environmental microbiology

Environmental Microbiology (2019) **21**(11), 4109–4124 doi:10.1111/1462-2920.14764

Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants

Dominik Forster,[1] Guillaume Lentendu,[1] Sabine Filker,[2] Elyssa Dubois,[1] Thomas A. Wilding[3] and Thorsten Stoeck[1*]

[1]*Department of Ecology, University of Kaiserslautern, Kaiserslautern, Germany.*

[2]*Department of Molecular Ecology, University of Kaiserslautern, Kaiserslautern, Germany.*

[3]*Scottish Association for Marine Science, Scottish Marine Institute, Oban, Scotland, UK.*

Summary

Effective and precise grouping of highly similar sequences remains a major bottleneck in the evaluation of high-throughput sequencing datasets. Amplicon sequence variants (ASVs) offer a promising alternative that may supersede the widely used operational taxonomic units (OTUs) in environmental sequencing studies. We compared the performance of a recently developed pipeline based on the algorithm DADA2 for obtaining ASVs against a pipeline based on the algorithm SWARM for obtaining OTUs. Illumina-sequencing of 29 individual ciliate species resulted in up to 11 ASVs per species, while SWARM produced up to 19 OTUs per species. To improve the congruency between species diversity and molecular diversity, we applied sequence similarity networks (SSNs) for second-level sequence grouping into network sequence clusters (NSCs). At 100% sequence similarity in SWARM-SSNs, NSC numbers decreased from 7.9-fold overestimation without abundance filter, to 4.5-fold overestimation when an abundance filter was applied. For the DADA2-SSN approach, NSC numbers decreased from 3.5-fold to 3-fold overestimation. Rand index cluster analyses predicted best binning results between 97% and 94% sequence similarity for both DADA2-SSNs and SWARM-SSNs. Depending on the ecological questions addressed in an environmental sequencing study with protists we recommend ASVs as replacement for OTUs, best in combination with SSNs.

Received 4 March, 2019; revised 25 July, 2019; accepted 25 July, 2019. *For correspondence. E-mail stoeck@rhrk.uni-kl.de; Tel. +49 631 2052502; Fax +49 631 205 2496.

Introduction

Ever since high-throughput-sequencing (HTS) has been introduced in molecular ecology, researchers have been looking for effective tools to evaluate the resulting sequence data in the context of diversity measures. The standard way of addressing this issue is to group sequencing reads obtained for example from environmental samples, into operational taxonomic units (OTUs; e.g. de Vargas *et al*. 2015; Stoeck *et al*. 2010). This grouping can be achieved either by relying on global clustering scores (Schloss *et al*., 2009; Edgar, 2010; Fu *et al*., 2012; Rognes *et al*., 2016) or on local clustering scores (Mahé *et al*., 2015). Although traditional clustering relies on fixed global clustering thresholds expressed, for example, as sequence similarity between aligned sequences, local clustering allows for a more fine-tuned evaluation by comparing all local pairs of nucleotides between the sequences and iteratively grouping them into OTUs. It is well known, though, that every kind of OTU is at best a bioinformatical approximation of a species (Schloss and Westcott, 2011; Tikhonov *et al*., 2015) and that we are far away from the one OTU – one species ideal. Currently available sequence grouping methods tend to allocate sequencing reads of the same species into multiple OTUs. Even though sequencing errors (Kunin *et al*., 2010) and intraspecific genetic heterogeneity may also contribute to diversity inflation in environmental HTS datasets, imprecise sequence grouping is the main cause of severe biodiversity overestimations (Flynn *et al*., 2015; Clare *et al*., 2016). Likewise, imprecise sequence grouping may also lead to biodiversity underestimations, when sequences of different species are grouped into the same OTU (Bachy *et al*., 2013; Grattepanche *et al*., 2014). This emphasizes the need for a more accurate grouping of sequences for providing more realistic estimates of species richness and diversity within a sample.

Recently, Callahan and colleagues (2017) presented an alternative approach by replacing OTUs with amplicon sequence variants (ASVs). The authors developed the open-source software package DADA2 (Callahan *et al*., 2016) to model and correct Illumina-sequenced amplicon errors. Bioinformatic tools that follow a similar denoising approach had already been introduced for 454 pyrosequencing datasets (Quince *et al*., 2009;

Reeder and Knight, 2010). Several terminologies for the description of these tools can be found. We refer to them as 'first-level sequence grouping algorithms', although 'pre-clustering algorithms' is also commonly found in the literature (e.g. Schloss *et al*., 2011). They rely on statistical model-based evaluation of HTS data to infer which sequence base-call differences represent true biological variants and which represent sequencing artefacts (Callahan *et al*., 2017; Knight *et al*., 2018). The quality filtering models allow resolutions down to single nucleotide differences between sequences and may thus affiliate error-prone sequences with an existing ASV. Thus, even though designed as a denoising tool, DADA2 is at the same time an elegant way of sequence grouping. Assigning sequences to existing ASVs is of further importance when samples of different studies are analysed in the same context. Since ASVs are supposed to be consistent biological entities, they provide high levels of reproducibility and comparability across independent studies (Callahan *et al*., 2016, 2017; Amir *et al*., 2017). Initial studies comparing ASVs against OTUs obtained by global clustering approaches from the same samples support these and other advantages of ASVs (Callahan *et al*., 2016; Utter *et al*., 2016; Allali *et al*., 2017; Needham *et al*., 2017; Nearing *et al*., 2018; Zoqratt *et al*., 2018). One congruent finding of all these studies was that distinctively fewer ASVs than OTUs were produced from the same samples, regardless of the sampled habitat. Furthermore, when comparing OTU clustering methods with ASV approaches, the latter could much more accurately reproduce a known diversity from mock communities (Callahan *et al*., 2016; Nearing *et al*., 2018; Xue *et al*., 2018).

Thus far, ASVs have never been directly compared with OTUs obtained from local clustering approaches. One of the currently most widely used local clustering algorithms is SWARM (Mahé *et al*., 2014, 2015). In contrast to heuristic global clustering algorithms, SWARM relies on single-linkage clustering and is input-order-independent, which results in much more robust OTU calling. Furthermore, its high clustering stringency allows separation of even highly similar sequences (Mahé *et al*., 2014). This stringency is reflected by a small local clustering threshold that is by default set to one nucleotide difference between two aligned sequences. SWARM-OTUs are created one after the other by adding sequences in an iterative process. As long as sequences with equal or less than the set nucleotide difference to any sequence already grouped into the OTU remain in the dataset, these sequences will be added and the OTU will not be closed. That is, SWARM does not only avoid the disadvantages of greedy heuristic global clustering algorithms, but also enables a very fine-grained grouping of sequences. This generates distinct OTUs that differ in as little as two nucleotides from one another. These advantages were recognized by several important studies on microbial communities, which have further demonstrated that SWARM scales exceptionally well even to the largest HTS datasets available to date (de Vargas *et al*., 2015; Mahé *et al*., 2017).

To further improve the congruency between species diversity and molecular diversity within a sample, several authors have proposed a two-level sequence grouping approach, employing either ASVs (Anslan *et al*., 2018; Jusino *et al*., 2018; Palmer *et al*., 2018) or OTUs (Forster *et al*., 2016) as a first-level of sequence grouping. For instance, single-linkage first-level sequence grouping followed by a subsequent second round of sequence grouping emerged as the most accurate strategy for defining OTUs in a study by Huse and colleagues (2010). The goal of this combined two-level sequence grouping approach is to fine-tune the obtained ASVs or OTUs to further improve biodiversity estimates for accurate species richness predictions within a sample. A very promising example for a second-level sequence grouping includes sequence similarity networks (SSNs; Forster *et al*., 2015). Sequence grouping in SSNs is achieved via pairwise sequence similarity scores. Two sequences are connected by an edge, if their similarity passes a defined value. A group of sequences connected in such a manner forms one connected component within the network, which is further interpreted as a network sequence cluster (NSC), the result of second-level sequence grouping in SSNs. Since the network approach is based on concepts from graph theory, there exists a full mathematical toolkit to evaluate central ecological and evolutionary theories (e.g. Forster *et al*., 2015; Corel *et al*., 2016; Lord *et al*., 2016). But it was not until recently that the strength of SSNs could be exploited for large HTS datasets. The all-versus-all pairwise sequence alignments on which the approach relies are time-intensive and computationally demanding (Bik *et al*., 2012; Sun *et al*., 2012) and the computational power as well as the tools became available for routine analyses only in the past years. After applying a first level of sequence grouping (OTUs or ASVs), a dataset is represented by fewer sequences and computational demands will exponentially decrease. Using this groundwork, the trade-off between computational demands and scientific benefits becomes less disadvantageous when SSNs are deployed as a second level of sequence grouping. Therefore, this strategy could enable SSN analyses of even the largest datasets while allowing a more in-depth analyses of OTUs than possible with heuristic clustering algorithms.

