environmental microbiology

Environmental Microbiology (2014) doi:10.1111/1462-2920.12610

Limits to robustness and reproducibility in the demarcation of operational taxonomic units

Thomas S. B. Schmidt, João F. Matias Rodrigues and Christian von Mering*
Institute for Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, Zürich 8057, Switzerland.

Summary

The demarcation of *operational taxonomic units* (OTUs) from complex sequence data sets is a key step in contemporary studies of microbial ecology. However, as biologically motivated 'optimal' OTU-binning algorithms remain elusive, many conceptually distinct approaches continue to be used. Using a global data set of 887 870 bacterial 16S rRNA gene sequences, we objectively quantified biases introduced by several widely employed sequence clustering algorithms. We found that OTU-binning methods often provided surprisingly non-equivalent partitions of identical data sets, notably when clustering to the same nominal similarity thresholds; and we quantified the resulting impact on ecological data description for a well-defined human skin microbiome data set. We observed that some methods were very robust to varying clustering thresholds, while others were found to be highly susceptible even to slight threshold variations. Moreover, we comprehensively quantified the impact of the choice of 16S rRNA gene subregion, as well as of data set scope and context on algorithm performance. Our findings may contribute to an enhanced comparability of results across sequence-processing pipelines, and we arrive at recommendations towards higher levels of standardization in established workflows.

Introduction

High-throughput sequencing technology has enabled the characterization of microbial communities at ever-increasing resolutions: individual environments have been probed to depths of millions of sequences, and even smaller-scale studies may routinely provide hundreds of thousands of reads. While cultivation-independent whole-genome sequencing has received increasing attention in the functional characterization of individual communities (The Human Microbiome Project Consortium, 2012a), targeted surveys for specific taxonomic marker genes, such as the 16S rRNA gene (Lane *et al.*, 1985; Olsen *et al.*, 1986), remain integral to many contemporary studies of microbial ecology. An essential first step in analysing targeted sequencing data sets is often the demarcation of basic units of diversity, ideally corresponding to 'true' microbial lineages that were present in the sample (Gevers *et al.*, 2005; Cohan, 2006; Koeppel *et al.*, 2008). However, in the absence of a unifying bacterial species concept (Doolittle and Papke, 2006; Achtman and Wagner, 2008; Doolittle and Zhaxybayeva, 2009), biologically motivated 'optimal' diversity unit definitions remain elusive, and a pragmatic approach is usually taken in practice: *operational taxonomic units* (OTUs), defined as clusters of 16S gene sequence similarity, are used to approximate microbial taxa. Because OTU demarcation from complex rRNA gene data sets is conceptually straightforward and often computationally efficient, OTUs are the backbone of established workflows for the ecological characterization of microbial communities, such as MOTHUR (Schloss *et al.*, 2009) and QIIME (Caporaso *et al.*, 2010).

As the identification of 'optimal' partitions of large 16S rRNA gene data sets remains an open problem, a wide variety of OTU-binning methods have been developed. Traditionally, *hierarchical-clustering algorithms* (implemented e.g. in MOTHUR, ESPRIT (Sun *et al.*, 2009) and HPC-CLUST (Matias Rodrigues and von Mering, 2014)) have been widely used, as have their *heuristic* approximations, which include CD-HIT (Li and Godzik, 2006; Fu *et al.*, 2012), UCLUST (Edgar, 2010), UPARSE (Edgar, 2013), ESPRIT-TREE (Cai and Sun, 2011), DYSC (Zheng *et al.*, 2012), MSCLUST (Chen *et al.*, 2013), MBKM (Wei *et al.*, 2012) and LSH (Rasheed *et al.*, 2013). Although these methods rely on 'hard' clustering thresholds, several 'soft-threshold' or 'threshold-less' approaches have been proposed, including CROP (Hao *et al.*, 2011), M-PICK (Wang *et al.*, 2013) and BEBAC (Cheng *et al.*, 2012). Moreover, several algorithms rely on additional external data, either in the form of reference OTUs (e.g.,

Received 25 April, 2014; accepted 21 August, 2014. *For correspondence. E-mail mering@imls.uzh.ch; Tel. +41 44 635 31 47; Fax +41 44 635 68 64.

'reference-based OTU picking' strategies as implemented in QIIME), or in the form of additional biological signals, notably ecotype simulation (Koeppel *et al*., 2008), the tree-based evolutionary placement algorithm (EPA)-poisson tree processes (PTP) (Zhang *et al*., 2013) or distribution-based clustering (Preheim *et al*., 2013).

Given this diversity of available methods, several studies have aimed to identify those OTU-binning strategies that provide the 'best' partitions with respect to different objectives. The arguably most straightforward parameter to optimize for is the total number of clusters, as OTU counts serve as basis for estimates of community richness. Consequently, various studies have benchmarked OTU definitions based on total cluster counts, usually by assessing their overestimation of diversity with respect to test sets of known taxonomic composition, obtained by simulation or sequencing of mock communities (Sun *et al*., 2009; Huse *et al*., 2010; Schloss, 2010; White *et al*., 2010; Barriuso *et al*., 2011; Bonder *et al*., 2012; Chen *et al*., 2013). Other benchmarking strategies include data set-internal quality measures (optimizing the ratio of specificity and sensitivity with respect to known input, e.g. by Schloss and Westcott, 2011; Li *et al*., 2012; Chen *et al*., 2013; Preheim *et al*., 2013) and external benchmarking against 'ground truth' data sets, optimizing for 'taxonomically pure' clusters (White *et al*., 2010; Cai and Sun, 2011; Sun *et al*., 2011; Bonder *et al*., 2012; Chen *et al*., 2013; Wang *et al*., 2013). More recently, OTU-binning methods have also been evaluated with respect to ecological consistency, by us and others (Koeppel and Wu, 2013; Schmidt *et al*., 2014).

Flexibility in 16S rRNA gene sequence processing workflows is introduced at several levels, notably when removing sequencing noise and filtering for chimeric sequences (e.g., Schloss *et al*., 2011; Bonder *et al*., 2012), by sequence alignment strategies (Schloss, 2010; 2012; White *et al*., 2010; Barriuso *et al*., 2011; Sun *et al*., 2011; Wang *et al*., 2011) and by sequence distance calculation (Schloss, 2010; Barriuso *et al*., 2011); however, these factors have been extensively discussed previously and are beyond the scope of our current study. Here, we are mainly concerned with the differences introduced by algorithmic choices, at the heart of the sequence-clustering step.

Given the large flexibility at different levels, considerable efforts have been made to integrate and standardize workflows into one-stop pipelines such as MOTHUR, QIIME or CD-HIT-OTU (Li *et al*., 2012). However, in spite of these efforts and of a substantial body of literature on benchmarks, 'optimal' OTU demarcation strategies remain elusive, and the choice of methods and parameters varies considerably between studies. In consequence, it is generally difficult to compare ecological descriptions across studies, and study design is sometimes redundant, implementing complementary data analysis strategies to control for effects on biological interpretation – for example, the *human microbiome project* data were analysed using multiple workflows, relying on MOTHUR [average linkage (AL) clustering] and QIIME-UCLUST (The Human Microbiome Project Consortium, 2012b). Moreover, in spite of substantial efforts to benchmark OTU demarcation strategies, surprisingly little is known about the systematic differences *between* methods. It is not clear how similar approaches are in terms of resulting cluster *composition* and how putative differences may bias biological interpretation beyond richness estimates. In other words, although OTU-binning strategies have been benchmarked extensively against varying concepts of 'optimality', systematic differences between methods are currently not well understood.

In this study, we explore limits to robustness and reproducibility in the demarcation of OTUs. We pursue a simple unifying question: how similar are different clustering methods? We approach this problem from various angles and quantify differences between five widely used clustering algorithms (average (AL), complete (CL) and single linkage (SL) clustering, as well as the heuristics CD-HIT and UCLUST) and for the recently published UPARSE, which implements adaptive on-the-fly chimera filtering. We selected these methods, because (i) they are capable of processing very large data sets, (ii) they are widely used in general and (iii) they rely on sequence data only (and not on external reference OTUs or additional phylogenetic or ecological signals to inform sequence clustering). We first revisit and quantify the observation that these methods generally provide diverging total cluster counts and complement this earlier finding by investigating how these differences propagate to the level of cluster size distributions. We then turn our focus towards cluster *composition* and investigate how concordant the methods are when partitioning the very same set of sequences. In other words, do different methods tend to form consistent clusters, i.e. do they group similar sets of sequences? We first approach this question anecdotally, by re-analysing the *human skin microbiome* (HSM) data set for an individual clustering threshold, for which we explore how differences between methods translate to biases in ecological descriptions. We then broaden the scope of investigation to studying a global, comprehensive survey of publicly available full-length 16S rRNA gene sequence data across a wide range of clustering thresholds. In particular, we assess how robust methods are against slightly changing thresholds and how reproducible partitions are across methods and thresholds. Finally, we assess robustness to changing clustering *context* (i.e., how does rich/sparse sequence space influence OTU demarcation?) and to the choice of 16S rRNA gene subregion.

Results

Quantitative differences between OTU definitions

When studying microbial communities, a crucial first step is often the characterization of local community complexity, richness and evenness, collectively referred to as *α-diversity*. Many measures of α-diversity rely on the total number of unique taxa observed in a sample (approximated by OTUs in practice), as well as their relative abundances. In consequence, several studies have used total cluster counts to benchmark OTU definitions (Sun *et al.*, 2009; Huse *et al.*, 2010; Schloss, 2010; White *et al.*, 2010; Barriuso *et al.*, 2011; Bonder *et al.*, 2012; Chen *et al.*, 2013).

To confirm and refine such previous observations, we clustered a global data set of 887 870 near full-length bacterial 16S rRNA gene sequences using six different methods: AL, CL, SL, CD-HIT, UCLUST and UPARSE. We observed systematic shifts in total cluster counts between methods (Fig. 1A; Table S2): when clustering to the same nominal sequence similarity threshold, SL provided the lowest, UCLUST the highest total OTU counts. All methods showed exponentially increasing counts with increasing clustering stringency, with over-exponential increases at very high similarities (≥ 98/99%). Interestingly, log-linear slopes were almost identical for AL, CL and CD-HIT, while UPARSE and SL diverged significantly; strikingly, the curve for UCLUST was not perfectly monotonous in the 96–97% threshold range.

