing of newly produced amplicons, starting from genomic DNA. Consecutively, they were added to the initial reference database of 17 sequences. In addition, spurious reads were identified by clustering the denoised non-chimeric reads using UCLUST. Eight spurious OTUs were detected in both datasets and were considered contaminants on manual inspection, leading to the removal of their members (45 reads in Mock-I and 36 reads in Mock-II).

The final error rates were calculated using the contaminant-free datasets with the updated reference database. An overall error rate per base of 0.14% for Mock-I and 0.12% for Mock-II was found for the non-chimeric reads (Table 1). More than half the reads in both datasets were entirely error-free (66% in Mock-I and 74% in Mock-II, see Supplementary Fig. S1 for the error distribution).

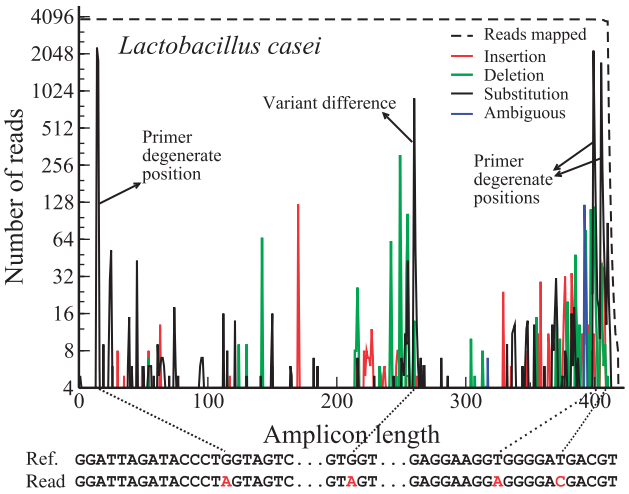

Fig. 2. Errors occurring along the *Lactobacillus casei* reference sequence. Mock-I reads, which were mapped uniquely to the *L. casei* reference sequence by BLASTN, were globally aligned to the reference amplicon. The high substitution peaks at the degenerate positions on the forward and reverse primers reveal non-specific priming. The substitution (G→A) peak at position 260 results from the abundant presence of a variant. The bottom of the figure shows an example with the reference and a read sequence

3.1.2 Estimating the abundance of chimeras 16S rDNA pyrosequencing includes a PCR amplification step prior to sequencing, resulting in the formation of a significant number of chimeras, which can be more than 45% of all reads (Haas *et al.*, 2011). Any realistic computational taxonomy analysis pipeline working with such datasets is required to address this issue. To determine the abundance of chimeric sequences, SEQERR and UCHIME algorithms were separately applied to Mock-I and Mock-II datasets. Using SEQERR, 14.5% (Mock-I) and 17.6% (Mock-II) of the raw reads were found to be chimeric (Table 1). The performance of UCHIME was evaluated on the basis of these sets of high-confidence chimeras detected by SEQERR. The sensitivity, precision and specificity of UCHIME in both datasets were found to be ~0.92, 0.98 and 1.00 (Supplementary Table S1).

3.2 *In silico* mock datasets

3.2.1 Mock-I-sim simulations Using the read length distribution, the abundance and forms of chimeric sequences, and the rate of errors found in the *in vitro* Mock-I dataset, we first replicated the Mock-I datasets *in silico*. Mock-I-sim, the resulting dataset, contained 39 600 reads after quality filtering, of which 5761 were chimeras generated by CHSIM. The length distribution of reads in Mock-I-sim and Mock-I were almost in complete agreement (Supplementary Fig. S2). The overall error rate in reads when simulated chimeras were discarded was the same as in non-chimeric Mock-I reads (0.14%, Table 1), and the distribution of these errors over the dataset was found to be similar to Mock-I and Mock-II (Supplementary Fig. S1). In addition, the overall error rate in reads after chimeras identified by UCHIME were filtered out, as well as UCHIME's performance in terms of sensitivity (0.96) and specificity (1.00), was very similar to the values obtained for Mock-I