Based on the available knowledge summarized above, we hypothesized that (i) DADA2-derived ASVs produce fewer molecular sequence clusters from individual protist species and that (ii) sequence similarity networks as a

second-level sequence grouping approach further improve the congruency between species diversity and molecular diversity as revealed by HTS sequencing. Thus far, these hypotheses have gone untested. If verified, DADA2 in combination with SSNs may allow for better biodiversity estimates in molecular environmental diversity surveys and, consequently, for less biased interpretations of diversity results obtained in such studies.

Our case study is based on a collection of 29 individual ciliate species, each of which was independently processed from cultivation to Illumina sequencing of the V9 18S rDNA region and data analysis. Ciliates were chosen as model organisms for unicellular eukaryotes, because they possess morphological features that largely allow for clear species differentiation (but see also e.g. Kumar and Foissner, 2016 for cryptic ciliate species) and because bioinformatic delineation of ciliate species has been thoroughly validated (e.g. Nebel *et al*., 2011; Dunthorn *et al*., 2014). Using the resulting HTS datasets, we directly compared first-level sequence grouping by an ASV-producing pipeline based on the algorithm DADA2, against first-level sequence grouping by an OTU-producing pipeline based on the algorithm SWARM and assessed the degree to which both algorithms matched the expected diversity (species richness of 29). For both algorithms we followed standard pipelines recommended by the respective developers, available at https://benjjneb.github.io/dada2/index.html for DADA2 and at https://github.com/frederic-mahe/swarm/wiki/Fred's-metabarcoding-pipeline for SWARM. We then applied sequence similarity networks as a second-level sequence grouping approach to DADA2-derived ASV and SWARM-derived OTU results to further improve the degree to which the expected species diversity can be obtained.

Results

First-level sequence grouping

Sequence grouping of the 29 species-specific ciliate datasets with the DADA2 pipeline resulted in 101 different ASVs (Table 1). This represented a 3.5-fold overestimation of the species diversity in the samples. For the species *Dexiotricha sp*., *Metanophrys sp*., *Trithigmostoma cucullulus* and *Vorticella sp*., DADA2 produced one ASV per species. For all other species, sequence grouping with DADA2 resulted in two or more ASVs, eight of which yielded five or more ASVs. The most ASVs were obtained for *Stentor coeruleus* (11). There was a moderate linear relationship between the abundance of reads and the resulting number of ASVs in each species-specific dataset (R^2 = 0.51, p-value <0.001).

Table 1. First-level sequence grouping results of each ciliate species-specific dataset.

Ciliate species	DADA2-ASVs	SWARM-OTUs
Chilodonella uncinata	5	9
Coleps hirtus hirtus	3	5
Deviata rositae	2	5
Dexiotricha sp.	1	6
Epistylis plicatilis	5	9
Euplotes sp.	5	6
Folliculina sp.	7	18
Fuscheria uluruensis	2	8
Gastrostyla steinii	2	4
Hypotrichida sp.	3	5
Lagynophrya acuminata	2	11
Metanophrys sp.	1	4
Oxytricha granulifera	3	7
Paramecium bursaria	7	10
Paramecium tetraurelia	3	9
Pelagodileptus trachelioides	3	7
Platynematum salinarum	3	15
Schmidingerothrix salinarum	7	11
Spathidium ascendens	4	19
Spathidium foissneri	3	4
Spirostomum ambiguum	2	3
Stentor coeruleus	11	12
Tetrahymena sp.	2	2
Tokophrya infusionum	2	2
Trithigmostoma cucullulus	1	1
Uroleptus willii	5	16
Urospinula succisa	4	8
Usconophrys sp.	2	7
Vorticella sp.	1	6

Results are shown for first-level grouping with both DADA2 and SWARM.

Sequence grouping of the 29 species-specific ciliate datasets with the SWARM pipeline resulted in 229 different OTUs (Table 1), overestimating the species diversity of the samples by a factor of 7.9. *Trithigmostoma cucullulus* was the only species for which SWARM produced one OTU per species. Two OTUs were obtained for the species *Tetrahymena sp*. and *Tokophrya infusionum*. Twenty-two of the species produced five or more OTUs, with most obtained for *Folliculina sp*. (18) and *Spathidium ascendens* (19). There was a weak linear relationship between the abundance of reads and the resulting number of OTUs in each species-specific dataset (R^2 = 0.11, p-value = 0.045).

In direct comparison, DADA2 produced 2.3-times fewer ASVs than SWARM produced OTUs (Table 1). We observed a weak linear relationship between the overestimation of DADA2 and the overestimation of SWARM across all species (R^2 = 0.31, p-value = 0.001). However, *Folliculina sp*. and *S. ascendens*, the two species with the largest overestimation in SWARM, resulted in only seven and four ASVs in DADA2 respectively. By contrast, *Stentor coeruleus*, the species with the largest overestimation in DADA2, also resulted in 12 SWARM-OTUs. For three species the number of ASVs and OTUs was identical (*Tetrahymena sp*., *T. cucullulus*, *T. infusionum*).

For the remaining 26 species fewer ASVs were produced than OTUs. There was not a single species for which more OTUs than ASVs were produced.

Quantitative evaluation of second-level sequence grouping

Sequence clusters originating from second-level sequence grouping are generally designated as NSCs (Fig. 1). When resulting from DADA2-ASVs, these NSCs are further described as ASV-NSCs, while NSCs resulting from SWARM-OTUs are further described as OTU-NSCs. The sequence similarity network approach successfully reduced the amount of sequence clusters from first-level sequence grouping in both DADA2 and SWARM. The reduction was higher at lower sequence similarity binning levels, since more pairs of sequences could be connected in the network, thus producing larger NSCs (i.e. comprising more first-level ASVs or OTUs respectively; see Fig. 2 and Supporting Information Table S1). At a binning level of 99% similarity, SSNs could reduce the amount from 101 ASVs to 65 ASV-NSCs (35.6% reduction). Although no reduction was possible for SWARM-derived OTUs at the 99% binning level, the 229 SWARM-derived OTUs were reduced by more than half (53.3% reduction) to 107 OTU-NSCs at the next-lower binning level of 98% similarity. DADA2-derived ASVs were also reduced by more than half (51.5% reduction) at 98% similarity. At the lowest tested binning level of 90%, SSNs could reduce the amount from 101 DADA2-derived ASVs to 16 ASV-NSCs (84.2% reduction) and from 229 SWARM-derived OTUs to 15 OTU-NSCs (93.4% reduction).

Overall, the results of the DADA2-SSN combination were notably closer to the ideal of one NSC per species than the results of the SWARM-SSN combination (Fig. 2, Supporting Information Table S1). But for binning levels lower than 95% similarity, the combination of SSNs with either DADA2 or SWARM produced similar amounts of NSCs. Except for the SWARM-SSN results at 95% similarity, which matched the ideally expected diversity of 29 species, all of these analyses underestimated the expected amount of diversity. For most lower binning levels, SWARM-SSNs resulted in fewer NSCs than DADA2-SSNs, thus diverging stronger from the expected amount of species than DADA2-SSNs (exceptions were 91% and 95% similarity). At all binning levels higher than 95% similarity, the SWARM-SSN combination produced distinctively more NSCs than the DADA2-SSN combination. At 96% similarity, the DADA2-SSN results (29 ASV-NSCs) matched the ideally expected diversity of 29 species. Up to similarity binning levels of 98% similarity, the amounts of ASV-NSCs moderately increased (up to 49 NSCs, meaning a 1.7-fold overestimation of the expected diversity). 99% of similarity is the first level at which more than twice as much ASV-NSCs were found as species were analysed (65 ASV-NSCs, 2.2-fold overestimation). The results at the 100% binning level were equivalent to the first-level sequence grouping results of DADA2 without SSNs. By contrast, the amount of OTU-NSCs in the SWARM-SSN approach increased much stronger at higher binning levels. When the binning level