These differences in overall cluster counts translated to differences in cluster size distributions. At a nominal similarity threshold of 97%, all tested methods provided differentially skewed OTU size histograms, with small OTUs (≤ 100 sequences) being notably overrepresented for UCLUST and underrepresented for SL (Fig. 1B). Indeed, although SL clustered 78.8% of sequences into the largest 1.5% of OTUs (> 100 sequences), the largest 1% of UCLUST OTUs contained only 40.3% of total sequences; all other methods provided intermediate behaviour (Fig. 1C).

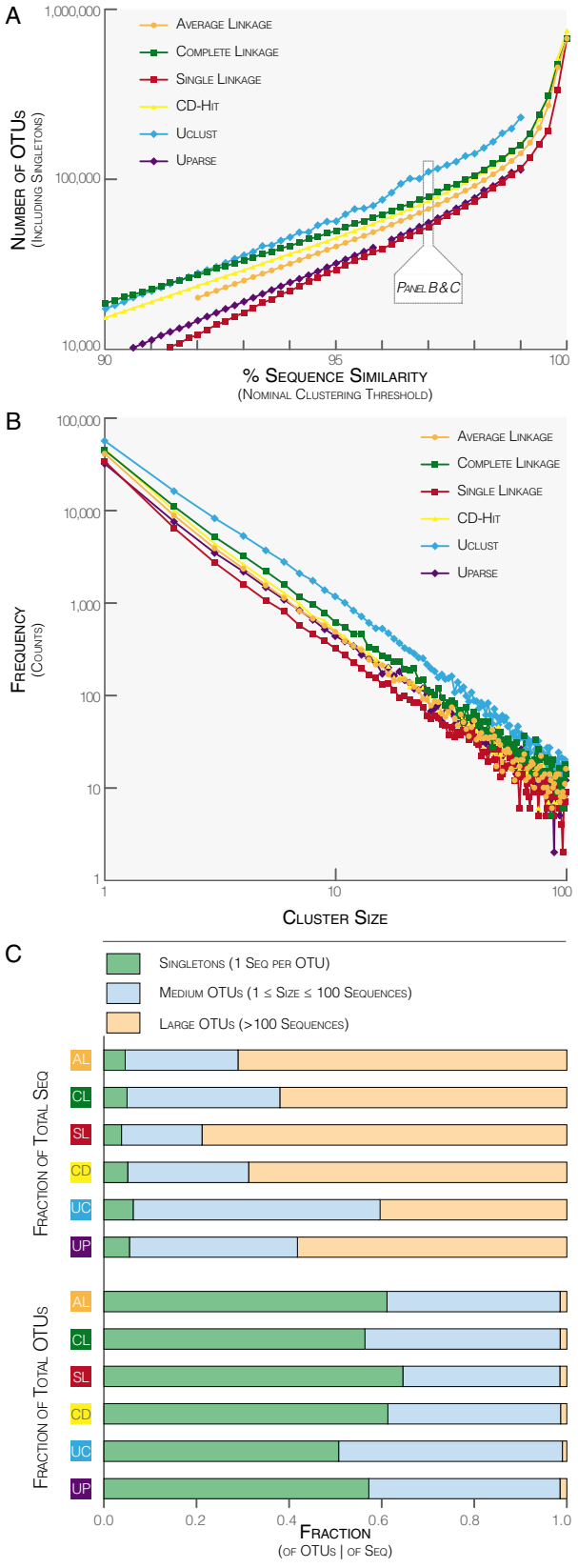


Fig. 1. Quantitative differences between clustering methods.
A. Differences in total OTU counts when clustering a global data set of 887 870 bacterial 16S rRNA gene sequences according to different methods. Note that UPARSE filtered for chimeric sequences differently than the other methods, which led to different numbers of sequences being clustered at different cut-offs (see Table S2 and Fig. S7). Moreover, UCLUST and UPARSE did not cluster to > 99% similarity, with additional missing data points for UPARSE (see Appendix S1).
B. Differences in OTU size distributions between methods when clustering to 97% nominal sequence similarity.
C. Differential dominance of singleton (1 sequence) and large OTUs (> 100 sequences) at 97% similarity. Methods differed in the fraction of total sequences (upper panel) and of total OTUs (lower panel) that fell into different OTU size categories.

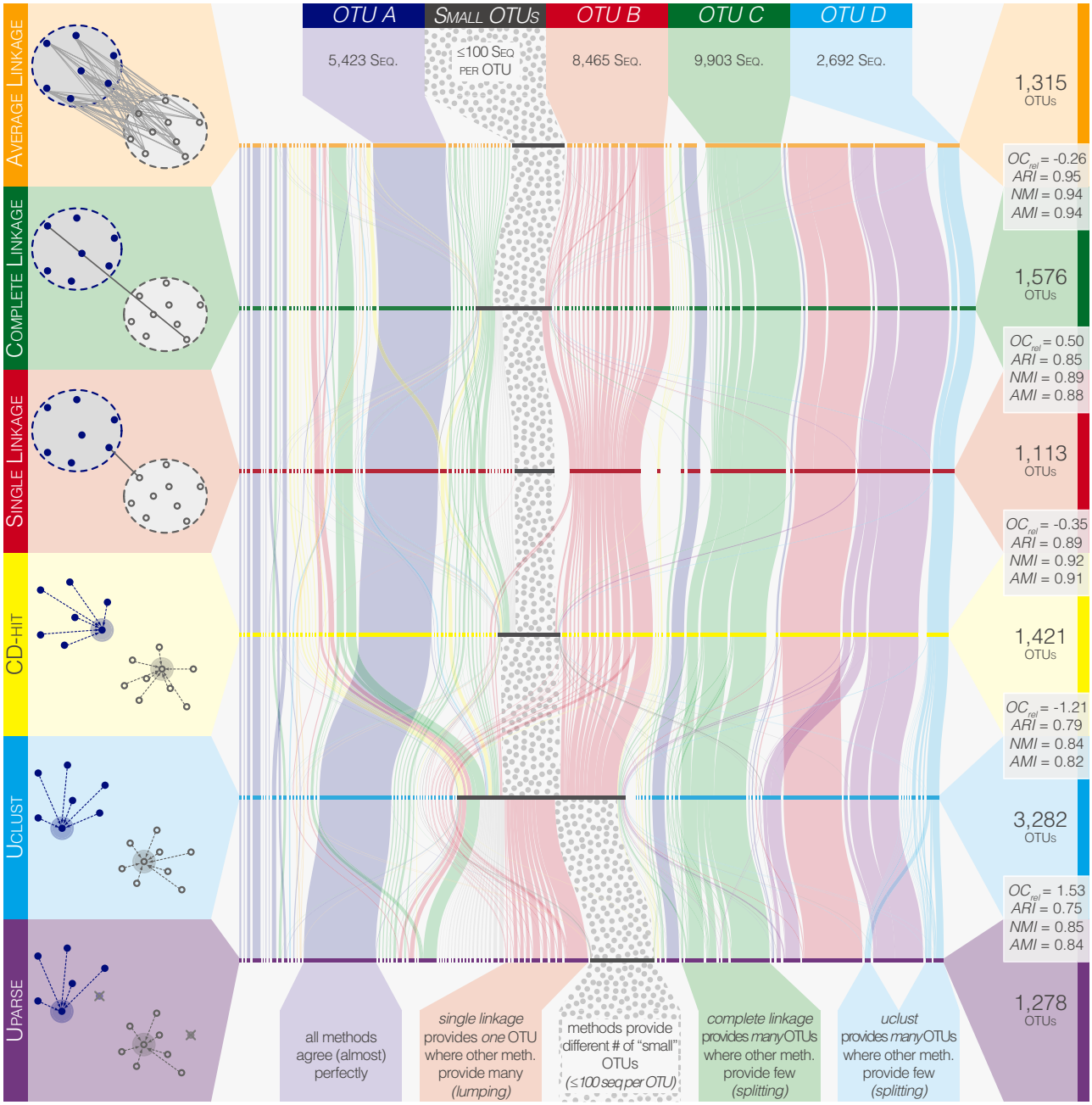


Fig. 2. Differences in OTU composition at an individual datapoint. There were 90 620 bacterial 16S rRNA gene sequences of the *human skin microbiome* (HSM) data set were clustered to 97% sequence similarity according to different methods (algorithms illustrated in left panel); note that here, we additionally used the '-id 0.97' option for UPARSE. OTUs are shown as horizontal bars per clustering method; sequences are represented as vertical bands. Colouring highlights different sequence subsets; the dotted grey band corresponds to sequences in small OTUs (≤ 100 sequences). Although some clusters (e.g., 'OTU A') were almost identical across methods, others were more discordant. Partition similarities (right panel) were quantified in terms of relative OTU counts (OC_{rel} as binary logarithm), AMI, NMI and ARI values; see also Table 1.

Differences in OTU composition between clustering methods

How do such differences in total cluster count and cluster size distribution translate to the level of individual OTUs? How similar are the various clustering methods with regard to the actual cluster composition? To approach these questions in a concrete example, we clustered 90 620 sequences of the deeply sequenced and frequently cited HSM data set (Grice *et al.*, 2009) to 97% sequence similarity and traced differences in clusterings between methods in an alluvial flow diagram (Fig. 2).

Table 1. Pairwise partition similarities at an individual datapoint.

	UPARSE		UCLUST		CD-HIT		SL		CL	
	AMI *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}
AL	0.963	0.870	0.842	0.816	0.957	0.949	0.943	0.936	0.938	0.916
	0.965	–0.207	0.854	1.320	0.960	0.112	0.945	–0.241	0.942	0.261
CL	0.948	0.902	0.809	0.793	0.930	0.931	0.885	0.852		
	0.951	–0.469	0.834	1.058	0.936	–0.149	0.888	–0.502		
SL	0.910	0.711	0.800	0.755	0.915	0.888				
	0.911	0.033	0.807	1.560	0.917	0.352				
CD-HIT	0.945	0.868	0.818	0.787						
	0.948	–0.319	0.837	1.208						
UCLUST	0.839	0.748								
	0.851	–1.528								

The *human skin microbiome* (HSM) data set was clustered to 97% nominal sequence similarity according to different methods (see also Fig. 2), and pairwise partition similarities were calculated as AMI, NMI, ARI and relative OTU counts (column-wise, as binary log-ratios, i.e. read 'CL provided $2^{0.261}$ times as many OTUs as AL'). Full pairwise partition similarities across thresholds are available in Tables S4–S6.