Table 1. Properties of the *in vitro* and *in silico* datasets

Datasets	Number of reads	Number of chimeras	Non-erroneous reads (%)	Insertion	Deletion	Substitution	Ambiguous	Total
Mock-I$^{\alpha}$	39 557	5729	66	0.059	0.041	0.043	0.0041	0.14
Mock-I$^{\beta}$	39 557	5401	66	0.064	0.046	0.071	0.0041	0.18
Mock-II$^{\alpha}$	37 932	6669	74	0.049	0.026	0.048	0.0022	0.12
Mock-II$^{\beta}$	37 932	6277	73	0.051	0.028	0.077	0.0022	0.16
Mock-I-sim$^{\alpha}$	39 600	5753	65	0.056	0.040	0.045	5.30E-06	0.14
Mock-I-sim$^{\beta}$	39 600	5535	65	0.059	0.042	0.058	5.60E-06	0.16
Mock-I-sim$^{\gamma}$	39 600	5761	65	0.056	0.040	0.043	5.40E-06	0.14

The number of raw reads and chimeras identified by SEQERR and UCHIME after quality filtering (on primers and bar-code errors only), respectively, are shown (Mock-I, Mock-II and Mock-I-sim), along with the number of chimeras generated by CHSIM (Mock-I-sim). The error rates per base (%) after the removal of chimeras by different methods are listed. α, chimera removal by SEQERR; β, chimera removal by UCHIME; γ, chimera removal based on CHSIM read labels.

Table 2. The sensitivity of UCHIME in *de novo* mode and the reference-based mode, and their union is shown for the Mock-I-sim dataset

Chimera checking mode	UCHIME's sensitivity for different top-parent similarity intervals			
	[89–92)%	[92–95)%	[95–98)%	[98–99)%
De novo CC	0.64	0.63	0.69	0.61
Reference CC	0.93	0.98	0.95	0.69
Union CC	0.97	0.99	0.98	0.82
De novo DCC	1.00	0.98	0.32	0.01
Reference DCC	0.97	0.96	0.31	0.01
Union DCC	1.00	0.98	0.32	0.01

A given percentage interval describes the sequence similarity of chimeras in that group to their parents. CC, chimera checked; DCC, denoised and chimera checked.

(Supplementary Table S1). We note that achieving such resemblance between Mock-I and Mock-I-sim datasets was possible only after extensive re-parameterization of Grinder, CHSIM and 454Sim packages and adjustments in several parts of the simulation pipeline.

To determine the effects of the order of the pre-processing steps, chimera checking was performed separately on denoised and non-denoised datasets, using the *de novo* and reference-based modes of UCHIME. The latter mode outperformed the first in terms of sensitivity, an outcome that was more pronounced on non-denoised sequences (Table 2). Taking the union set of chimeras detected by the two modes resulted in an improved sensitivity in chimera detection and did not inflate the number of false-positive chimeras.

When the reads were denoised prior to chimera checking, the majority of chimeric sequences with high similarity to their parents were altered by denoising. These reads with altered sequences were not detected as chimeras by UCHIME, resulting in a >40% decrease in UCHIME's sensitivity with respect to the original read labels (Table 2, DCC approach). However, these denoised chimeras did not inflate the diversity estimation, as clustering the reads using UCLUST did not yield any unexpected OTUs. Indeed, chimeric sequences with high parent similarity are likely to be denoised to a non-chimeric state.

3.2.2 CORE-sim simulations Having been able to replicate the *in vitro* Mock-I dataset in considerable detail, we extended the simulations to generate datasets of more realistic diversity. To obtain two reference databases of different compositional properties, $CORE_1$ and $CORE_2$ were derived from the CORE oral microbiome database (see Section 2.3.3). $CORE_1$ contained amplicons that are >97% similar to each other, as well as different amplicons from the same taxonomic lineage. To probe the limitations of a database with such properties, the $CORE_2$ database was constructed, where the reference sequences were <97% similar to each other and had unique taxonomic lineages. Using each database, 50 representative datasets of ~40 000 amplicons were simulated from randomly selected reference sequences. To assess the taxonomy analysis pipeline, each dataset was pre-processed in five different ways: no cleaning (NC), denoised (D), chimera checked (CC), denoised and chimera checked (DCC) and chimera checked and denoised (CCD). The resulting reads were clustered separately by 11 clustering algorithms at a 97% similarity threshold. CRUNCHLUST and swarm do not use percentage similarity, but Levenshtein distance. We set this parameter to 11 based on the average read length (368 nt) and 3% divergence. The clusters were evaluated by calculating the NMI score and the number of OTUs. For each simulated dataset, a non-erroneous dataset was created by discarding chimeric reads