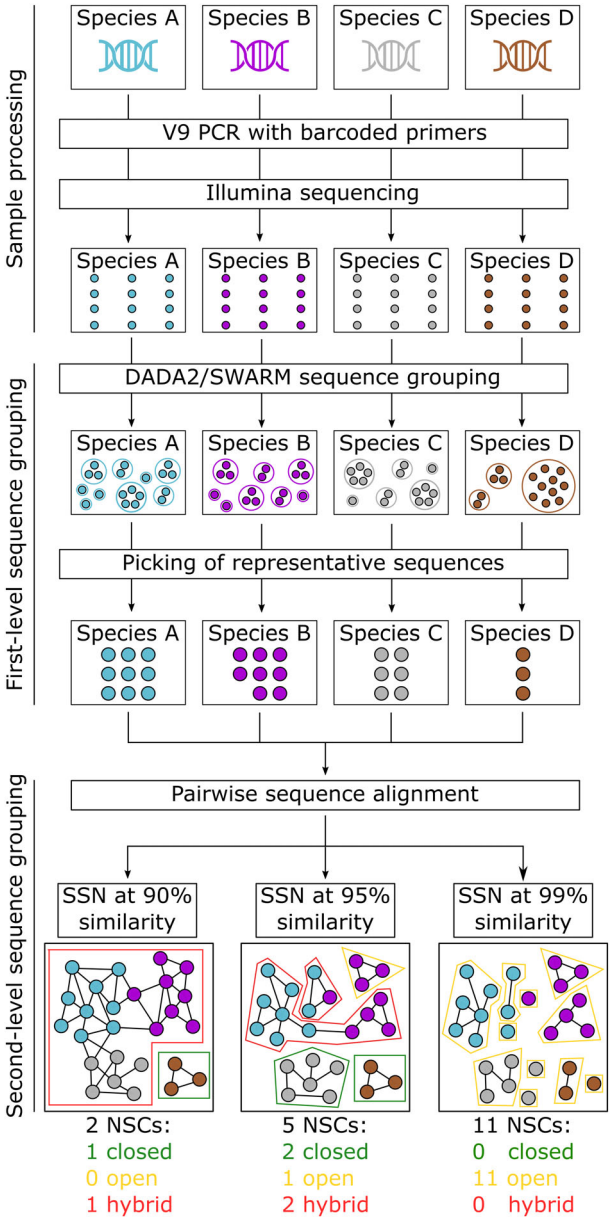

Fig. 1. Schematic workflow of the two-level sequence grouping approach. The workflow is shown for four hypothetical species and includes sample processing, first-level sequence grouping in either DADA2 or SWARM and second-level sequence grouping with SSNs. A unique colour was chosen for each of the species to highlight the separate handling of each dataset. Second-level sequence grouping also displays in how far network sequence clusters (NSCs) can be used to distinguish between closed, open and hybrid NSCs.

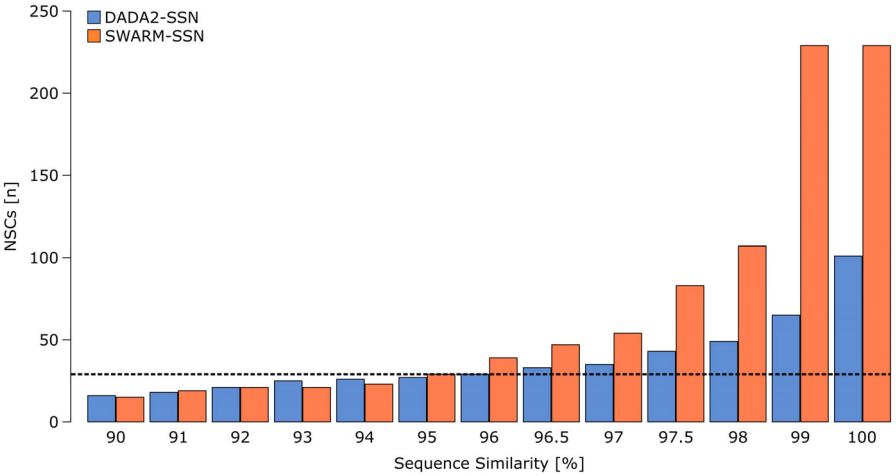

Fig. 2. Results of second-level sequence grouping at different sequence similarity binning levels. The bars show the gradual increase of network sequence clusters (NSCs) with increasing similarity binning level. Please note that the *x*-axis contains 0.5% steps between 96% and 98% similarity. DADA2-SSN results are displayed in blue, SWARM-SSN results in orange. The dashed line indicates the ideally expected diversity richness of 29 NSCs. Raw values of the bars can also be found in Supporting Information Table S1.

was set to 97% similarity, which is a commonly used threshold for sequence grouping of ciliate data, the SWARM-SSN approach produced 54 OTU-NSCs, which nearly doubles the expected diversity of the dataset (1.9-fold overestimation). At 98% similarity, 107 OTU-NSCs were produced, which exceeded the number of ASVs produced by DADA2 without a second-level grouping in SSNs (Fig. 2).

The application of an abundance filter could further decrease the overestimation of diversity by the second-level sequence grouping (Supporting Information Fig. S1, Supporting Information Table S2). For each species-specific dataset, we discarded every OTU that accounted for less than 0.01% of the total read abundance in that dataset. Thereby, there was only little effect on the SSN results at binning levels lower than 95% similarity. At higher binning levels, the effect was more apparent on the outcome of the SWARM-SSN approach. The diversity overestimation at the 100% similarity binning level decreased from a 7.9-fold overestimation without abundance filter, to a 4.5-fold overestimation when the abundance filter was applied. For the DADA2-SSN approach, the diversity overestimation decreased at the same binning level from 3.5-fold without abundance filter to 3-fold with abundance filter.

Qualitative evaluation of second-level sequence grouping

Species-specific datasets were used for first-level sequence grouping to preclude DADA2-derived ASVs or SWARM-derived OTUs that contained sequences from more than one dataset (Fig. 1). The preclusion did not apply for SSN analyses since representative sequences of either ASVs or OTUs from all species were pooled for all-versus-all pairwise sequence analyses during second-level sequence grouping. Therefore, we distinguished between three different types of NSCs (Fig. 1): (i) closed NSCs, which contained all DADA2-derived ASVs or SWARM-derived OTUs of one species and none from any other species; (ii) open NSCs, which contained some, but not all, DADA2-derived ASVs or SWARM-derived OTUs of one species and none from any other species; (iii) hybrid NSCs, which contained DADA2-derived ASVs or SWARM-derived OTUs of at least two different species. The goal is to determine which approach maximizes the number of closed NSCs and minimizes the number of hybrid NSCs for each species-specific dataset. The distinction between hybrid, open and closed NSCs should not be confounded with the terminology used for defining reference databases in some OTU clustering methods (for more information on the latter see e.g. Bik *et al*., 2012). Figures 3A and 3B illustrate the results at each binning level from 90% to 100% sequence similarity (see also Supporting Information Table S3). Independent of the first-level sequence grouping algorithm, most closed NSCs were produced at intermediate binning levels. Except for 91% similarity, the DADA2-SSN approach produced at each binning level more closed NSCs than the SWARM-SSN approach. The maxima of 24 closed ASV-NSCs and 20 closed OTU-NSCs were detected at a binning level of 94% similarity. The least number of closed NSCs was in both cases detected at the 100% similarity level (four closed ASV-NSCs, one closed OTU-NSC). Most hybrid NSCs were produced at low similarity binning levels. At 90% similarity we detected maxima of four hybrid ASV-NSCs and OTU-NSCs. In general, both approaches produced similar numbers of hybrid NSCs at each binning level, with decreasing numbers of hybrid NSCs at increasing similarity binning levels. For the DADA2-SSN approach one hybrid NSC was consistently observed even at binning levels up to 99% similarity; for the SWARM-SSN approach no more hybrid NSCs were observed at binning levels higher than 97% similarity. By contrast, increasing similarity binning levels led to

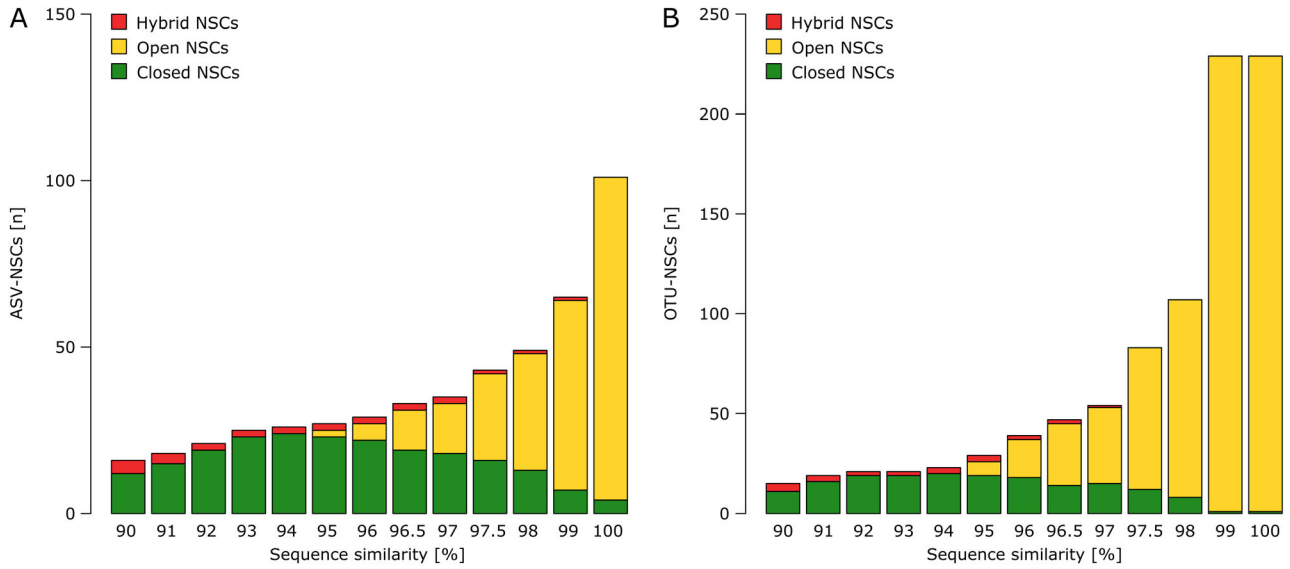

Fig. 3. Qualitative evaluation of second-level sequence grouping with the DADA2-SSN approach (A) and the SWARM-SSN approach (B). The evaluation focused on the questions how many closed network sequence clusters (NSCs; green), open NSCs (yellow) and hybrid NSCs (red) were found. Ideally, a binning level should be chosen at which both the number of closed NSCs is maximal and the number of hybrid NSCs is minimal. The bars show the gradual changes of NSC types with increasing similarity binning levels.

increasing numbers of open NSC for both approaches. Open NSCs were observed in neither approach up to a binning level of 94%. Starting from two open ASV-NSCs and seven open OTU-NSCs at 95% similarity, the numbers of open NSCs steadily increased until reaching 97 open ASV-NSCs and 228 open OTU-NSCs at 100% similarity. The maximal number of open ASV-NSCs was surpassed in the SWARM-SSN approach already at a binning level of 98% similarity (amounting to 99 OTU-NSCs).