Clearly, the tested OTU definitions provided markedly distinct partitions with respect to both cluster composition and cluster counts and sizes. Although some clusters (e.g., *'OTU A'*) were almost identical between partitions, other sequence subsets showed characteristic behaviour for the different clustering methods. SL tended to produce large, comprehensive clusters (e.g., *'OTU B'*), lumping together sequences that were split into multiple smaller OTUs by the other methods; this is in line with the generally *inclusive* SL algorithm (see also Fig. 2, left panel). In contrast, both CL (e.g., *'OTU C'*) and UCLUST (e.g., *'OTU D'*) tended to split sequences into more and smaller clusters; in particular, these methods also clustered more sequences into 'small' OTUs (≤ 100 sequences, marked with grey dots), which is in line with their 'splitting' behaviour and the above observations on total cluster counts and size distributions. However, in spite of highly fluctuating partitions between methods, there were remarkably few cases of truly discordant clustering (sequences that completely traversed OTU boundaries), at least at the given resolution. Rather, differences between sets were almost always due to differential 'lumping' or 'splitting' of OTUs, although in some cases, the heuristic methods generated counterintuitive sub-partitions (e.g., sequences from *'OTU C'* clustered into one large UCLUST OTU that contained parts of several smaller UPARSE and CD-HIT OTUs).

We used several measures to quantify the observed differences between sets (Fig. 2, right panel, and Table 1): we assessed pairwise set similarity in terms of relative total OTU count (as binary log ratio, OC_{rel}), as well as in terms of cluster composition, using Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI) and the *Adjusted Rand Index* (ARI; see Appendix S1). We observed that UCLUST and SL provided the most dissimilar partitions (NMI = 0.81, AMI = 0.80, ARI = 0.76, $OC_{rel} = \pm 1.56$), whereas AL and CD-HIT were the most similar (NMI = 0.96, AMI = 0.96, ARI = 0.95, $OC_{rel} = \pm 0.11$).

Qualitative differences between clustering methods may bias biological interpretation

To test how these differences in cluster counts and composition may influence biological interpretation, we re-analysed the HSM data set with respect to different ecological parameters. For the 21 skin sites sampled in the original HSM study, we estimated local diversity (α-diversity) based on three widely used measures: (i) the *Chao1* index, an abundance-based richness estimator that corrects for rare (unseen) classes (Chao, 1984), (ii) the *inverse Simpson* index, a classical abundance-weighted diversity measure (Simpson, 1949) and (iii) the entropy-based *Shannon* index (Shannon, 1948). Moreover, we assessed pairwise community similarity between habitats (β-diversity) using three different methods: (i) the abundance-informed *Sørensen–Dice–Czekanowski* (SDC) similarity index, which is closely related to the more well-known *Bray–Curtis* dissimilarity (Dice, 1945; Bray and Curtis, 1957), (ii) the *Morisita–Horn* (MH) overlap index (Horn, 1966) and (iii) Chao's *abundance-based Jaccard* (J_{abd}) index, which corrects for rare classes (Chao *et al.*, 2004); see Appendix S1 for more details. The results across different clustering methods are shown in Fig. 3 and Table S3.

We observed that clustering methods generally provided highly divergent α-diversity estimates: e.g., Chao1 richness estimates differed by up to 7.4-fold for individual samples ('inguinal crease', UCLUST versus SL, Fig. 3A). Average shifts in diversity estimates were

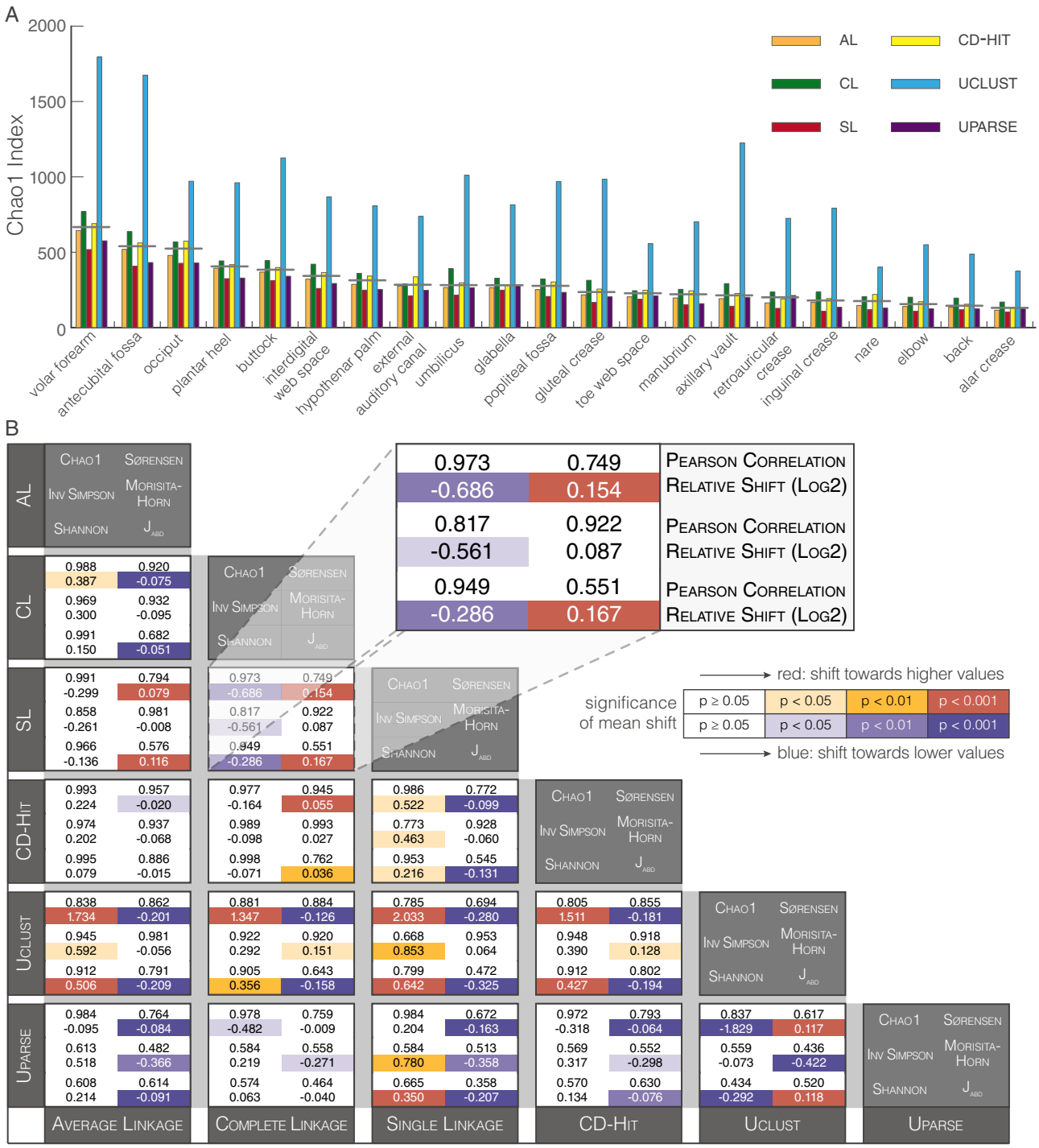


Fig. 3. Clustering methods introduce differential bias to ecological data descriptions.
A. *Chao1* richness estimates for the 21 skin sites of the HSM data set when clustering to 97% sequence similarity using different methods. Horizontal grey bars indicate median values.
B. Biases to estimators of local diversity (*Chao1, inverse Simpson* and *Shannon* indices) and community similarity (*Sørensen–Dice–Czekanowski*, *Morisita–Horn* and *abundance-corrected Jaccard* indices; see Appendix S1) for the same 21 skin sites. For every pair of clustering methods, Pearson correlations (upper cells) indicate differences in diversity ranking between skin sites per index. Average shifts in absolute diversity estimates per index ('relative shift', lower cells) between methods are calculated 'row-wise' (e.g., for the highlighted cell, read 'SL provided lower *Chao1* estimates than CL'). Cell colouring indicates significance levels and direction of shifts between methods (Mann–Whitney *U*-test). Raw data on diversity estimates are provided in Table S3.

Table 2. Partition similarities may predict trends in ecological data description.

	Inv Simpson		Shannon		Chao1		Soerensen-Dice		Morisita-Horn		J_{abd}	
	Correlation	*Shifts*	*Correlation*	*Shifts*	*Correlation*	*Shifts*	*Correlation*	*Shifts*	*Correlation*	*Shifts*	*Correlation*	*Shifts*
AMI	0.543	0.673	0.611	0.891	0.884	0.866	0.642	0.864	0.493	0.381	0.606	0.858
NMI	0.541	0.671	0.610	0.895	0.885	0.870	0.648	0.870	0.487	0.383	0.611	0.864
OC_{rel}	0.687	0.631	0.762	0.787	0.886	0.986	0.740	0.771	0.579	–0.411	0.669	0.712
ARI	0.802	0.770	0.861	0.910	0.861	0.814	0.837	0.882	0.767	0.688	0.757	0.893

Pairwise partition similarities between methods (OC_{rel}, AMI, NMI and ARI), as shown in Table 1, were correlated with trends and shifts in α- and β-diversity estimates between methods across skin habitats (shown as Pearson correlations and relative shifts in Fig. 3B). In other words, each value in the table indicates how well partition similarities correlated with similarities in ecological data descriptions (Spearman rank correlation).

often systematic across skin sites and statistically significant (Fig. 3B): UCLUST provided significantly higher α-diversity estimates than other methods (binary log ratio, 1.347–2.033 for Chao1, 0.073–0.853 for inverse Simpson, 0.292–0.642 for Shannon), while SL estimated systematically lower α-diversities and the other methods provided intermediate behaviour. Nevertheless, all methods generally ranked the 21 samples similarly by α-diversity. In particular, AL, CL and CD-HIT provided very similar diversity trends (Pearson correlation across skin sites, 0.977–0.993 for Chao1, 0.969–0.989 for inverse Simpson and 0.991–0.998 for Shannon, Fig. 3B), whereas for UCLUST and UPARSE, trends were considerably less similar to other methods (Pearson correlation, lower limit of 0.434, Shannon index UCLUST versus UPARSE).

We observed similar effects for estimates of β-diversity. When comparing all pairwise community similarities between skin samples, AL, CL and CD-HIT provided very similar trends (Pearson correlation, 0.920–0.957 for SDC, 0.932–0.993 for MH, J_{abd} generally lower; Fig. 3B), whereas UCLUST, UPARSE and SL provided lower correlations to other methods. Interestingly, MH estimates correlated very well (> 0.9) between all methods except UPARSE, while J_{abd} provided comparatively low correlations (0.358–0.886); the latter effect is probably due to the correction for 'rare' (unseen) taxa implemented in J_{abd}, which will differentially distort community similarities according to OTU abundance distributions. SDC and J_{abd} estimates of community similarity were systematically and significantly lower for UCLUST (Mann–Whitney *U*-test, $P < 3 * 10^{-10}$) and higher for SL ($P < 1.4 * 10^{-14}$) when compared with all other methods. Systematic shifts in estimated community similarity among AL, CD-HIT and CL were far less pronounced, but sometimes statistically significant.