and replacing the erroneous reads with the non-erroneous reference amplicon sequences from which they originated. These datasets were used to obtain reference clusterings for each clustering algorithm. Additionally, we set the number of expected OTUs to 200 for $CORE_2$ simulations because the sequences in $CORE_2$ are <97% similar to each other.

The distinctive compositions of $CORE_1$ and $CORE_2$ databases are reflected in the clustering results. In simulations performed using the $CORE_1$ database, OTUs formed by clustering the non-erroneous (reference) datasets resulted in species-level NMI scores between 0.93 and 0.97 (Fig. 3A). This outcome is expected when amplicons from different taxonomic lineages that are ≥97% similar to each other cluster together. In contrast, the average NMI score per clustering algorithm for non-erroneous datasets in $CORE_2$ simulations ranged between 0.98 and 1.00 (Fig. 3B), indicating that almost all reads were correctly clustered when reads with different taxonomic labels were <97% similar to each other. Likewise, for all $CORE_2$ datasets, a given clustering algorithm produced more consistent numbers of OTUs than in $CORE_1$ datasets. The variance in the number of OTUs formed by the clustering algorithms for the reference datasets also shows the difference between these databases (Fig. 3E and F).

At species level (Fig. 3A and B), clustering algorithms expectedly yielded higher NMI scores when clustering the non-erroneous datasets than the pre-processed datasets. At genus level, however, denoising followed by chimera checking (CCD), as well as chimera checking followed by denoising (DCC) resulted in NMI scores higher than reference clusterings (Fig. 3C and D). This was found to be a consequence of denoising, where sequences of different species of the same genus, which were not assigned to the same cluster at a 97% similarity threshold in reference clusterings, were clustered together in the case of the DCC and CCD approaches. Correspondingly, the number of OTUs formed by sequences of the same genus in the reference clusterings was higher than the number of OTUs formed by the DCC and CCD approaches. This also explains why all clustering algorithms (applied after DCC or CCD) underestimate the number of expected OTUs: denoising decreases the divergence of very similar sequences (Fig. 3E and F). When the average number of OTUs for the 50 $CORE_2$ simulations (per algorithm) after the CCD pre-processing approach is considered, the largest diversity underestimation with respect to the number of expected OTUs (200) was observed with CROP (13.6%), whereas DNACLUST, USEARCH, USEARCH Optimal and swarm provided the smallest underestimation (3.4%). When the same comparison is performed between the average number of OTUs obtained in the reference clusterings and after the CCD approach, CLUSTOM yielded the largest overestimation (4.2%), whereas CRUNCHCLUST resulted in the smallest underestimation (1.7%), suggesting that CRUNCHCLUST is rather robust against the remaining noise. CROP resulted in a 2.9% overestimation of the number of OTUs obtained in the reference clustering (173 versus 168 OTUs). The run times of the clustering algorithms are given in Supplementary Table S2.

The NC approach resulted in the lowest NMI scores along with diversity overestimations as high as 20-fold (Fig. 3E and F), confirming earlier clusterings (Bonder *et al*., 2012). Chimera checking or denoising alone was insufficient to obtain accurate clustering. The removal of chimeras was the most important step because the clustering accuracy deteriorated far more by the presence of low-parent similarity chimeras (D) than by the occurrence of erroneous reads (CC). The low error rate in the NC sequences and successful chimera removal by UCHIME gave rise to high NMI scores in the CC approach. This is especially salient in $CORE_2$ simulations, where achieving a good agreement with ground truth labels is more likely because reference sequences in these simulations were at least 3% distant from each other, making inaccurate clustering due to sequence errors less probable.