A detailed evaluation of the DADA2-SSN results (Fig. 3A) revealed that at binning levels from 94% to 99% sequence similarity, one hybrid ASV-NSC was repeatedly produced. This always comprised *Hypotrichida sp*. and *Oxytricha granulifera* (as well as *Urospinula succisa* at binning levels lower than 97%). Another hybrid ASV-NSC, comprising *Spathidium ascendens* and *Spathidium foissneri*, was continuously detected until a binning level of 97%, after which the two species from the same genus no longer formed a hybrid ASV-NSC. The hybrid NSC patterns of the SWARM-SSN approach were similar to the ones observed in the DADA2-SSN approach. However, hybrid OTU-NSCs comprising *Hypotrichida sp., O. granulifera* and *U. succisa* were only detected up to 95% similarity. Of these three, only *Hypotrichida sp*. and *U. succisa* also aggregated into hybrid OTU-NSCs at 96% and 96.5% similarity. Hybrid OTU-NSCs comprising *S. ascendens* and *S. foissneri* were detected from 93% to 97% similarity. Starting from the 96% similarity binning level, the amount of open OTU-NSCs was always higher than the amount of closed OTU-NSCs, with increasing numbers of open OTU-NSCs and decreasing numbers of closed OTU-NSCs towards higher similarity binning levels.

The cluster distribution of NSCs was compared against an artificial perfect NSC cluster distribution, in which only closed NSCs existed. The results of this comparison are expressed as Rand index (RI) and adjusted Rand index (ARI) values in Fig. 4. For both indices, the DADA2-SSN approach resulted in higher values than the SWARM-SSN approach. A notable exception are the values for the binning level of 96% similarity at which the maxima for the SWARM-SSN approach were observed. The maximum RI and ARI values for the DADA2-SSN approach were observed at the 94% sequence similarity binning level. The progression of RI values with a distinct plateau phase at intermediate binning levels is similar for DADA2-SSNs and SWARM-SSNs. For DADA2-SSNs the phase starts at 93% similarity (RI = 0.9871) and reaches until 97% similarity (RI = 0.9907). For SWARM-SSNs, the plateau phase is shorter and reaches from 95% similarity (RI = 0.9873) to 97% similarity (RI = 0.985). At binning levels higher than 97%, RI values slowly start to decrease, while ARI values drastically decrease. For both approaches, RI values were lowest at the 90% similarity and ARI values were lowest at the 100% similarity binning level.

Applying an abundance-filter affected the quantitative output of the second-level sequence grouping stronger than the qualitative output. The removal of low abundant DADA2-derived ASVs or SWARM-derived OTUs resulted in first place in a general decrease of open NSCs, which was coupled to a slight increase of closed NSCs at similarity binning levels of 97% and higher (Supporting Information Figs S2A and S2B). But this did not lead to different trends comparing RI values calculated from abundance-filtered and non-abundance-filtered data

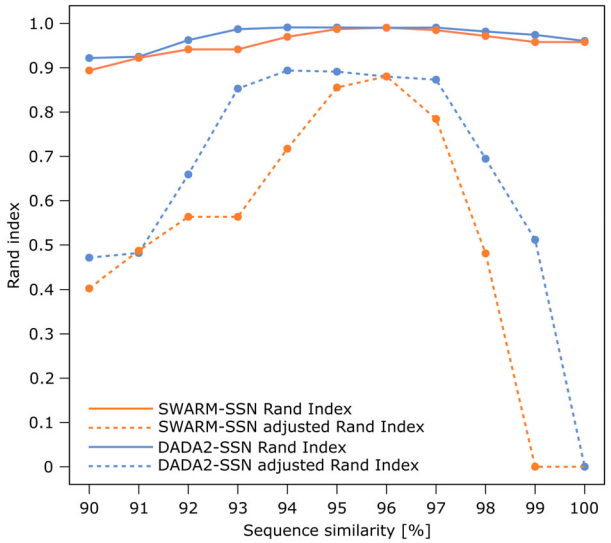

Fig. 4. Rand index (RI) and adjusted Rand index (ARI) results for second-level sequence grouping. The graphs show the values for comparing the observed cluster distributions of the DADA2-SSN (in blue) and SWARM-SSN (in orange) approaches against a perfect cluster distribution of the same data in which one cluster existed for each of the 29 species under study. The values were calculated for all tested sequence similarity binning levels. RI results are displayed as a straight line, ARI results as a dashed line.

(Supporting Information Fig. S3). There was, however, a trend towards higher ARI values for the DADA-SSN approach and lower ARI values for the SWARM-SSN approach compared between abundance-filtered and non-abundance-filtered data.

We also tested the complete sequence grouping workflow for DADA2 and SWARM when merging all species-specific datasets into one ciliate dataset before first-level sequence grouping. However, the results were not substantially different from those obtained when species-specific datasets were used for first-level sequence grouping (Supporting Information Table S5, Supporting Information Figs S4 and S5). A stronger effect could be observed when all species-specific datasets were merged before first-level sequence grouping and, in addition, an abundance-filter was applied before second-level sequence grouping in SSNs (Supporting Information Table S5). Since the initial idea of this study was to analyse distinct ciliate species in the same way and each species-specific dataset indeed represented a sample of a distinct ciliate species, we decided to focus on the results obtained from species-specific datasets.

Discussion

First-level sequence grouping results overestimate species richness

A major goal of most environmental HTS studies is to provide realistic estimates of biodiversity within a sample. However, it is not a trivial task to place the tremendous amount of resulting data into an ecologically meaningful context. One of the main difficulties is the delineation of species based on HTS datasets. Ciliates have been used as model organisms for addressing this problem in the past, but so far all comparisons of morphospecies richness and OTU richness within the same samples have been incongruent (Bachy *et al*., 2013; Grattepanche *et al*., 2014; Stoeck *et al*., 2014). Although some studies relying on clone library and Sanger sequencing technologies reported an underestimation of diversity (Bachy *et al*., 2013; Grattepanche *et al*., 2014), more recent studies relying on short HTS reads reported a diversity overestimation (Stoeck *et al*., 2014; Flynn *et al*., 2015; Clare *et al*., 2016). The same trend can be inferred from the results of our study: the DADA2 as well as the SWARM algorithm delineated highly similar sets of sequences obtained from species-specific Illumina-sequencing datasets and assigned to the same taxonomic hit, into multiple ASVs or OTUs. However, of the two tested first-level sequence grouping approaches, DADA2-derived ASVs came much closer to the known number of species than SWARM-derived OTUs.

The lower level of diversity overestimation in DADA2 is a consequence of the algorithm's error-model based approach (Callahan *et al*., 2017; Knight *et al*., 2018). Not only does DADA2 exclude a large fraction of reads based on statistical models from the analyses before assigning ASVs, it also performs a sequencing artefact correction. Together, these steps aim at removing all spurious reads and retaining only those variants that are not a product of erroneous sequencing. Conclusions about whether or not an ASV represents a true organismal variant are still hard to draw, since ciliates and other microbial species are known for intraspecific and intraindividual sequence polymorphism (e.g. Miao *et al*., 2004; Coleman, 2005; Gong *et al*., 2013; Wang *et al*., 2017). There may be cases in which DADA2 retains artificial sequences, just as well as there are some cases in which DADA2 discards true organismal variants. For instance, previous work of Gong and colleagues (2007) reported an intraspecific polymorphism with a mean sequence divergence of 1.6% for the SSU rRNA gene within the ciliate genus *Gastrostyla*. Given that their results were based on sequences of three individuals, this may only represent a fraction of the complete intraspecific polymorphism for these organisms. In our analyses, we detected only two ASVs for *Gastrostyla steinii* while we detected considerably more ASVs of ciliates (e.g. 11 ASVs for *Stentor coeruleus*) for which no high variation rates of intraspecific polymorphism have been reported (Kusch, 1998; Zhang *et al*., 2012). Thus, it is possible that some of the more divergent true organismal variants of *G. steinii* may have been error-corrected by DADA2.