Thus, the choice of clustering method clearly had a significant impact on the ecological characterization of the HSM data set. Generally, UCLUST, UPARSE and SL deviated most in their descriptions of the data, whereas AL and CD-HIT, and to a lesser extent also CL, were more similar in diversity estimates between themselves. Next, we tested how these trends between clustering methods were captured by different measures of partition similarity (provided in Table 2). In other words, we asked whether differences in cluster counts and cluster composition between methods were predictive of differences in ecological descriptions of the data set. We found that in general, relative OTU counts between partitions (OC_{rel}) and ARI, AMI and NMI correlated well with trends in Chao1 (Spearman correlation of pairwise partition similarities with diversity correlation across skin sites, 0.861–0.886, Table 2). Correlations with trends in inverse Simpson, Shannon, SDC and J_{abd} were less pronounced, and trends in MH were moderately captured only by the ARI (Spearman correlation, 0.767). Systematic shifts in diversity estimates between methods in general corresponded well to trends in ARI, AMI and NMI for all diversity estimates except MH and inverse Simpson. Relative OTU counts were particularly poor indicators for shifts in the latter two estimators, but reasonably good indicators for the OTU count-correcting indices Chao1 and J_{abd}, as expected. Thus, AMI, NMI and ARI between partitions provided by different clustering methods were generally indicative both of differential trends in diversity estimates, as well as of their systematically shifted absolute values. In other words, differences between methods in cluster composition may in part explain biased diversity estimation between methods; when comparing two clusterings of the same data, AMI, NMI and ARI may in general indicate how biased these sets will be for downstream ecological descriptions.

General trends in robustness, reproducibility and similarity

In the previous sections, we have discussed how the choice of clustering method may influence the biological interpretation of 16S rRNA gene sequence data. However, these findings are arguably anecdotal: they pertain to only one data point (clustering to 97% nominal similarity) for a defined model data set. To generalize our observations, we clustered the 90 620 sequences in the HSM data set to similarity thresholds ranging from 90% to 100% (in steps of 0.2%) and calculated pairwise partition similarities for all combinations of clustering methods and thresholds as AMI, NMI and ARI (Fig. S1–S3; Tables S4–S6). Moreover, to further broaden the scope of investigation, we performed a similar experiment on a global data set of 887 870 bacterial 16S rRNA gene sequences, sampled from a wide array of environments (Fig. 4, Figs S4 and S5; Tables S7–S9).

We observed that AL, CD-HIT and CL generally provided high partition similarities between themselves over wide cut-off ranges (AMI/NMI/ARI ≥ 0.9), whereas SL, UCLUST and UPARSE provided reasonably high partition similarities to other methods for the HSM data set, but considerably lower similarities for the global data set. In most cases, maximum partition similarities between two given methods were *off diagonal*, indicating that nominal clustering thresholds were not directly equivalent across algorithms. For all methods, and at any threshold, best matches to SL partitions were shifted towards higher nominal thresholds for SL; in other words, when clustering e.g. to 97% similarity using CL, the most similar SL partition was at $>97\%$ SL clustering. The reverse was true for UCLUST, and to a lesser extent CL: for these methods, maximum similarities to other methods were shifted towards lower nominal clustering thresholds (e.g., AL clustering at 97% most similar to CL $< 97\%$). These effects are in line with the inclusive ('lumping') and exclusive ('splitting') nature of the SL and CL/UCLUST algorithms. UPARSE provided comparatively low similarities with all other methods across tested thresholds; this is likely due to UPARSE's on-the-fly chimera filtering, which removed more sequences than our UCHIME-based pipeline for the other methods – although we corrected for this effect by calculating partition similarity based only on shared sequences. Moreover, UPARSE reproducibly crashed when clustering to individual intermediate thresholds (such as 94%, 96% or 98.2%; see Appendix S1).

To test how robust clustering methods were against slight changes in similarity thresholds, we re-clustered the same data sets, but randomized the order of sequences (comparisons against 'self', diagonals in Fig. 4). When clustering twice to the same threshold, all methods provided nearly identical partitions; note that for the heuristic methods, this is probably due to forced or internal sequence sorting, rather than deterministic algorithm behaviour. However, even slight variations in threshold (increments of 0.2% corresponded to ~ 2.6 differences across 1301 alignment columns) had strong impacts on UCLUST and UPARSE. Effects on CD-HIT and SL were less drastic, while CL and AL were robust even to wider threshold variations.

We observed similar trends in robustness against varying thresholds when comparing partition similarities across methods. For UCLUST and UPARSE, similarity to other methods generally fluctuated with slightly changing clustering thresholds (vertically/horizontally 'striped' profiles in Fig. 4), whereas similarities between the other tested methods were generally more robust. At very high similarity thresholds ($\geq 99\%$), partition similarities dropped markedly for all methods, both when comparing between methods, as well as between different runs for the same method; this is likely due to very low levels of clustering at high stringencies (few sequences are clustered, most remain in singletons or very small OTUs).

As differences in total OTU counts have previously been used to benchmark clustering methods, we tested how predictive relative OTU counts were of differences in cluster composition (Spearman correlation of absolute binary log OTU count ratios, OC_{rel}, and AMI/NMI/ARI across thresholds, Table 3). We found that OC_{rel} correlated well with AMI, NMI and ARI for methods that provided generally similar partitions (e.g., AL and CD-HIT, correlation 0.858–0.945). However, for most pairwise comparisons of methods, only moderate or low correlations were observed: in particular, when comparing generally dissimilar methods, the difference in total OTU counts was a weak indicator of differences in cluster composition (e.g., SL and UPARSE, correlation 0.012–0.310). Indeed, the generally high AMI, NMI and ARI values across wide threshold ranges for some methods (in particular pairwise comparisons of AL, CL and CD-HIT) indicated that these methods provided partitions that were similar in spite of marked differences in OTU count. In other words, even though forming different total numbers of clusters, these methods tended to agree in OTU composition.

Finally, we assessed differential reproducibility between clustering methods, using partition similarities to other methods as a common reference. For all pairs of clustering methods, we correlated pairwise similarities to all other tested methods across thresholds; in other words, we asked how predictive the partition similarity of method A to a reference method X at a given threshold T was for the partition similarity of method B to method X at T% clustering. We found that AL and CD-HIT behaved the most similarly (Spearman correlation, 0.991–0.993 for AMI and NMI; ARI generally lower; Table 4), whereas both

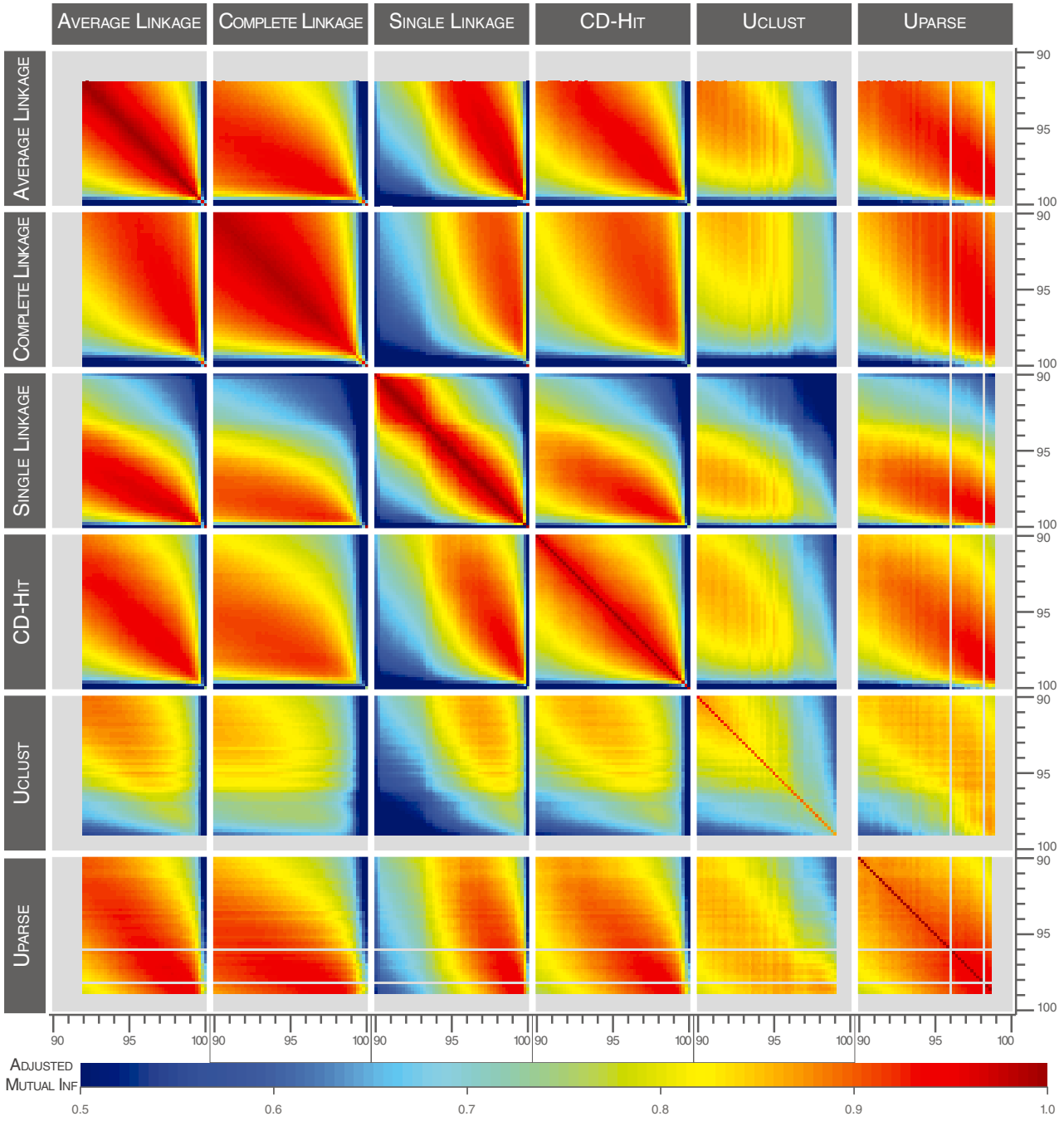


Fig. 4. Differences in cluster composition between methods across wide threshold ranges. A global data set of 887 870 16S rRNA gene sequences was clustered to thresholds ranging from 90% to 100% sequence similarity, in steps of 0.2% (corresponding to ~ 2.6 differences across the full alignment length). Pairwise partition similarities between methods and across thresholds were calculated as Adjusted Mutual Information (AMI). To calculate partition similarities of methods 'against themselves' (subplots on diagonal), clustering was re-run with randomized order of sequences. Note that algorithm memory requirements prohibited AL clustering of the full set to < 92% similarity and UCLUST/UPARSE clustering to > 99% similarity; moreover, UPARSE consistently crashed when clustering the data set to 96.0% or 98.2% similarity (grey lines in the corresponding plots; see Appendix S1). Equivalent plots on NMI and ARI similarities, and for the HSM data set, are provided as Figs S1–S5. Raw data on partition similarities are provided in Tables S4–S9.