The best agreement with the number of OTUs formed in reference clusterings was obtained by the DCC and CCD approaches. Although hardly any difference was observed in the number of OTUs formed between the two approaches, the clusters formed using the latter approach showed consistently higher NMI scores. This was due to a large portion (~10%) of undetected chimeras in the DCC approach: denoising not only converted the chimeras into non-chimeric sequences but also resulted in disagreements between the non-detected chimeras and their taxonomic labels. We observed the same denoising effect for undetected chimeras in the CCD approach; however, because their number was substantially lower (~10-fold), the NMI score was not affected.

The differences (NMI scores and number of OTUs) between the various clustering algorithms became less pronounced when they were applied after denoising (D, DCC, CCD). Four exceptions here were CLUSTOM, CRUNCHCLUST, CROP and M-pick, which in both $CORE_1$ and $CORE_2$ simulations resulted in slightly lower species-level NMI scores than other algorithms, whereas at genus level they achieved slightly higher NMI scores. Along with these, the relatively small number of OTUs formed by CLUSTOM and CROP, especially for NC datasets, is indicative of over-clustering, which in this case occurs when sequences from two species of the same genus with <97% similarity are assigned to the same cluster. In contrast, clustering by ESPRIT-Tree and TBC resulted in consistently higher species- and genus-level NMI scores after NC and CC approaches, as well as good estimations of the reference number of OTUs after the DCC and CCD approaches.

4 DISCUSSION

Pre-processing and clustering are two major steps in the analysis of pyrosequenced 16S rDNA amplicons, which have a dramatic influence on the accuracy of downstream efforts such as diversity estimation. In this study, we aimed to improve our understanding of how these methods transform the sequencing datasets into community profiles. To this end, we analyzed data from two pyrosequencing runs that were performed for a 15-species mock microbial community. Data properties derived from these analyses were adapted in computer simulations to overcome the main shortcoming of *in vitro* studies, namely, the absence of ground truth. Error rates, error distributions, read length distributions, and percentage of chimeras are good examples of basic data features that can be incorporated *in silico*. To the best of our knowledge, this is the first study where such characteristics of a given pyrosequencing dataset are replicated with high precision in the raw experimental output (flowgrams). This, for example, resulted in similar chimera detection sensitivity for *in vitro* Mock-I and *in silico* Mock-I-sim data.