In contrast to the DADA2-pipeline, the SWARM algorithm does not perform any denoising but is purely designed for grouping sequences. As such, SWARM will use any sequence provided in the input dataset without deciding whether or not the sequence may be a true organismal variant or a sequencing artefact. Denoising is a very sensible and important step that has to be performed by other bioinformatic tools, preferentially on results obtained after sequence grouping in SWARM (Mahé *et al*., 2014). SWARM's focus is on finding smallest differences between sequences and enabling a very fine-scaled resolution of genetic diversity in a sample. As reflected by our results, this is counterproductive when looking at alpha diversity or species richness within a sample. The *Spathidium ascendens* dataset of 19 SWARM-derived OTUs was the most severe example of over-splitting in our data. Since the dataset of *S. ascendens* was subjected to the same bioinformatic treatment as the other datasets with standard SWARM parameters, it is unlikely that this high SWARM-OTU number is an artefact of the SWARM algorithm. The inflation of OTUs may be caused by intraspecific sequence polymorphism in *S. ascendens*, but this feature has not been studied for this species before. For other species which yielded high numbers of OTUs, intraspecific and intraindividual sequence polymorphism is documented. Among the order Heterotrichida, polymorphic sites were found within the V9 region of the 18S rDNA (Wang *et al*., 2017), which could explain our finding of 18 SWARM-OTUs for *Folliculina sp*. Sequence polymorphism is also a widespread feature in the genus *Paramecium* (Coleman, 2005). At least some part of the SWARM-OTUs from *Paramecium bursaria* (10 OTUs) and *Paramecium tetraurelia* (9 OTUs) may reflect this true intraspecific genetic diversity. Using dataset replicates, as proposed by Prosser (2010), might help for deciding which sequence is artificial and which is a true organismal variant. But even this strategy cannot be the *ultima ratio* for species-specific datasets from single cell sequencing, since intraindividual genetic diversity may not contain all variants of intraspecific genetic diversity.

Second-level sequence grouping with sequence similarity networks improves diversity estimates

The shortcomings for species diversity estimations of both first-level sequence groupings (DADA2-ASVs and SWARM-OTUs) were distinctively alleviated when sequence similarity networks were used as second-level sequence grouping. Even though species richness was still overestimated, the extent of overestimation decreased through the application of SSNs. The greater effect of SSNs, in terms of reducing the amount of first-level sequence grouping results, could be observed for SWARM. These findings corroborate predictions of an earlier study, in which SWARM and SSNs had not been used in combination, but as two different means of sequence grouping (Forster *et al*., 2016). Interestingly, though, no reduction of SWARM's first-level sequence grouping results could be achieved at sequence similarity binning levels of 99% and higher. This result can be attributed to our application of SWARM on short V9 gene regions of ciliates, which reach an average length of 120 nucleotides (Dunthorn *et al*., 2012). Because a single nucleotide difference was used on these sequences in SWARM's first-level sequence grouping step, SWARM-OTUs already contained all pairs of sequences with sequence similarities higher than 99% to each other. Thus, a further reduction by SSNs in second-level sequence grouping was not possible at this binning level. To rely on local nucleotide differences and avoid global sequence similarity values (typically 97%) is a central aspect of SWARM (Mahé *et al*., 2014, 2015). Although this increases the resolution of genetic diversity, the higher resolution comes at the cost of generating numerous OTUs characterized by similar sequences. Sequence similarity networks are well suited for reducing the apparent number of OTUs, since they offer a straightforward way of grouping SWARM-derived OTUs while treating each of them as an equal entity and allowing further downstream comparisons of these entities (e.g. as in Forster *et al*., 2015, 2016). The same also applies to DADA2-derived ASVs. As outlined in the discussion of the DADA2 approach, first-level sequence grouping in DADA2 resulted in a much smaller overestimation of species richness than SWARM. Although there was inherently less room for further reduction of the diversity overestimation by SSNs, the DADA2-SSN combination did produce diversity estimates that better mirrored the species richness than those obtained from the SWARM-SSN combination.

Previous studies (e.g. Huse *et al*., 2010; Bonder *et al*., 2012) predicted that two levels of sequence grouping will have positive effects on the accuracy with which taxonomic units are defined from HTS datasets, but without using sequence similarity networks. Huse and colleagues conducted an initial single-linkage sequence grouping followed by a second-level sequence grouping, similar to the combination of SWARM and SSNs. Although our results confirm their predictions about the positive effect of the two-level sequence grouping strategy, the combination of SWARM (for single-linkage first-level grouping) with SSNs (for second-level grouping) was not as accurate for defining species-specific groups as the DADA2-SSN combination. The latter combination is somewhat similar to the strategy used by Bonder and colleagues (2012): for a mock dataset of 15 species, they reduced the number of OTUs by 93.4% when a denoising

step was employed before final sequence grouping. Likewise, our combination of DADA2 as a denoising step and SSNs for sequence grouping led to a reduction of OTUs from first- to second-level grouping by 74.3% at the binning level with the highest RI and ARI score (94% similarity). Further reduction of NSCs would be possible, but would not lead to a qualitatively better output. The lower NSC reduction rate in the current study compared with the study of Bonder and colleagues (2012) is merely a product of the more effective denoising by DADA2 (Callahan *et al*., 2016; Knight *et al*., 2018).

DADA2 has recently been introduced as an effective first-level sequence grouping and denoising tool and combined with different sequence grouping algorithms (Frøslev *et al*., 2017; Anslan *et al*., 2018; Jusino *et al*., 2018; Palmer *et al*., 2018), but none of the studies used SSNs. Frøslev and colleagues (2017) decided to combine DADA2 with hierarchical sequence grouping in VSEARCH, followed by a post-clustering treatment based on ecological patterns to remove erroneous groups of sequences. In contrast to this, we decided to combine DADA2 with hierarchical sequence grouping in VSEARCH and analysed the resulting pairwise sequence similarities in SSNs. Without post-clustering treatment, both approaches result in quite similar patterns of diversity overestimation, although to a lesser extent by the DADA2-SSN approach on the ciliate dataset. When post-clustering is taken into account, the approach of Frøslev and colleagues led to an underestimation of diversity by one-third at their suggested level of 97% sequence similarity. At the same level of similarity, the DADA2-SSN combination overestimated the diversity by one-fifth (six ASV-NSCs more than species expected). In addition, our approach led to more accurate diversity estimates and an underestimation of diversity by only three ASV-NSCs at the binning level with the highest RI and ARI scores (94% similarity). The marginal differences between RI and ARI scores at intermediate similarity binning levels (e.g. RI difference of 0.0002 and ARI difference of 0.0026 between 94% and 95% similarity), indicate that there is not one universally applicable level for optimal sequence grouping. Instead, there exists a range of sequence similarities, at which qualitatively highly similar and equally precise sequence grouping can be achieved. To identify the best binning level, we advise to create SSNs on such a range of similarities and evaluate the outcome for best sequence grouping results. Our data suggests that the use of more conservative and lower similarity binning levels has a positive effect towards more accurate diversity estimates of ciliates while also allowing for more precise species delineation. Similar conclusions can be drawn from the SWARM-SSN approach, which yielded the best results at 96% similarity and thus, below the 97% similarity threshold widely used for clustering ciliate sequence data. Although our two-level sequence grouping strategy can easily be adapted to datasets of other taxonomic groups or other barcode gene regions, it is important to note that the similarity binning level, which produced the best output for our ciliate V9 18S rDNA dataset, is not generalizable to other datasets without prior tests. The extent of genetic diversity varies among different taxonomic groups (e.g. Brown *et al*., 2015) and even within ciliates, different sequence similarity thresholds have been shown to be more effective for delineating species when working with datasets of different hypervariable regions of the 18S rDNA (Dunthorn *et al*., 2012).

All-versus-all pairwise sequence alignments in hierarchical sequence grouping are a prerequisite for attaining the advantages of sequence similarity networks (Forster *et al*., 2015, 2016; Corel *et al*., 2016). One of the main benefits for using hierarchical instead of heuristic sequence grouping is that an over-splitting of diversity is avoided (Mahé *et al*., 2014; Flynn *et al*., 2015). This benefit displays in the avoidance of diversity over-splitting in our study as well. By contrast, an over-splitting was observed in the study of Frøslev and colleagues (2017) when no additional post-clustering of the sequence grouping results was conducted. A further benefit of using networks for second-level grouping is that they allow for detailed evaluation of the information provided within each NSC. For instance, the persistent grouping of *Spathidium ascendens* and *Spathidium foissneri* in hybrid NSCs is not only explained by them belonging to the same genus, but can be further related to rapid radiation events and incomplete lineage sorting in the order Spathidiida (Vďačný *et al*., 2014). By considering the internal structure of NSCs, one can thus draw additional ecological and evolutionary conclusions about the organisms under study.