Table 3. Differences in OTU counts are a poor predictor of differences in cluster composition.

	UPARSE	UCLUST	CD-HIT	SL	CL
	AMI *NMI* *ARI*	*AMI* *NMI* *ARI*	*AMI* *NMI* *ARI*	*AMI* *NMI* *ARI*	*AMI* *NMI* *ARI*
AL	0.884 0.852 0.147	0.716 0.820 0.363	0.945 0.944 0.858	0.869 0.826 0.748	0.607 0.575 0.234
CL	0.866 0.857 0.702	0.626 0.794 0.626	0.573 0.535 0.267	0.687 0.608 0.574	
SL	0.310 0.215 0.012	0.766 0.668 0.612	0.830 0.755 0.628		
CD-HIT	0.705 0.643 0.203	0.725 0.768 0.407			
UCLUST	0.936 0.922 0.645				

For every pair of methods, differences in OTU counts (as absolute binary log ratios) and in cluster composition (as AMI, NMI and ARI) were correlated across thresholds (Spearman rank correlation). In other words, every value in the table indicates how predictive differences in OTU counts were of (AMI/NMI/ARI) differences in cluster composition.

methods were very similar to CL (0.853–0.915 and 0.830–0.899) and UPARSE (0.918–0.965 and 0.934–0.976). Moreover, CL also correlated well with UCLUST (0.892–0.924) and UPARSE (0.813–0.825). In contrast, SL provided low, and sometimes even slightly negative, correlations to other methods. We observed similar trends when assessing absolute differences in partition similarities relative to other methods (Fig. S6). Notably, comparisons of relative OTU counts across methods provided very different correlations between methods, indicating that similar total OTU counts were generally not predictive of similar cluster composition.

A matter of perspective: sequence space and context affect OTU clustering

When comparing general trends in partition similarity for the well-defined HSM data set (Figs S1–S3) and the 'global' set of sequences (Fig. 4, Figs S4 and S5), we observed that similarities between methods depended on the data set: pairwise similarities and robustness to wide-range threshold changes were generally lower for the global set of 887 870 sequences than for the HSM. We hypothesized that these effects were due to differences in sequence space and context between the sets: the HSM set was arguably 'local', in the sense that it represented a more focused survey of microbial diversity than the comprehensive 'global' set.

To explore this hypothesis, and to quantify the differential impact of sequence context on clustering methods, we investigated two 'local' sets of sequences: (i) the HSM data set, as described above, and (ii) an artificial data set of 53 999 sequences from 18 samples of *broad ecological range* (BER; see Table S1). These sets covered very distinct sequence spaces: sequences in the HSM set shared significantly higher pairwise similarities than expected for the global set (Mann–Whitney *U*-test, $P << 10^{-16}$, Fig. 5A), whereas for the BER set, sequences were significantly *less* similar ($P << 10^{-16}$). In other words, the HSM was indeed a more 'compact' subset of the global sequence set, whereas BER sequences were more dispersed, as illustrated in the toy sequence space representation in Fig. 5.

We re-clustered both the HSM and BER sets twice – once in the presence and once in the absence of the remaining sequences from the global set. We found that

Table 4. Pairwise similarities between clustering methods, expressed as shared trends in partition similarities to other methods.

	UPARSE		UCLUST		CD-HIT		SL		CL	
	AMI *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}	*AMI* *NMI*	*ARI* OC_{rel}
AL	0.918 0.965	0.760 –0.664	0.840 0.777	0.190 –0.417	0.993 0.991	0.960 –0.017	0.806 0.599	0.481 0.399	0.915 0.853	0.479 0.738
CL	0.813 0.825	0.539 –0.383	0.892 0.924	0.595 –0.103	0.899 0.830	0.488 0.252	0.444 0.029	–0.271 0.661		
SL	0.668 0.447	0.110 –0.057	0.310 –0.043	–0.586 0.219	0.743 0.498	0.376 0.582				
CD-HIT	0.934 0.976	0.826 0.361	0.829 0.790	0.350 0.633						
UCLUST	0.754 0.811	0.465 0.736								

For every pair of methods, partition similarities (as OC_{rel}, AMI, NMI and ARI) across all methods and thresholds were correlated (Pearson correlation). For example, the value for AMI correlations between AL and CL was calculated from pairwise AMI similarities of AL to all methods across thresholds, which were correlated to AMI similarities of CL partitions across methods and thresholds. In other words, values in the table indicate how similarly the two methods behave, using partition similarities across methods and thresholds as common reference.

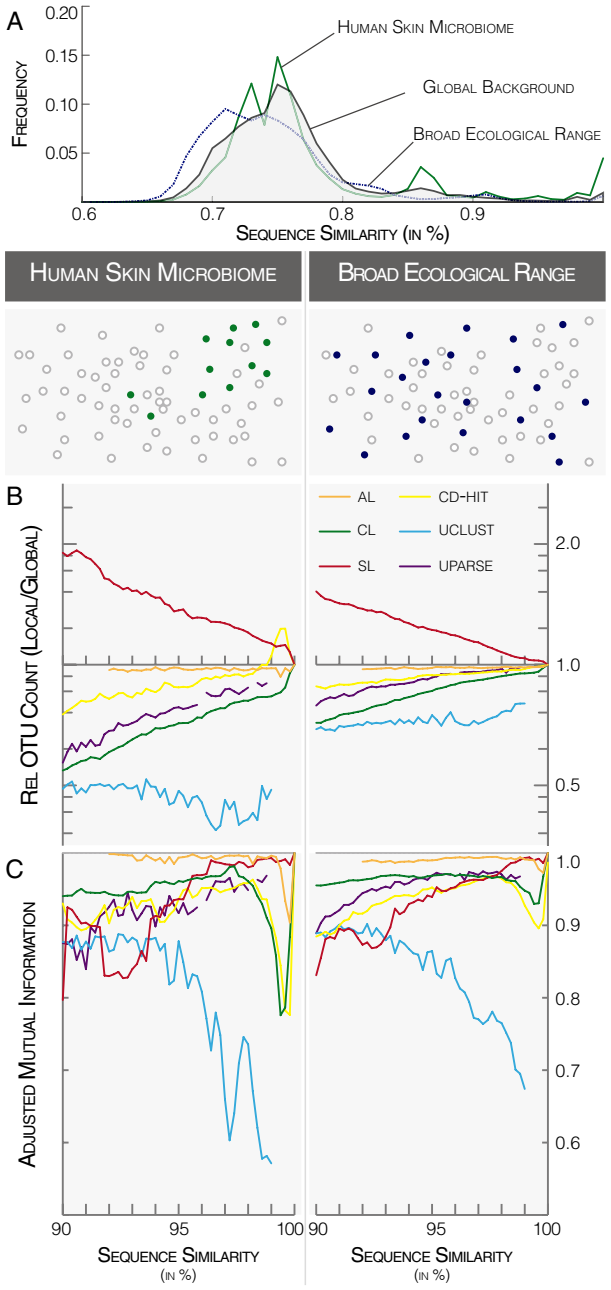


Fig. 5. Robustness to clustering context.
A. The HSM and an artificially generated data set of *broad ecological range* (BER, see Table S1) were extracted as 'local' subsets from the global set of 887 870 16S rRNA gene sequences. Pairwise internal sequence similarities were calculated based on 10 randomly drawn sets of 10 000 sequences per data set. Compared with the global background, internal sequence similarities were significantly higher for the HSM set and significantly lower for the BER set ($P << 10^{-16}$, Mann–Whitney *U*-test). This corresponds to the patterns of filled dots (HSM/BER) over grey circles (global background) in the lower panel in A.
B. Relative OTU counts when clustering the HSM and BER sets in the presence ('global context') and absence ('local context') of the full global sequence space.
C. Partition similarities between local and global context, expressed as Adjusted Mutual Information (AMI).

methods were differentially robust to clustering context, both in terms of total OTU counts (Fig. 5B) and cluster composition (Fig. 5C). Although AL, and to a lesser extent CD-HIT, provided very similar cluster counts and compositions across thresholds regardless of context (AMI ≥ 0.9), SL and in particular UCLUST were more strongly affected, providing up to twofold more (SL) or less (UCLUST) OTUs. CL provided diverging total cluster counts, but OTUs were highly similar by composition (AMI ≥ 0.94 at thresholds ≤ 98%). Generally, all methods except SL provided fewer OTUs under 'local' clustering, likely due to the absence of sequences that 'broke' OTUs into subclusters when partitioning a richer, global sequence space. In contrast, for SL, a richer context would provide 'stepping-stone' sequences, connecting OTUs by nearest neighbour similarity that remained separated under sparse, local context. Moreover, we observed that effects were generally more pronounced with decreasing thresholds (decreasing clustering stringency), except for UCLUST, which showed inverse behaviour (higher AMI towards lower thresholds). Finally, AL, CD-HIT and CL showed a pronounced drop in partition similarity at very high thresholds (≥ 99%); note that UCLUST and UPARSE did not cluster the global set to these high stringencies.

We observed that effects were generally more pronounced for the HSM than for the BER set. Thus, the interpretation of focused data sets of ecologically similar samples, such as different skin habitats, may be more susceptible to clustering context than the arguably less realistic use case of an ecologically broad set with dispersed sequence space. However, clustering context did have a significant impact in both scenarios.