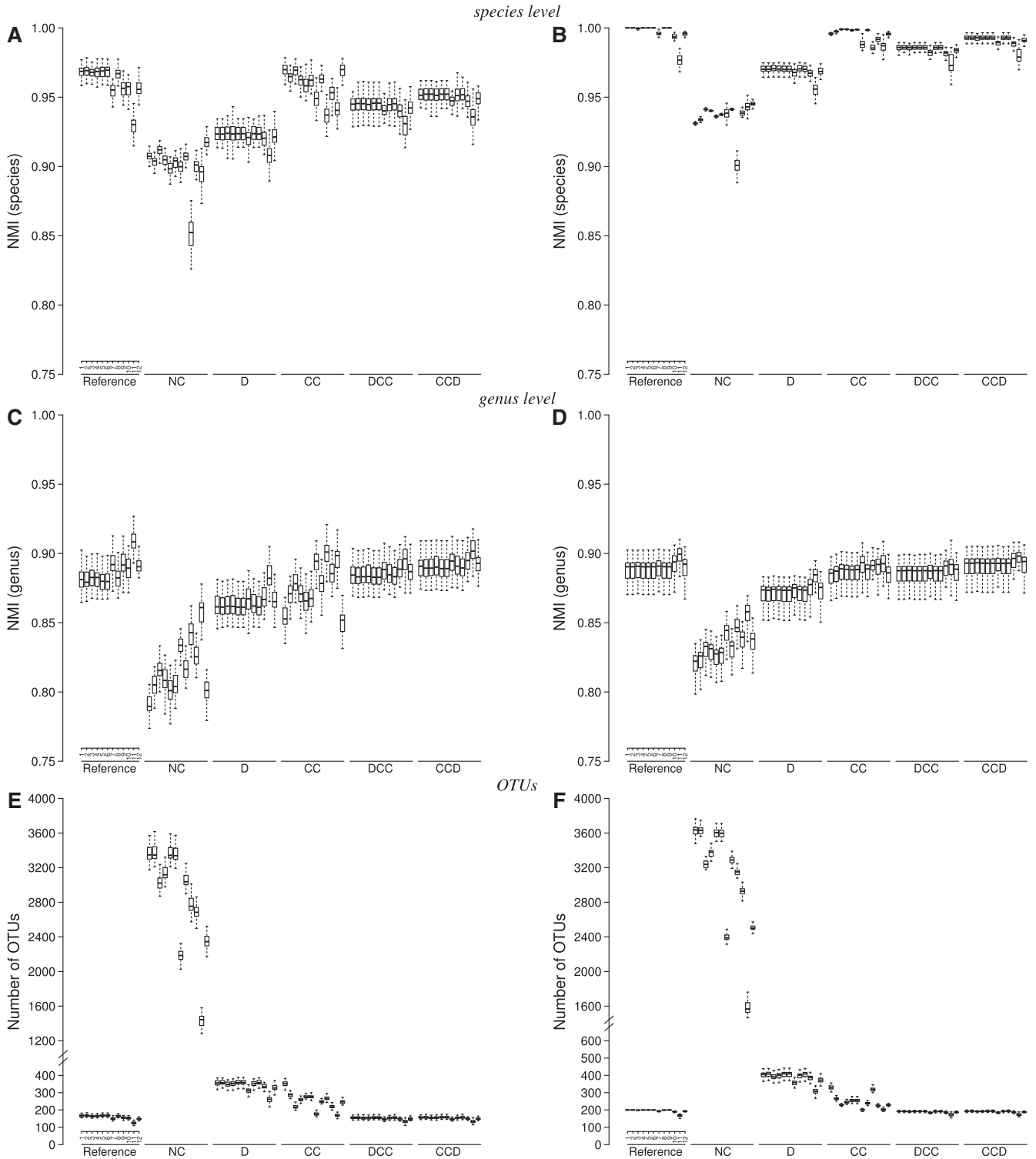

Fig. 3. Clustering results after each pre-processing approach. Species- and genus-level NMI scores and the number of OTUs formed by each algorithm after each pre-processing are shown for the 50 $CORE_1$ (**A**, **C**, **E**) and 50 $CORE_2$ (**B**, **D**, **F**) simulations. Reference, reference clustering of non-chimeric and non-erroneous reads; NC, no cleaning; D, denoised; CC, chimera checked; DCC, denoised and chimera checked; CCD, chimera checked and denoised. The order of the clustering algorithms in all pre-processing blocks is as follows: (1) CD-HIT, (2) DNACLUST, (3) ESPRIT-Tree, (4) UCLUST, (5) USEARCH, (6) USEARCH Optimal, (7) CLUSTOM, (8) TBC, (9) swarm, (10) CRUNCHCLUST, (11) CROP and (12) M-pick