Although SSNs emerged as a powerful tool for improving first-level sequence grouping results, we have to state that it remained impossible to perfectly reproduce the diversity in the samples. This conclusion is not unexpected, because evolutionary processes can only be approximated, but not predicted by bioinformatic algorithms and molecular proxies. Different lineages may evolve at different rates (Brown *et al*., 2015) or gene transfer may occur (Bapteste and Boucher, 2008), all of which complicate the evaluation of sequencing data and the estimation of diversity from an environmental community dataset. There is, however, still room for improvement and drawing even more accurate pictures when working with SSNs as second-level sequence grouping. Our results indicated that SWARM-derived OTUs are strongly affected by subsequent denoising steps. For SWARM sequence grouping and denoising, we followed the same strategy used in several benchmarks studies (e.g. de Vargas *et al*., 2015; Mahé *et al*., 2017), but still observed a distinct overestimation of species richness. To limit the overestimation, we propose additional filtering steps relying on sequence abundances. Other studies have shown

that abundance filter can efficiently remove noise from HTS datasets (Quince *et al*., 2009, 2011; Reeder and Knight, 2010; Bokulich *et al*., 2013; Auer *et al*., 2017). Likewise, an abundance-filter step is also implemented in DADA2 (Callahan *et al*., 2016). Applying an abundance filter had a positive effect on the outcome of the DADA2-SSN and especially on the outcome of the SWARM-SSN strategy (see Figs S1–S3 and Supporting Information Table S2). The effect was further enhanced when species-specific datasets were merged before first-level sequence grouping and an abundance-filter was applied before second-level sequence grouping in SSNs. But even without merging the datasets, the ASV-NSCs and OTU-NSCs reflected the real species diversity quantitatively closer after the application of an abundance filter. The filter had little effect, however, on the formation of hybrid and closed NSCs. Most of all, the formation of open NSCs was distinctively reduced. Since open NSCs result from ASV or OTU variants, which cannot be affiliated with ASV or OTU variants from the same species, this implies that many open NSCs were actually low abundant variants of a given species. They possibly emerge from either sequencing errors or real but rare genetic polymorphisms. In how far the application of the abundance filter and the removal of these variants can be justified with regard to an environmental study and become a standard step of the second-level sequence grouping by SSNs remains to be tested in the future.

Conclusion

The consequences and relevance of notably different outputs using DADA2 and SWARM in ecological studies remains to be tested in real case scenarios using field samples. Although some specific ecological questions (such as results of beta diversity analyses) may not suffer, others (such as results of alpha diversity analyses or the identification of ecologically relevant key species) may be highly compromised by different information obtained from DADA2 and SWARM. Beyond that, second-level sequence grouping with sequence similarity networks clearly improved diversity estimates compared with first-level sequence grouping. We expect that future studies will benefit from implementing this strategy, especially by relying on the combination of DADA2 and sequence similarity networks.

Experimental procedures

Ciliate specimen collection

Single ciliate cells were hand-picked from either pure cell cultures or from environmental culture samples, leading to a collection of 29 different ciliate species (Table 1, see Supporting Information Table S4 for material sources). To allow for testing species delineation at higher taxonomic levels, six of the known ciliate classes were covered, including Heterotrichea, Litostomatea, Oligohymenophorea, Phyllopharyngea, Prostomatea and Spirotrichea. To test the species delineation on lower taxonomic levels, the collection also included two species from each of the genera *Paramecium* and *Spathidium*. Each of the 29 species was treated independently in our workflow to create one species-specific high-throughput sequencing (HTS) dataset per ciliate for downstream analyses (Fig. 1).

DNA extraction, amplification and HTS

For each ciliate species, DNA was extracted from individually picked specimens from pure cultures, enrichments or environmental samples using Qiagen's DNeasy Blood and Tissue Kit, followed by a clean-up step with Qiagen's MinElute Polymerase Chain Reaction (PCR) Purification Kit. Both kits were used according to the manufacturer's instructions (Qiagen; Hilden, Germany). Extracted DNA was then amplified in a semi-nested PCR targeting the hypervariable V9 region of the 18S rDNA. The V9 is a routinely used barcode gene region for eukaryotic community analyses (de Vargas *et al*., 2015) and has been specifically tested for delineation of ciliate species (Dunthorn *et al*., 2012, 2014). In the first step of the semi-nested PCR, we used the ciliate specific forward primer CilF (5′-TGGTAGTGTATTGGACWACCA-3′; Lara *et al*., 2007) in combination with the universal eukaryotic 18S reverse primer EukB (5′-TGATCCTTCTGCAGGTTCACCTAC-3′; Medlin *et al*., 1988). This initial step was especially important for ciliate species from environmental samples to avoid amplification of non-target eukaryotic organisms. The CilF-EukB protocol consisted of an initial denaturation step of 30 s (s) at 98°C, followed by 35 cycles of 10 s at 98°C, 30 s at 62°C and 60 s at 72°C; the final extension lasted for 300 s at 72°C. After the first step of the semi-nested protocol, resulting PCR products were cleaned once more with Qiagen's MinElute PCR Purification Kit according to the manufacturer's instructions. In the second step of the semi-nested PCR, the CilF-EukB products were amplified with a universal eukaryotic V9 forward primer (5′-GTACACACCGCCCGTC-3′; Lane, 1991) and EukB as reverse primer. Each primer pair was tagged with one of 10 different barcodes, so that the resulting species-specific barcoded sequences could easily be retraced in downstream analyses. The V9-EukB protocol consisted of an initial denaturation step of 30 s at 98°C, followed by 25–35 cycles (adjusted for each species, see Supporting Information Table S4) of 10 s at 98°C, 20 s at 64°C and 25 s at 72°C; the final extension lasted for 300 s at 72°C. Once the barcoded V9 PCR products were successfully

amplified, samples were pooled into libraries that contained up to 10 different barcodes (one for each ciliate species). Libraries were paired-end sequenced on either an Illumina MiSeq or NextSeq platform (see Supporting Information Table S4 for details), each generating 250-base pair reads. HTS was conducted by SeqIT (Kaiserslautern, Germany).

After sequencing, libraries were split into single species datasets using CUTADAPT v1.18 (Martin, 2011) according to the barcodes initially applied during the semi-nested PCR. Filtering of reads with an expected barcode and removal of those was followed by matching the primer sequences at both 5′ and 3′-ends in CUTADAPT, as well. Reads were kept if they exactly matched the forward or reverse V9 primers at their 5′-end and, at the same time, if they exactly matched the reverse or forward V9 primers near to the 3′-end ('linked adapter' approach in CUTADAPT). Primer sequences and overhang at 5′-end were removed during this filtering process to keep only the targeted V9 region. Reads were oriented in the same direction using VSEARCH v2.8.0 (Rognes *et al*., 2016). Thus, the first paired libraries only contained reads oriented in the forward V9 primer direction and the second paired libraries only contained reads oriented in the reverse V9 primer direction. The barcode- and primer-filtering produced 29 single species paired libraries (for the species *Epistylis plicatilis* two libraries existed, but the data were merged for subsequent steps) which were used as input for the downstream first-level sequence grouping approaches.

The sequence data were deposited at the Sequence Read Archive of the National Center for Biotechnology Information (NCBI) and are available under accession number PRJNA548847. All bioinformatic procedures and commands for statistical analyses are provided in HTML format as Supporting Information (File S1, Supporting Information Table S6).

Quality filtering and first-level sequence grouping with DADA2

Reads were truncated to their first 100 bp and filtered with a maximum expected error of 0.2 using the DADA2 package v1.8 in R v3.5.1 (Callahan *et al*., 2016; R Core Team, 2018). Then error rates were learned for each sequencing run separately, reads were dereplicated per single species paired libraries and the core DADA2 denoising algorithm was applied on each read of each single species' paired library using their respective sequencing run error model. Denoised reads were paired for each single species by requesting a minimum overlap of 50 bp and by allowing a maximum of five mismatches using the function *mergePairs* in DADA2. ASVs produced in this manner were merged among all 29 single species libraries and chimeras were *de novo* removed from this ASV table using the consensus approach from the *removeBimeraDenovo* function.

Quality filtering and first-level sequence grouping with SWARM

Demultiplexed libraries, from which barcodes and primers were previously removed, were used for first-level sequence grouping with SWARM. For each paired-end, species-specific library, we precisely followed the publicly available instructions at https://github.com/frederic-mahe/swarm/wiki/Fred's-metabarcoding-pipeline. In short, paired-end reads were merged and subsequently dereplicated into amplicons using VSEARCH. First-level sequence grouping in SWARM v2.0.5 (Mahé *et al*., 2015) was performed on the amplicons with *-d 1* and the fastidious option *-f*. The resulting SWARM-OTUs were subjected to a *de novo* chimera detection in UCHIME (Edgar *et al*., 2011) and singletons were removed from the output. From each non-chimeric SWARM-OTU we extracted the seed amplicon as representative sequence for downstream taxonomic assignment.