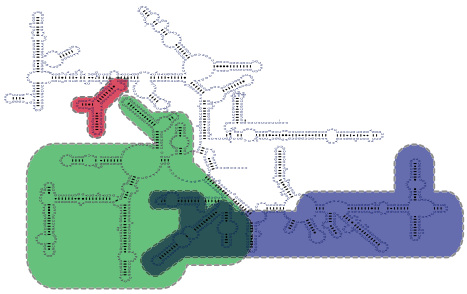
Clustering methods are differentially robust to the choice of 16S rRNA gene subregion

Many contemporary studies in microbial ecology rely on sequencing of short, hypervariable subregions of the 16S rRNA gene, rather than of the full-length molecule, mostly for reasons of throughput and cost-efficiency. To test how the choice of 16S rRNA gene subregion may affect clustering methods, we extracted data sets on subregions V2–V3, V3–V5 and V6 from the global alignment (see Experimental procedures and Fig. 6A). While V2–V3 and V3–V5 approximate the subregions used in the *Human Microbiome Project* (The Human Microbiome Project Consortium, 2012b), V6 has been a popular target in many ILLUMINA-based studies (see e.g. Huse *et al.*, 2010).

We found that partitions based on the different subregions generally diverged from clusterings of full-length 16S rRNA gene sequences. Across tested methods, V2–V3 and V6 generally provided more OTUs than

A

	Length	NT Position (Infernal)	NT Position (E. coli ref.)
FL	1301	107 - 1408	105 - 1372
V2-V3	429	107 - 536	105 - 514
V3-V5	553	378 - 931	357 - 906
V6	60	1012 - 1072	986 - 1045



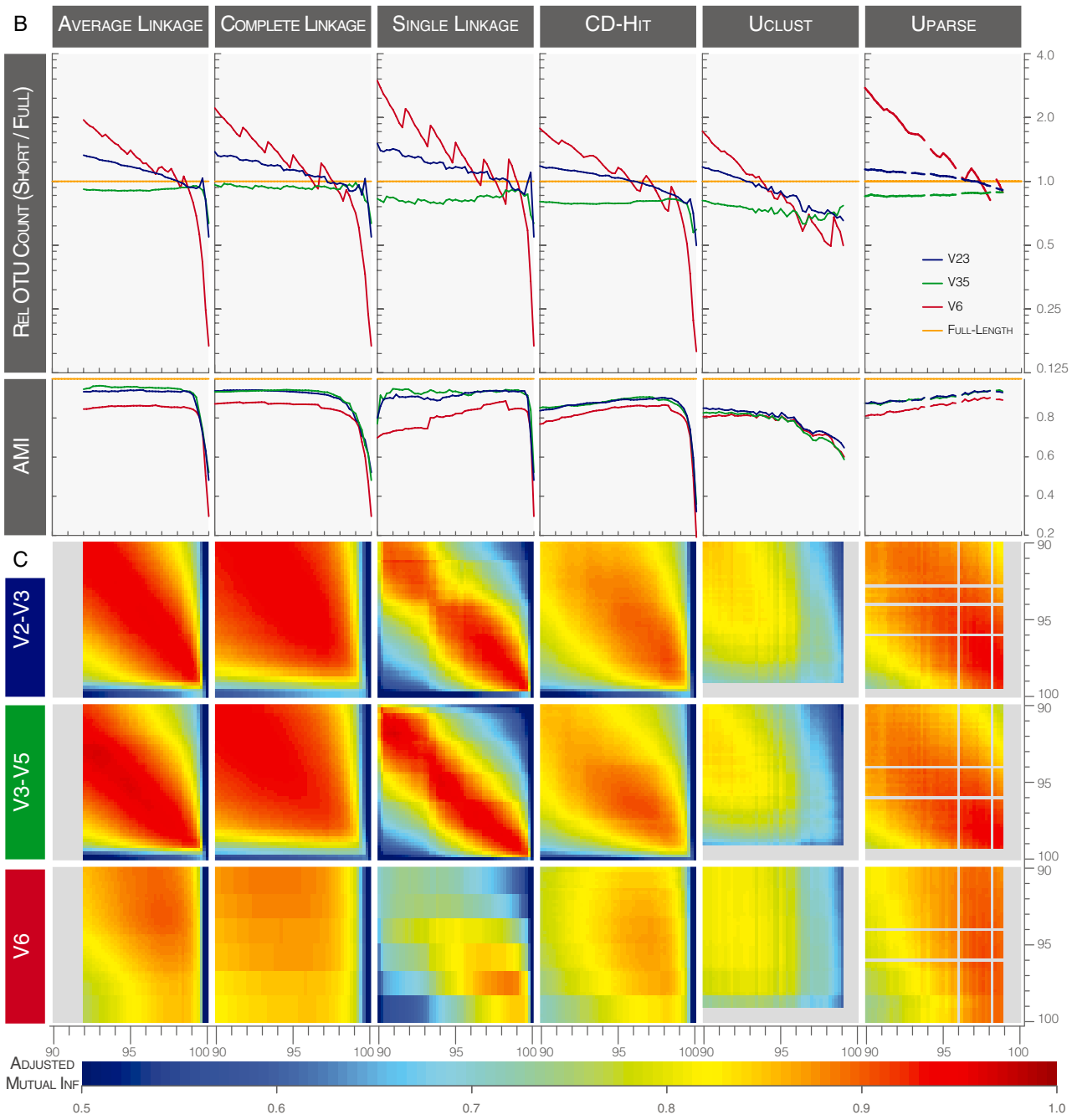


Fig. 6. Robustness to the choice of 16S rRNA gene subregion.
A. Extraction of selected hypervariable subregions from near full-length 16S rRNA gene sequence alignments ('FL'). Sequence length (left column) and nucleotide positions in the SSU-ALIGN bacterial 16S model (middle column, Nawrocki, 2009), and the *E. coli* reference 16S rRNA sequence (right column) for subregions V2–V3, V3–V5 and V6 were chosen following Schloss (2010), and indicated in the secondary structure resolved *E. coli* 16S rRNA sequence on the right (modified from an image kindly provided by Harry Noller, University of California, Santa Cruz).
B. Relative OTU counts of subregion clustering over full-length clustering and partition similarities as Adjusted Mutual Information (AMI) when clustering to the same nominal similarity threshold according to different methods.
C. Partition similarities across clustering thresholds reveal differential trends in robustness between methods.

full-length sequences at lower thresholds, but fewer OTUs at higher stringencies, whereas V3–V5 provided consistently fewer clusters (Fig. 6B). Differences in OTU count were more pronounced for V6 than V2–V3, whereas V3–V5 was generally least affected; note that these results are in line with earlier findings by Schloss (2010) on different alignment and distance calculation methods for a smaller test data set, and with findings by Kim and colleagues (2011) on differences between subregions. The notable 'spikes' in V6 relative OTU counts and the generally discrete behaviour for V2–V3 and V3–V5 were due to sequence length effects at the given resolution: e.g. for the 60 bp long V6, a single nucleotide mismatch corresponds to 1.67% sequence distance. When clustering to the same nominal threshold, V2–V3 and V3–V5 were generally more similar to full-length sequences in cluster composition (AMI $\geq$ 0.8–0.9, Fig. 6B) than V6, although partition similarity usually dropped markedly at very high thresholds ($\geq$ 99%).

Clustering methods were differentially robust to the choice of 16S rRNA gene subregion. While AL was least affected in terms of relative OTU counts, AL, CL, SL and UPARSE were the most robust in terms of cluster composition. UCLUST showed very similar (but comparatively low) partition similarities to full-length clustering for all tested subregions. When comparing partition similarities across varying thresholds (Fig. 6C), we found that AL and in particular CL were the most robust to changing stringencies, both on full-length and subregion clusterings; SL, and to a lesser extent, CD-HIT and UPARSE, were more susceptible, in particular for V6, whereas UCLUST provided generally lower similarities in cluster composition. Both UPARSE and UCLUST, and to a lesser extent SL, were susceptible to slight threshold changes in either subregion or full-length clusterings (horizontal/vertical 'stripes' in Fig. 6C).

Thus, in spite of notable differences in total cluster counts, some tested methods were surprisingly robust to the choice of 16S rRNA gene subregion in terms of cluster composition, in particular AL and CL. In other words, even when using shorter reads (containing less information), OTUs were often composed overall similarly, although there were clear differences between tested methods.

Discussion

Reproducibility of results is paramount to any empirical field of research. Scientific findings are generally required to be robust to the choice of experimental approach, and 'true' phenomena should be observable using independent methodologies, within reasonable limits. Drummond (2009) has formalized this notion by pointing out a conceptual distinction between the *replicability* of results and the *reproducibility* of findings. He contends that although the latter is a necessary prerequisite of scientific endeavour, the former is indeed less instructive. In other words, the exact replication of an experiment *ceteris paribus* (i.e., replicability) is less informative than the corroboration of findings by reproduction in an independent setup (i.e., reproducibility, Casadevall and Fang, 2010).

We believe that these considerations are highly relevant to microbial ecology – which is currently not only an empirical, but indeed a data-driven research field. In this study, we have focused on the reproducibility of OTU demarcation from complex sequencing data sets: when repeating an experiment under different sequence clustering parameters, how much bias is introduced simply by the choice of methodology? In other words, how robust are biological findings to the choice of clustering method? We found that OTU demarcation may indeed be *replicable*: different methods provided (almost) identical partitions when twice clustering the exact same sets of sequences, but in randomized order (Fig. 4, diagonals). However, trends in *reproducibility* were less clear.

We quantified the variability in OTU demarcation on various complementary levels. In a first, basic approach, we confirmed previous observations on diverging total OTU counts and cluster size distributions between methods (Fig. 1). However, total cluster counts are a summary statistic of limited biological significance with respect to OTU composition (Table 3) and higher-level ecological data descriptions (Fig. 3, Table 2). Rather, differences between clusterings are more meaningfully described as differences in cluster *composition*. We explored these on an anecdotal level for an exemplary data point (Fig. 2), for which we also quantified biases to

higher-level ecological data descriptions (Fig. 3). We then generalized our observations to a global data set, and to the choice of clustering threshold (Fig. 4, Figs S1–S5), clustering context (Fig. 5) and targeted 16S rRNA gene subregion (Fig. 6).

When viewed across all these tests, hierarchical AL and heuristic CD-HIT clustering were generally the most similar pair of methods. This is both surprising and remarkable, as CD-HIT relies on several computationally efficient shortcuts that are expected to reduce accuracy. Both AL and CD-HIT also showed generally similar behaviour to CL. Similarities in cluster composition among these three methods were robust to (wide) changes in clustering threshold, indicating that these methods provided surprisingly reproducible clusterings. These high levels of similarity among AL, CL and CD-HIT are remarkable, in particular when considering that these methods diverged considerably in terms of total OTU counts across thresholds.