The sequencing error rate (0.14%) in our *in vitro* and simulated datasets is low compared with other reported values, such as 1.07% (Gilles *et al*., 2011) or 0.61% (Schloss *et al*., 2011). This may appear to be limiting the resolution at which the performance of different clustering algorithms can be compared for the pre-processing approaches that include denoising. Other than CLUSTOM and CROP, there was no prominent difference between the outcomes of different algorithms in the DCC and CCD approaches, and they all yielded underestimations of the expected number of OTUs [minimum (min): 3.4% with CCD-swarm; maximum (max): 13.7% with DCC-CROP for $CORE_2$, averaged over all 50 simulations]. Highly erroneous $CORE_2$ simulations with a 6-fold higher error rate (0.86%) did not help differentiate between clustering methods when denoising was applied either (results not shown). In this case, all methods excluding CLUSTOM, CRUNCHCLUST and CROP yielded a small (min: 0.7% DCC-TBC; max: 5.8% CCD-USEACH), but significant, overestimation of the expected number of OTUs (200) instead of an underestimation (M-pick could not be included due to its large memory requirement). These findings suggest that studies where accurate estimations of the sample complexity is of crucial importance should include a thorough analysis to appropriately adjust the sequence/flowgram similarity thresholds used for denoising, preferably with the guidance of an experimental mock dataset, to avoid over- or underestimations. This is supported by an earlier study (Gaspar and Thomas, 2013), which thoroughly examined the spectrum of changes caused by a variety of pyrosequencing noise-removal tools. The authors report that some of these changes were inconsistent with removing noise and could potentially lead to undesired outcomes such as removing rare variants. Similarly, Bakker *et al*. (2012) provide a discussion on the effects of subsampling and error correction, where they observed the loss of rare OTUs as a denoising outcome. Because the datasets used in both studies are derived from environmental samples lacking a ground truth, results presented here can be regarded as complementary with respect to such earlier predictive conclusions.

Recently, Bonder *et al*. (2012) analyzed the diversity in human saliva samples using a taxonomy-independent approach. After denoising, chimera checking and clustering with ESPRIT-Tree, they reported 306 OTUs. In our 50 simulated datasets of closely related sequences ($CORE_1$), DCC followed by ESPRIT-Tree clustering, on average, yielded a 6.0% underestimation of the sample diversity (min: 1.3%; max: 12.1%). In 50 simulations of more divergent sequences ($CORE_2$), the same approach resulted in a 4.2% underestimation on average (min: 1.0%; max: 7.0%). When our *in silico* findings are projected on the results reported by Bonder *et al*., we predict the microbial diversity in this saliva dataset to be slightly higher, namely, between 319 and 326 OTUs.

Although we did not observe major differences in the performance of different clustering algorithms after the DCC and CCD approaches, the NC and CC approaches, as well as the calculation of the NMI scores at different taxonomic levels, provide useful information for the comparison of different clustering methods. An indication of accurate clustering is a high NMI score at both the species and genus level for a given dataset. A method that clusters sequences at similarity thresholds that are lower than a specified value (e.g. 97%) will have a high NMI score at genus level, whereas the species-level NMI score will be low. When the CORE-sim results are evaluated in this context, ESPRIT-Tree and TBC are identified as the two best clustering algorithms of the 11 considered in this study. They consistently achieved the highest NMI scores at both taxonomic levels in pre-processing approaches that do not include denoising, while simultaneously providing very close approximations after the DCC or CCD treatments to the reference number of OTUs.

Here, we assessed different data pre-processing strategies in combination with several clustering algorithms that are commonly used in taxonomy analysis. Based on our as well as previous results (Bonder *et al*., 2012; Schloss *et al*., 2011), we suggest an analysis workflow, which (after quality filtering) starts with chimera removal. We showed that the most accurate chimera detection is obtained when the union of chimeras detected by UCHIME *de novo* and reference-based modes are taken, where the reference database should be extensive, quality checked [e.g. SILVA (Quast *et al*., 2013) or Greengenes (DeSantis *et al*., 2006)] and preferably trimmed to the HVR of interest (Brandt *et al*., 2012). As the current UCHIME algorithm is remarkably fast, chimera checking can be performed before denoising, resulting in 2-fold shorter denoising time due to the reduction in unique flowgrams after chimera removal. The resulting reads should be corrected for sequencing errors, preferably with a denoising algorithm that is fine-tuned for the error rate that is specific to the dataset. The chimera-checked and denoised reads can be clustered with ESPRIT-Tree, TBC or a significantly faster clustering algorithm such as USEARCH or swarm in the case of run-time constraints.

ACKNOWLEDGEMENTS

The authors thank Mark Buijs for preparing the amplicons for Sanger sequencing and the authors of 454Sim for their help in re-tuning the 454Sim parameters. The authors also thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster.