Taxonomic assignment of first-level sequence grouping results

ASVs and representative seed sequences (i.e. the most abundant amplicon) of SWARM OTUs were annotated using BLAST (Altschul *et al*., 1990) against a modified version of the PR2-derived database provided in de Vargas and colleagues (2015). This reference database was specifically designed for taxonomic assignment of V9 sequences via VSEARCH (i.e. by containing only references trimmed to the V9 region). We manually trimmed and added ciliate reference sequences that were still missing in the database. The added references represent NCBI GenBank entries deposited under accession numbers AB558117, AF429900, AF508776, FJ998037, KC991098, KF301567, KF411460, KF733753, KF733756, KF878932, KU525298, MG589318. Only ASVs and OTUs assigned with at least 90% sequence similarity to the most abundant ciliate species in each species-specific library were kept for further analyses, while sequences of prey, parasitic or mutualistic organisms present in the samples were discarded.

Second-level sequence grouping with sequence similarity networks

Identical second-level sequence grouping steps were performed independently for the DADA2 and SWARM first-level sequence grouping outputs (Fig. 1). Representative sequences of all species-specific target ASVs or

OTUs were first pooled to create one dataset for either DADA2 and SWARM. The initial step for sequence similarity network construction employed all-versus-all pairwise sequence analyses of these datasets in VSEARCH (Rognes *et al*., 2016) using the settings *-allpairs_global*, *-iddef 1* and a similarity cutoff of 90%. This resulted in an edge table in which each line represented a pair of sequences that shared a sequence similarity of at least 90% to each other. From the edge table, unweighted and undirected SSNs were calculated in R with the package *igraph* v1.2.2 (Csardi and Nepusz, 2006). To determine the sequence similarity which gave the maximal congruence between the number of NSCs and the number of ciliate species, similarity binning levels for SSN construction from 90% to 100% were tested in single percentage steps, except between 96% and 98% for which we tested 0.5 percentage steps. This was because previous studies suggested the best cutoff level for species delineation at this range of sequence similarities (e.g. Worden, 2006; Caron *et al*., 2009). Every node in a SSN represented one DADA2-derived ASV or one SWARM-derived OTU representative sequence. Every edge in a SSN represented a sequence similarity between two ASVs or two representative sequences that exceeded the applied binning level (Fig. 1, see also Forster *et al*., 2015). Thus, NSCs either represented connected components, that is, a cluster of ASVs or OTU representative sequences that could be further grouped based on the applied binning level; or a single node, that is, a single ASV or OTU representative sequence which sequence similarity to any other ASV or OTU representative sequence in the dataset was lower than the applied binning level. For instance, a binning level of 100% similarity reproduced the results from the first-level sequence grouping, meaning that the SSNs consisted exclusively of single nodes that represented DADA2-derived ASVs or SWARM-derived OTUs.

Previous studies indicated that the implementation of an abundance filter can be beneficial for increasing the accuracy of sequence grouping (e.g. Quince *et al*., 2009, 2011; Reeder and Knight, 2010; Bokulich *et al*., 2013; Auer *et al*., 2017). The rationale behind this is that sequences which emerge from organisms that are actually occurring in a sample should be much more abundant in a HTS dataset, than sequences which represent sequencing artefacts or elusive contaminations. Following this idea, we tested in a separate approach if an abundance filter could improve the outcome of our two-level sequence grouping strategy. For this test, we removed before SSN construction from each species' dataset all DADA2-derived ASVs with an abundance of less than 0.01% with regard to all sequences of that dataset. For abundance filtering in SWARM we first set SWARM's *-b* option to the species-specific 0.01% abundance threshold of each dataset during first-level sequence grouping, then removed before SSN construction all SWARM-OTUs with an abundance of less than 0.01% with regard to all sequences of that dataset. All other second-level sequence grouping steps in this test were performed as outlined above.

Statistical evaluation of sequence grouping

All statistical tests were run in R v3.5.1. The focus of the statistical evaluation for second-level sequence grouping was to identify the similarity binning level, which maximized the number of closed NSCs and minimized the number of hybrid NSCs. To express this level mathematically, we listed the NSC membership of each representative sequence and applied both the RI (Rand, 1971) and Hubert's and Arabie's ARI (Hubert and Arabie, 1985). In short, the RI compares the congruence between two cluster distributions, whereas the ARI is the corrected-by-chance version of the RI. The values of the indices range from 0 to 1, with 0 describing completely different cluster distributions and 1 describing perfectly matching cluster distributions between two sets of data. In this study, we compared the sequence grouping of representative sequences within NSCs for each binning level to the optimal distribution of representative sequences, in which case DADA2-derived ASVs and SWARM-derived OTUs exclusively form species-specific, closed NSCs (that is 29 NSCs, one for each species). RI and ARI were calculated with the R package clues v0.5.9 (Chang *et al*., 2010).

Acknowledgements

We thank Gianna Pitsch, Bettina Sonntag, Thomas Prös-chold, Ewa Przyboś and the staff of CCAP (SAMS, Oban, Scotland) for providing cultured ciliate species. Furthermore, we thank Hans-Werner Breiner and Sara Neß for their help with laboratory works. We are grateful for the constructive criticism of three anonymous reviewers. This study was supported by a research grant of the Carl Zeiss Stiftung to D.F. and an ASSEMBLE PLUS project to T.S. under European Union's Horizon 2020 research and innovation program (Grant Agreement No. 730984). The authors declare that there is no conflict of interest. Open access funding enabled and organized by Projekt DEAL.

[Correction added on 20 October 2020, after first online publication: Projekt Deal funding statement has been added.]

References

Allali, I., Arnold, J.W., Roach, J., Cadenas, M.B., Butz, N., Hassan, H.M., *et al*. (2017) A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiol* **17**: 194.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Xu, Z.Z., *et al*. (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2**: e00191–e00216.

Anslan, S., Nilsson, R.H., Wurzbacher, C., Baldrian, P., Tedersoo, L., and Bahram, M. (2018) Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. *MycoKeys* **39**: 29–40.

Auer, L., Mariadassou, M., O'Donohue, M., Klopp, C., and Hernandez-Raquet, G. (2017) Analysis of large 16S rRNA illumina data sets: impact of singleton read filtering on microbial community description. *Mol Ecol Resour* **17**: e122–e132.

Bachy, C., Dolan, J.R., López-García, P., Deschamps, P., and Moreira, D. (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *ISME J* **7**: 244–255.

Bapteste, E. and Boucher, Y. (2008) Lateral gene transfer challenges principles of microbial systematics. *Trends Microbiol* **16**: 200–207.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R., and Thomas, W.K. (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol Evol* **27**: 233–243.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., *et al*. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* **10**: 57–59.

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

Brown, E.A., Chain, F.J.J., Crease, T.J., MacIsaac, H.J., and Cristescu, M.E. (2015) Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecol Evol* **5**: 2234–2251.

Callahan, B.J., McMurdie, P.J., and Holmes, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* **11**: 2639–2643.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D., *et al*. (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75**: 5797–5808.

Chang, F., Qiu, W., Zamar, R.H., Lazarus, R., and Wang, X. (2010) Clues: an R package for nonparametric clustering based on local shrinking. *J Stat Softw* **33**: 1–16.

Clare, E.L., Chain, F.J.J., Littlefair, J.E., and Cristescu, M.E. (2016) The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome* **59**: 981–990.

Coleman, A.W. (2005) *Paramecium aurelia* revisited. *J Eukaryot Microbiol* **52**: 68–77.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016) Network-thinking: graphs to analyze microbial complexity and evolution. *Trends Microbiol* **24**: 224–237.

Csardi, G., and Nepusz, T. (2006) The igraph software package for complex network research. *Int J Complex Syst* **1695**: 1–9.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., *et al*. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Dunthorn, M., Klier, J., Bunge, J., and Stoeck, T. (2012) Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J Eukaryot Microbiol* **59**: 185–187.

Dunthorn, M., Otto, J., Berger, S.A., Stamatakis, A., Mahé, F., Romac, S., *et al*. (2014) Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Mol Biol Evol* **31**: 993–1009.

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

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.

Flynn, J.M., Brown, E.A., Chain, F.J.J., MacIsaac, H.J., and Cristescu, M.E. (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecol Evol* **5**: 2252–2266.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., *et al*. (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol* **13**: 16.

Forster, D., Dunthorn, M., Stoeck, T., and Mahé, F. (2016) Comparison of three clustering approaches for detecting novel environmental microbial diversity. *PeerJ* **4**: e1692.

Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., and Hansen, A.J. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat Commun* **8**: 1188.

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.

Gong, J., Dong, J., Liu, X., and Massana, R. (2013) Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* **164**: 369–379.

Gong, J., Kim, S.-J., Kim, S.-Y., Min, G.-S., Roberts, D.M., Warren, A., and Choi, J.-K. (2007) Taxonomic redescriptions of two ciliates, *Protogastrostyla pulchra n. g., n. comb*. and *Hemigastrostyla enigmatica* (Ciliophora: Spirotrichea, Stichotrichia), with phylogenetic analyses based on 18S and 28S rRNA gene sequences. *J Eukaryot Microbiol* **54**: 468–478.

Grattepanche, J.-D., Santoferrara, L.F., McManus, G.B., and Katz, L.A. (2014) Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends Microbiol* **22**: 432–437.