In contrast, SL, UCLUST and UPARSE diverged more strongly in their behaviour from all other methods. Indeed, the 'inclusive' SL algorithm is a conceptual outlier in the tested set of methods, as it implements a fundamentally different clustering regime than the other, more 'exclusive' methods. Similarly, UPARSE can be considered an outlier, as it implements adaptive on-the-fly chimera filtering and effectively clustered different sets of sequences than the other tested methods. Indeed, UPARSE filtering for chimeric sequences was far more restrictive than the UCHIME-based protocol that was used for the other five methods: depending on the clustering threshold, UPARSE removed ~ 30–50% of sequences as 'chimeric' from the global data set (compared with ~ 20% for the UCHIME-based pipeline; Fig. S7). This is surprising when considering that the input data were full-length, high-quality and often curated or pre-filtered sequences; indeed, even for the 16S rRNA gene sequences deposited to the REFSEQ database as part of complete genomes, UPARSE flagged up to 31.7% as 'chimeric', depending on the clustering threshold (2.6% for UCHIME; Fig. S7). Notably, UPARSE-filtered sets for all tested thresholds were entirely contained in the UCHIME-filtered set (all sequences retained by UPARSE were also retained by UCHIME; '$\text{UPARSE} \subseteq \text{UCHIME}$'). Thus, lower similarities of UPARSE relative to other methods in cluster composition, as well as in overall behaviour could be expected – in particular, when also considering the observations on data set scope and clustering context (Fig. 5). In contrast, although conceptually related to CD-HIT and CL, UCLUST clearly diverged across many tests: with respect to other methods, it provided significantly shifted diversity estimates (Fig. 3) and deviating cluster composition, in particular at higher (biologically more relevant) clustering thresholds (Fig. 4).

Although it is tempting to interpret these findings in terms of *cluster quality*, we note that a notion of 'true' or 'false' clustering requires more than pairwise comparisons between methods, and has been addressed elsewhere, by us and others (e.g., Sun *et al.*, 2011; Koeppel and Wu, 2013; Schmidt *et al.*, 2014). Rather, partition similarities across thresholds may inform the comparison of results across studies, as they allow an assessment of the bias introduced by clustering method. In contrast, the observed trends in robustness to changing parameters may indeed be interpreted in either way: in terms of comparability across studies, but also as (quality) attributes of clustering methods. In particular, CL, AL and CD-HIT were surprisingly robust to changes in clustering threshold, clustering context and the choice of subregion, whereas SL, UPARSE and especially UCLUST were more strongly affected. By design, previous unidimensional benchmarking studies focusing on different concepts of partition 'optimality' did not capture these trends in robustness between methods.

There have been great efforts to address the reproducibility issue through increased levels of standardization: software pipelines such as MOTHUR or QIIME provide comprehensive protocols to analyse microbial ecology data sets. However, these efforts have arguably enhanced *replicability* rather than *reproducibility*, by providing widely adopted defaults. Furthermore, we note that QIIME relies on UCLUST as default clustering method – the method that was consistently the most sensitive to any parameter change across our tests. Thus, while QIIME aims to enhance comparability of findings across studies, at the level of sequence clustering, it probably achieves the reverse. In contrast, the default clustering method implemented in MOTHUR is AL, which was among the most robust methods in our tests.

Reference-based OTU demarcation is another approach to standardization, which has recently received increasing attention: sequences are mapped to pre-clustered reference sets of curated 16S rRNA gene sequences, provided e.g. by the RIBOSOMAL DATABASE PROJECT (Cole *et al.*, 2013), GREENGENES (DeSantis *et al.*, 2006) and SILVA (Yilmaz *et al.*, 2013) databases. We note that the global data set used in our study closely resembles these reference sets in size, scope, sequence length and pre-processing. Thus, our results are potentially relevant for the generation of reference OTU sets and when designing reference-based OTU demarcation frameworks: the 'quality' of reference-picked OTUs directly depends on the quality of pre-clustering of the reference set. The GREENGENES and SILVA databases, which are the default reference sets in QIIME and for the Earth Microbiome Project (Gilbert *et al.*, 2010), are pre-clustered to 97% and 99% similarity using UCLUST. Moreover, one main difference between reference-based OTU

demarcation and *de novo* clustering is arguably clustering context – an effect that has previously been ignored or underestimated.

In view of the many parameter choices in sequence processing pipelines, how can reproducibility of results be enhanced in practice? Based on our findings, we suggest that researchers may want to resort to deliberately redundant study/analysis designs. Several recent studies, e.g. the *Human Microbiome Project*, have indeed relied on complementary analysis pipelines, but this is usually not the case. Redundancy may be introduced at many levels, e.g. in the choice of sequenced rRNA gene subregions, and at every level of sequence processing, and we recommend that researchers implement at least two complementary analysis pipelines. Biological findings that are robust to independent methodologies are arguably more dependable than any single-track analysis.

Experimental procedures

Sequence data and preprocessing

We generated a comprehensive global 16S rRNA gene sequence data set as described previously (Schmidt *et al.*, 2014); see Appendix S1 for further details. In short, we parsed available full-length 16S rRNA gene sequences from National Center for Biotechnology Information (NCBI) GENBANK (Benson *et al.*, 2013) and from the genomes available in the NCBI Reference Sequence Database (REFSEQ, Pruitt *et al.*, 2011). After removing ~ 20% of total sequences that were flagged as chimeric by UCHIME (Edgar *et al.*, 2011), we aligned the remaining sequences to a reference model for the bacterial 16S rRNA gene (provided in the package SSU-ALIGN, Nawrocki, 2009) using INFERNAL (Nawrocki *et al.*, 2009) and pruned away any terminal nucleotides that aligned outside of two manually chosen, well-conserved start and end positions. After these steps, our data set comprised 887 870 aligned, near full-length bacterial 16S rRNA gene sequences; the filtered data set is available online (http://meringlab.org/suppdata/2014-otu_robustness/).

From this global data set, we extracted two smaller, 'local' data sets for in-depth analyses, the HSM data set (Grice *et al.*, 2009), comprising 90 620 sequences after filtering and alignment; and an artificial data set of broad ecological range (BER), combining 53 999 sequences from 18 studies focusing on distinct, unrelated environments (see Table S1).

Moreover, we generated three global data sets of *'short read'* sequences, by extracting subregions V2–V3 (pos 107–536 in the INFERNAL model, length 429 nt, corresponding to 105–514 in the *Escherichia coli* 16S gene sequence; *E. coli* reference positions as used by Schloss, 2010), V3–V5 (378–931, length 553 nt, *E.coli* 357–906) and V6 (1012–1072, length 60 nt, *E. coli* 986–1045).

Sequence clustering into OTUs

We clustered sequences into OTUs using three hierarchical clustering algorithms (average, complete and single linkage) and three heuristic methods (CD-HIT, UCLUST and UPARSE). For every method, we clustered to varying thresholds between 90% and 100% sequence identity (in steps of 0.2%; 92–100% for AL, 90–99% for UCLUST and UPARSE; see Appendix S1). We generated OTU sets using CD-HIT (version 4.5.4, Build 2012-08-25, Fu *et al.*, 2012) in CD-HIT-EST mode (default for the CD-HIT-OTU pipeline) from unaligned sequences using standard parameters. The UCLUST (version 6.0.307, Edgar, 2010) series of OTU sets were generated from unaligned sequences using the UCLUST software with the *cluster_fast* option and standard parameters. As UPARSE (Edgar, 2013) combines OTU clustering with on-the-fly filtering for chimeric sequences, we used the full, unaligned, non-chimera-filtered sequence data set for UPARSE runs and subsequently mapped shared sequences between UPARSE partitions and the UCHIME-filtered data set used for the other clustering methods. Hierarchical AL, CL and SL clustering were performed using our recently developed software package HPC-CLUST (Matias Rodrigues and von Mering, 2014), using the 'onegap' sequence distance calculator (counting gaps as single mismatches). See Appendix S1 for additional details and parameters.

Assessing OTU set similarity

We assessed pairwise similarities between OTU sets using three distinct measures, namely Normalized Mutual Information (NMI) (Fred and Jain, 2003), Adjusted Mutual Information (AMI) (Vinh *et al.*, 2010) and the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985); see Appendix S1 for further details. All three measures quantify the similarity in *cluster composition* between partitions (OTU sets): AMI, NMI and ARI values of 1 indicate perfectly identical clusterings.

Acknowledgements

We thank Mark Robinson, University of Zürich, for insightful discussions and for pointing us to the Adjusted Rand Index, as well as Damian Szklarczyk and Alexander Roth for helpful discussions during the preparation of the manuscript. This work was supported by an ERC starting grant (UMICIS/242870) and by the Swiss National Science Foundation (31003A_135688). The authors declare that no conflict of interests exists.

References

Achtman, M., and Wagner, M. (2008) Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* **6:** 431–440.

Barriuso, J., Valverde, J.R., and Mellado, R.P. (2011) Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics* **12:** 473.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013) GenBank. *Nucleic Acids Res* **41:** D36–D42.

Bonder, M.J., Abeln, S., Zaura, E., and Brandt, B.W. (2012) Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* **28:** 2891–2897.

Bray, J.R., and Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr* **27:** 325–349.

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., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7:** 335–336.

Casadevall, A., and Fang, F.C. (2010) Reproducible science. *Infect Immun* **78:** 4972–4975.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scand J Stat* **11:** 265–270.

Chao, A., Chazdon, R.L., Colwell, R.K., and Shen, T.-J. (2004) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* **8:** 148–159.

Chen, W., Cheng, Y., Zhang, C., Zhang, S., and Zhao, H. (2013) MSClust: a multi-seeds based clustering algorithm for microbiome profiling using 16S rRNA sequence. *J Microbiol Methods* **94:** 347–355.

Chen, W., Zhang, C.K., Cheng, Y., Zhang, S., and Zhao, H. (2013) A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS ONE* **8:** e70837.

Cheng, L., Walker, A.W., and Corander, J. (2012) Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Res* **40:** 5240–5249.

Cohan, F.M. (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc Lond B Biol Sci* **361:** 1985–1996.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., *et al*. (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42:** D633–D642.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72:** 5069–5072.

Dice, L.R. (1945) Measures of the amount of ecologic association between species. *Ecology* **26:** 297–302.

Doolittle, W.F., and Papke, R.T. (2006) Genomics and the bacterial species problem. *Genome Biol* **7:** 116.

Doolittle, W.F., and Zhaxybayeva, O. (2009) On the origin of prokaryotic species. *Genome Res* **19:** 744–756.