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

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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.

Angly,F.E. *et al.* (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**, e94.

Bakker,M.G. *et al.* (2012) Implications of pyrosequencing error correction for biological data interpretation. *PLoS One*, **7**, e44357.

Balzer,S. *et al.* (2010) Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–i425.

Bonder,M.J. *et al.* (2012) Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics*, **28**, 2891–2897.

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.
Crielaard,W. *et al.* (2011) Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. *BMC Med. Genomics*, **4**, 22.
DeSantis,T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
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.
Gaspar,J.M. and Thomas,W.K. (2013) Assessing the consequences of denoising marker-based metagenomic data. *PLoS One*, **8**, e60458.
Ghodsi,M. *et al.* (2011) DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, **12**, 271.
Gilles,A. *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, 245.
Griffen,A.L. *et al.* (2011) CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLoS One*, **6**, e19051.
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.
Hao,X. *et al.* (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, **27**, 611–618.
Hartmann,M. *et al.* (2010) V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *J. Microbiol. Methods*, **83**, 250–253.
Hartmann,M. *et al.* (2012) Significant and persistent impact of timber harvesting on soil microbial communities in Northern coniferous forests. *ISME J.*, **6**, 2199–2218.
Hugenholtz,P. and Huber,T. (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.*, **53**, 289–293.
Huse,S.M. *et al.* (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, **12**, 1889–1898.
Hwang,K. *et al.* (2013) CLUSTOM: a novel method for clustering 16S rRNA next generation sequences by overlap minimization. *PLoS One*, **8**, e62623.
Kembel,S.W. *et al.* (2012) Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.*, **8**, e1002743.
Kirchman,D.L. *et al.* (2010) The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ. Microbiol.*, **12**, 1132–1143.
Kittelmann,S. *et al.* (2013) Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. *PLoS One*, **8**, e47879.
Klindworth,A. *et al.* (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.*, **41**, e1.
Kraneveld,E.A. *et al.* (2012) The relation between oral *Candida* load and bacterial microbiome profiles in Dutch elderly. *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.
Le Chatelier,E. *et al.* (2013) Richness of human gut microbiome correlates with metabolic markers. *Nature*, **500**, 541–546.
Lee,J.H. *et al.* (2012) TBC: a clustering algorithm based on prokaryotic taxonomy. *J. Microbiol.*, **50**, 181–185.
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.
Lysholm,F. *et al.* (2011) An efficient simulator of 454 data using configurable statistical models. *BMC Res. Notes*, **4**, 449.
Quast,C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
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.
Rice,P. *et al.* (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
Richter,D.C. *et al.* (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**, e3373.
Roesch,L.F. *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.*, **1**, 283–290.
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.
Schloss,P.D. *et al.* (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*, **6**, e27310.
Sergeant,M.J. *et al.* (2012) High-throughput sequencing of 16S rRNA gene amplicons: effects of extraction procedure, primer length and annealing temperature. *PLoS One*, **7**, e38094.
Siqueira,J.F. Jr *et al.* (2012) Pyrosequencing as a tool for better understanding of human microbiomes. *J. Oral Microbiol.*, **4**, 10743.
Somboonna,N. *et al.* (2012) Metagenomic profiles of free-living archaea, bacteria and small eukaryotes in coastal areas of Sichang island, Thailand. *BMC Genomics*, **13** (**Suppl. 7**), S29.
Sun,Y. *et al.* (2012) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform.*, **13**, 107–121.
Tamaki,H. *et al.* (2011) Analysis of 16S rRNA amplicon sequencing options on the Roche/454 next-generation titanium sequencing platform. *PLoS One*, **6**, e25263.
Wang,X. *et al.* (2013) M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformatics*, **14**, 43.
White,J.R. *et al.* (2010) Alignment and clustering of phylogenetic markers—implications for microbial diversity studies. *BMC Bioinformatics*, **11**, 152.
Woo,P.C.Y. *et al.* (2008) Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clin. Microbiol. Infec.*, **14**, 908–934.