Hubert, L., and Arabie, P. (1985) Comparing partitions. *J Class* **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.

Jusino, M.A., Banik, M.T., Palmer, J.M., Wray, A.K., Xiao, L., Pelton, E., *et al*. (2018) An improved method for utilizing high-throughput amplicon sequencing to determine the diets of insectivorous animals. *Mol Ecol Resour* **19**: 176–190.

Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., *et al*. (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* **16**: 410–422.

Kumar, S., and Foissner, W. (2016) High cryptic soil ciliate (Ciliophora, Hypotrichida) diversity in Australia. *Eur J Protistol* **53**: 61–95.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

Kusch, J. (1998) Local and temporal distribution of different genotypes of pond-dwelling *Stentor coeruleus*. *Protist* **149**: 147–154.

Lane, D.J. (1991) 16S/23S rRNA sequencing. In *Nucleic Acid Techniques in Bacterial Systematics*, Stackebrandt, E., and Goodfellow, M. (eds). Chichester, UK: John Wiley and Sons.

Lara, E., Berney, C., Harms, H., and Chatzinotas, A. (2007) Cultivation-independent analysis reveals a shift in ciliate 18S rRNA gene diversity in a polycyclic aromatic hydrocarbon-polluted soil. *FEMS Microbiol Ecol* **62**: 365–373.

Lord, E., Cam, M.L., Bapteste, É., Méheust, R., Makarenkov, V., and Lapointe, F.-J. (2016) BRIDES: a new fast algorithm and software for characterizing evolving similarity networks using breakthroughs, roadblocks, impasses, detours, equals and shortcuts. *PLoS One* **11**: e0161474.

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., *et al.* (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat Ecol Evol* **1**: 0091.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**: e593.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**: e1420.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Medlin, L., Elwood, H.J., Stickel, S., and Sogin, M.L. (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* **71**: 491–499.

Miao, W., Fen, W.-S., Yu, Y.-H., Zhang, X.-Y., and Shen, Y.-F. (2004) Phylogenetic relationships of the subclass Peritrichia (Oligohymenophorea, Ciliophora) inferred from small subunit rRNA gene sequences. *J Eukaryot Microbiol* **51**: 180–186.

Nearing, J.T., Douglas, G.M., Comeau, A.M., and Langille, M.G.I. (2018) Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**: e5364.

Nebel, M., Pfabel, C., Stock, A., Dunthorn, M., and Stoeck, T. (2011) Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. *Environ Microbiol Rep* **3**: 154–158.

Needham, D.M., Sachdeva, R., and Fuhrman, J.A. (2017) Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME J* **11**: 1614–1629.

Palmer, J.M., Jusino, M.A., Banik, M.T., and Lindner, D.L. (2018) Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ* **6**: e4925.

Prosser, J.I. (2010) Replicate or lie. *Environ Microbiol* **12**: 1806–1810.

Quince, C., Lanzén, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., *et al*. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P. J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**: 38.

Rand, W.M. (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* **66**: 846–850.

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

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.

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., Ryabin, 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.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., and Richards, T.A. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19**: 21–31.

Stoeck, T., Breiner, H.-W., Filker, S., Ostermaier, V., Kammerlander, B., and Sonntag, B. (2014) A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology. *Environ Microbiol* **16**: 430–444.

Sun, Y., Cai, Y., Huse, S.M., Knight, R., Farmerie, W.G., Wang, X., and Mai, V. (2012) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinformatics* **13**: 107–121.

Tikhonov, M., Leach, R.W., and Wingreen, N.S. (2015) Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* **9**: 68–80.

Utter, D.R., Mark Welch, J.L., and Borisy, G.G. (2016) Individuality, stability, and variability of the plaque microbiome. *Front Microbiol* **7**: 564.

Vďačný, P., Breiner, H.-W., Yashchenko, V., Dunthorn, M., Stoeck, T., and Foissner, W. (2014) The chaos prevails: molecular phylogeny of the Haptoria (Ciliophora, Litostomatea). *Protist* **165**: 93–111.

Wang, C., Zhang, T., Wang, Y., Katz, L.A., Gao, F., and Song, W. (2017) Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. *Proc R Soc B* **284**: 20170425.

Worden, A.Z. (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat Microb Ecol* **43**: 165–175.

Xue, Z., Kable, M.E., and Marco, M.L. (2018) Impact of DNA sequencing and analysis methods on 16S rRNA gene bacterial community analysis of dairy products. *mSphere* **3**: e00410–e00418.

Zhang, W.-J., Lin, Y.-S., Cao, W.-Q., and Yang, J. (2012) Genetic diversity and variance of *Stentor coeruleus* (Ciliophora: Heterotrichea) inferred from inter-simple sequence repeat (ISSR) fingerprinting. *J Eukaryot Microbiol* **59**: 157–162.

Zoqratt, M.Z.H.M., Eng, W.W.H., Thai, B.T., Austin, C.M., and Gan, H.M. (2018) Microbiome analysis of Pacific white shrimp gut and rearing water from Malaysia and Vietnam: implications for aquaculture research and management. *PeerJ* **6**: e5826.

Supporting Information

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

File S1. Supplementary codes in HTML format.

Fig. S1. Results of second-level sequence grouping at different sequence similarity binning levels after abundance-filtering. The bars show the gradual increase of NSCs with increasing similarity binning level, similar to the results shown without abundance filter in Fig. 2. Please note that the x-axis contains 0.5% steps between 96% and 98% similarity. DADA2-SSN results are displayed in blue, SWARM-SSN results in orange. The dashed line indicates the ideally expected diversity richness of 29 NSCs.

Fig. S2. Qualitative evaluation of second-level sequence grouping with the DADA2-SSN approach (A) and the SWARM-SSN approach (B) after abundance-filtering. Similar to Figs 3A and B, the evaluation focused on the questions how many closed NSCs (green), open NSCs (yellow) and hybrid NSCs (red) were found when working with abundance-filtered data. Ideally, a binning level should be chosen at which both the number of closed NSCs is maximal and the number of hybrid NSCs is minimal. The bars show the gradual changes of NSC types with increasing similarity binning levels.

Fig. S3. Rand index (RI) and adjusted Rand index (ARI) results for second-level sequence grouping after abundance-filtering. The graphs show the values for comparing the observed cluster distributions of the DADA2-SSN (in blue) and SWARM-SSN (in orange) approaches after an initial abundance filtering step of the data, against a perfect cluster distribution of the same data in which one cluster existed for each of the 29 species under study. The values were calculated for all tested sequence similarity binning levels. RI results are displayed as a straight line, ARI results as a dashed line.

Fig. S4. Qualitative evaluation of second-level sequence grouping with the DADA2-SSN approach (A) and the SWARM-SSN approach (B) when species-specific datasets were merged before first-level sequence grouping. Similar to Figs 3A and 3B, as well as Figs S2A and S2B, the evaluation focused on the questions how many closed NSCs (green), open NSCs (yellow) and hybrid NSCs (red) were found. Ideally, a binning level should be chosen at which both the number of closed NSCs is maximal and the number of hybrid NSCs is minimal. The bars show the gradual changes of NSC types with increasing similarity binning levels.

Fig. S5. Rand index (RI) and adjusted Rand index (ARI) results for second-level sequence grouping, when species-specific datasets had been merged before first-level sequence grouping. The graphs show the values for comparing the observed cluster distributions of the DADA2-SSN (in blue) and SWARM-SSN (in orange) approaches, against a perfect cluster distribution of the same data in which one cluster existed for each of the 29 species under study. The values were calculated for all tested sequence similarity binning levels. RI results are displayed as a straight line, ARI results as a dashed line.

Table S1. NSC numbers at different sequence similarity binning levels. The values document the gradual increase of ASV-NSCs and OTU-NSCs with increasing similarity binning level, which is also displayed in Fig. 2.

Table S2. Difference between non-abundance filtered and abundance-filtered data before second-level sequence grouping. For testing the effect of an abundance filter, all DADA2-ASVs or SWARM-OTUs which amounted to less than 0.01% of the sequences of a species were filtered from the dataset. That is, the numbers shown here are also equivalent to the number of NSCs at the 100% binning level for second-level sequence grouping with and without abundance filter.

Table S3. Numbers of NSC types at each similarity binning level. The amounts of closed, open and hybrid NSCs shown in this table reflect the bars shown in Figs 3A and 3B.

Table S4. Sample information for each of the 29 ciliate species. In addition to species name and taxonomy, the table also shows the source of the sample, to which GenBank accession number the best hit refers and with which Illumina platform it was sequenced.

Table S5. NSC numbers at different sequence similarity binning levels when species-specific datasets were merged before first-level sequence grouping. The values document the gradual increase of ASV-NSCs and OTU-NSCs with increasing similarity binning level. NSC numbers at the similarity binning level of 100% are equivalent to the outcome of first-level sequence grouping after merging species-specific ciliate datasets. The third and fourth columns show NSC numbers when species-specific datasets were merged before first-level sequence grouping and an abundance-filter was applied before second-level sequence grouping.

Table S6. Library information on all used datasets.