Drummond, C. (2009) *Replicability is Not Reproducibility: Nor is It Good Science*. Proc Eval Meth Mach Learn Workshop 26th ICML. Montreal, Quebec, Canada.

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

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10:** 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194–2200.

Fred, A., and Jain, A.K. (2003) Robust data clustering. In *Proc IEEE Conference Comp Vision Pattern Recognition*. Madison, WI, USA: IEEE Computer Society, pp. II/128–II/133.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28:** 3150–3152.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., *et al*. (2005) Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3:** 733–739.

Gilbert, J., Meyer, F., Antonopoulos, D.A., Balaji, P., Brown, C.T., Brown, C.T., *et al*. (2010) Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand Genomic Sci* **3:** 243–248.

Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., *et al*. (2009) Topographical and temporal diversity of the human skin microbiome. *Science* **324:** 1190–1192.

Hao, X., Jiang, R., and Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* **27:** 611–618.

Horn, H.S. (1966) Measurement of 'overlap' in comparative ecological studies. *Amer Naturalist* **100:** 419–424.

Hubert, L., and Arabie, P. (1985) Comparing partitions. *J Classification* **2:** 193–218.

Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12:** 1889–1898.

Kim, M., Morrison, M., and Yu, Z. (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods* **84:** 81–87.

Koeppel, A., Perry, E.B., Sikorski, J., Krizanc, D., Warner, A., Ward, D.M., *et al*. (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci USA* **105:** 2504–2509.

Koeppel, A.F., and Wu, M. (2013) Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res* **41:** 5175–5188.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82:** 6955–6959.

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.

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform* **13:** 656–668.

Matias Rodrigues, J.F., and von Mering, C. (2014) HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* **30:** 287–288.

Nawrocki, E.P. (2009) Structural RNA homology search and alignment using covariance models. PhD Thesis. St Louis, USA: Washington University School of Medicine.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25:** 1335–1337.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40:** 337–365.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M.A., Smirnova, T., Nosrat, B., *et al*. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40:** D571–D579.

Preheim, S.P., Perrotta, A.R., Martin-Platero, A.M., Gupta, A., and Alm, E.J. (2013) Distribution-based clustering: using ecology to refine the Operational Taxonomic Unit. *Appl Environ Microbiol* **79:** 6593–6603.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40:** D130–D135.

Rasheed, Z., Rangwala, H., and Barbará, D. (2013) 16S rRNA metagenome clustering and diversity estimation using locality sensitive hashing. *BMC Syst Biol* **7** (Suppl. 4)**:** S11.

Schloss, P.D. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* **6:** e1000844.

Schloss, P.D. (2012) Secondary structure improves OTU assignments of 16S rRNA gene sequences. *ISME J* **7** (Suppl. 4)**:** 511.

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., Westcott, S.L., Rabyn, T., Hall, J.R., Hartmann, M., Hollister, E.B., *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., Gevers, D., and Westcott, S.L. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6:** e27310.

Schmidt, T.S.B., Matias Rodrigues, J.F., and von Mering, C. (2014) Ecological consistency of SSU rRNA-based Operational Taxonomic Units at a global scale. *PLoS Comput Biol* **10:** e1003594.

Shannon, C.E. (1948) A mathematical theory of communication. *AT&T Tech J* **27:** 623–656.

Simpson, E.H. (1949) Measurement of diversity. *Nature* **163:** 688–688.

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M.L., McKendree, W., and Farmerie, W. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37:** e76.

Sun, Y., Cai, Y., Huse, S.M., Knight, R., Farmerie, W.G., Wang, X., and Mai, V. (2011) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform* **13:** 107–121.

The Human Microbiome Project Consortium (2012a) A framework for human microbiome research. *Nature* **486:** 215–221.

The Human Microbiome Project Consortium (2012b) Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207–214.

Vinh, N.X., Epps, J., and Bailey, J. (2010) Information theoretic measures for clustering comparison: variants, properties, normalization and correction for chance. *J Mach Learn Res* **11:** 2837–2854.

Wang, X., Cai, Y., Sun, Y., Knight, R., and Mai, V. (2011) Secondary structure information does not improve OTU assignment for partial 16S rRNA sequences. *ISME J* **6:** 1277–1280.

Wang, X., Yao, J., Sun, Y., and Mai, V. (2013) M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformatics* **14:** 43.

Wei, D., Jiang, Q., Wei, Y., and Wang, S. (2012) A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics* **13:** 174.

White, J.R., Navlakha, S., Nagarajan, N., Ghodsi, M.-R., Kingsford, C., and Pop, M. (2010) Alignment and clustering of phylogenetic markers – implications for microbial diversity studies. *BMC Bioinformatics* **11:** 152.

Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., *et al*. (2013) The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. *Nucleic Acids Res* **42:** D643–D648.

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* **29:** 2869–2876.

Zheng, Z., Kramer, S., and Schmidt, B. (2012) DySC: software for greedy clustering of 16S rRNA reads. *Bioinformatics* **28:** 2182–2183.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Adjusted Mutual Information (AMI) between methods across thresholds when clustering the HSM data set. Equivalent to Fig. 4 in the main text. Raw AMI values provided in Table S4.

Fig. S2. Normalized Mutual Information (NMI) between methods across thresholds when clustering the HSM data set. Equivalent to Fig. 4 in the main text. Raw NMI values provided in Table S5.

Fig. S3. Adjusted Rand Index (ARI) between methods across thresholds when clustering the HSM data set. Equivalent to Fig. 4 in the main text. Raw ARI values provided in Table S6.

Fig. S4. Normalized Mutual Information (NMI) between methods across thresholds when clustering the global data set of 887 870 16S rRNA sequences. Equivalent to Fig. 4 in the main text. Raw NMI values provided in Table S8.

Fig. S5. Adjusted Rand Index (ARI) between methods across thresholds when clustering the global data set of 887 870 16S rRNA sequences. Equivalent to Fig. 4 in the main text. Raw ARI values provided in Table S9.

Fig. S6. Pairwise similarities between clustering methods, expressed as absolute differences in partition similarities to other methods. For every pair of clustering methods, differences in partition similarities (expressed as AMI) to all methods across thresholds are shown as histograms. For example, the top left subgraph shows differences between AL and all other methods; it shows that CD-HIT and AL provided very similar AMI values against other methods across thresholds, although CD-HIT AMI values tended to be slightly lower

(peak shifted to the left). In other words, the subplots indicate how similarly the pairs of methods behaved, using partition similarities to other methods across thresholds as reference.

Fig. S7. Differential filtering for 'chimeric' sequences by UCHIME and UPARSE. The 16S rRNA gene sequence data set used in this study was filtered for chimeric sequences using two different protocols, based on UCHIME and UPARSE (see Experimental procedures and Appendix S1). For the UCHIME workflow, filtering was performed in '*uchime_ref*' mode, using a custom reference database of non-chimeric sequences, generated directly from the full set of unfiltered 16S rRNA sequences (see Appendix S1). This custom reference database was tailored to the present sequence data set, thus allowing for more stringent chimera checking than general-purpose databases such as the frequently used GOLD database (Pagani *et al*., 2012; http://www.genomesonline.org). UPARSE implements similarity threshold-dependent on-the-fly chimera filtering at the time of clustering, followed by UCHIME filtering of cluster seed sequences (see Appendix S1 for detailed parameters). We observed that for all tested thresholds, UPARSE-filtered sets were entirely contained within the UCHIME-filtered sequence set ('UPARSE $\subseteq$ UCHIME').
A. The global, unfiltered data set contained 6760 16S rRNA gene sequences from fully sequenced genomes deposited in the curated NCBI REFSEQ database. Out of these, UCHIME retained 97.4% as 'non-chimeric' when filtering against our custom reference database and 99.6% when filtering against the GOLD database. In contrast, UPARSE retained 53.5–76% of reads, depending on the clustering threshold.
B. For the entire, global data set, UCHIME retained 80.3% 'non-chimeric' sequences when filtering against the custom database and 97.2% against the GOLD database, whereas UPARSE retained 48.2–70.9%, depending on the threshold.

Table S1. Composition of an artificial 'local' data set of *broad ecological range* (BER). A total of 53 999 16S rRNA gene sequences from 18 studies were selected to generate the data set; additional information and detailed references are provided in the table.

Table S2. Total OTU counts per method when clustering a global data set of 887 870 16S rRNA sequences. OTU counts are given per method for different thresholds. For UPARSE, the respective number of clustered sequences that mapped to the UCHIME-filtered data set is also provided (as UPARSE implements differential on-the-fly chimera filtering, removing different sets of sequences at different clustering thresholds; see also Fig. S7).

Table S3. Estimates of α- and β-diversity when clustering the HSM data set to 97% sequence similarity according to different methods. The table provides raw data of diversity estimates across 21 skin sites for every method. Moreover, trends between methods (Pearson correlation of diversity estimates), absolute shifts (as binary log-ratio) and shift significance (as *P*-values in one-sided Mann–Whitney *U*-tests) are also provided for every index. Finally, an overview of the HSM data set (sequence counts and internal sequence similarities per habitat), as well as pairwise partition similarities between methods at 97% clustering are provided.

Table S4. Adjusted Mutual Information (AMI) between methods across thresholds when clustering the HSM data set. Values as shown in Fig. S1.

Table S5. Normalized Mutual Information (NMI) between methods across thresholds when clustering the HSM data set. Values as shown in Fig. S2.

Table S6. Adjusted Rand Index (ARI) between methods across thresholds when clustering the HSM data set. Values as shown in Fig. S3.

Table S7. Adjusted Mutual Information (AMI) between methods across thresholds when clustering the global data set of 887 870 16S rRNA sequences. Values as shown in Fig. 4 in the main text. Missing values for AL at thresholds $< 92\%$ and UCLUST/UPARSE $> 99\%$ as clustering to these thresholds was prohibited by memory requirements (see Appendix S1).

Table S8. Normalized Mutual Information (NMI) between methods across thresholds when clustering the global data set of 887 870 16S rRNA sequences. Values as shown in Fig. S4. Missing values for AL at thresholds $< 92\%$ and UCLUST/UPARSE $> 99\%$ as clustering to these thresholds was prohibited by memory requirements (see Appendix S1).

Table S9. Adjusted Rand Index (ARI) between methods across thresholds when clustering the global data set of 887 870 16S rRNA sequences. Values as shown in Fig. S5. Missing values for AL at thresholds $< 92\%$ and UCLUST/UPARSE $> 99\%$ as clustering to these thresholds was prohibited by memory requirements (see Appendix S1).

Appendix S1. Supplementary methods.