A peer-reviewed version of this preprint was published in PeerJ on 25 February 2016.

Forster D, Dunthorn M, Stoeck T, Mahé F. 2016. Comparison of three clustering approaches for detecting novel environmental microbial diversity. PeerJ 4:e1692 https://doi.org/10.7717/peerj.1692

Comparison of three clustering approaches for detecting novel environmental microbial diversity

Dominik Forster, Micah Dunthorn, Thorsten Stoeck, Frédéric Mahé

Department of Ecology, University of Kaiserslautern, Kaiserslautern, Germany

Co-corresponding authors:

Dominik Forster, dforster@rhrk.uni-kl.de

Frédéric Mahé, mahe@rhrk.uni-kl.de

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1414v1 | CC-BY 4.0 Open Access | rec: 2 Oct 2015, publ: 2 Oct 2015

ABSTRACT

Discovery of novel diversity in high-throughput sequencing (HTS) studies is a central task in environmental microbial ecology. To evaluate the effects that amplicon clustering methods have on novel diversity discovery, we clustered an environmental marine protist HTS dataset of protist reads together with accessions from the taxonomically curated PR^2 reference database using three *de novo* approaches: sequence similarity networks, USEARCH, and Swarm. The novel diversity uncovered by each clustering approach differed drastically in the number of operational taxonomic units (OTUs) and the number of environmental amplicons in these novel diversity OTUs. Global pairwise alignment comparisons revealed that numerous amplicons classified as novel by USEARCH and Swarm were actually highly similar to reference accessions. Using graph theory we found additional novel diversity within OTUs that would have gone unnoticed without further using their underlying network topologies. Our results suggest that novel diversity inferred from clustering approaches requires further validation, whereas graph theory provides a powerful tool for microbial ecology and the analyses of environmental HTS datasets.

Keywords: Environmental diversity inventories; Novel diversity; High-throughput sequencing data; Sequence clustering; Bioinformatics

INTRODUCTION

High-throughput sequencing (HTS) technologies have fundamentally changed our perceptions and concepts of environmental protist diversity (Amaral-Zettler et al., 2009; de Vargas et al., 2015; Logares et al., 2014; Massana et al., 2015; Stoeck et al., 2009). Current HTS surveys can analyze protist communities by targeting specific molecular markers, resulting in datasets of many millions of sequencing reads that can address community-comparative, ecosystem-functioning, and novel-diversity questions (Dunthorn et al., 2014b). The detection of novel diversity, in specific, uses a strategy that detects reads distantly related to previously sequenced species (*e.g.* Berney et al., 2013; Dunthorn et al., 2014b; Edgcomb et al., 2011b; Filker et al., 2014; Hartikainen et al., 2014; Gimmler and Stoeck 2015). While the detection and description of novel protists is a central task since our understanding of their diversity is far from complete (Pawlowski et al., 2012), our ability to detect novel diversity in molecular environmental studies is affected by the way in which we cluster reads into operational taxonomic units (OTUs).

A traditional method of constructing *de novo* OTUs is by the popular program USEARCH (Edgar, 2010), though several other similar alternatives exist (*e.g.* Fu et al., 2012; Ghodsi et al., 2011; Schloss et al., 2009). USEARCH and these other related programs initiate OTUs by selecting an amplicon (*i.e.*, a dereplicated read) to serve as a centroid. Every amplicon within a global similarity value from the centroid, based on a pairwise comparison score, joins the OTU. The OTU is then closed, and its radius (or diameter, depending on the method used) from the centroid is that global similarity value. There is no consensus on which global similarity value should be used because taxa evolve at different rates (Brown et al., 2015; Caron et al., 2009; Nebel et al., 2011); a 97% value is commonly used in protist studies

(Edgcomb et al., 2011a; Massana et al., 2015), although higher values are also used (Egge et al., 2015).

A second method of constructing *de novo* OTUs is by sequence similarity networks (Forster et al., 2015). Each node in these networks stands for one amplicon, and two nodes are connected by an edge only if their amplicons are within a single global similarity value that is computed by pairwise alignment scores. Groups of nodes can form enclosed connected components that can be used as OTUs (Forster et al., 2015). Since additional nodes are added iteratively, the radius of a connected component is not pre-defined, but can be any value, including higher than the global similarity value. As with USEARCH, there is no agreement upon which global similarity threshold should be used. Unlike USEARCH, sequence similarity networks result in OTUs that exhibit an internal network topology that can be further evaluated by graph theory analyses (Junker and Schreiber, 2011; Newman, 2010).

A third method to define *de novo* OTUs is by the program Swarm (Mahé et al., 2015, 2014). Unlike USEARCH and sequence similarity networks, Swarm relies on an iterative, single-linkage algorithm that clusters using a local value. This value is user-defined (1 by default), and corresponds to the maximum number of differences due to substitutions or insertions/deletions in a global pairwise alignment. Swarm begins by choosing a starting amplicon for an OTU and links all amplicons with *d* or less differences. Then those newly linked amplicons are themselves linked to all amplicons with *d* or less differences and so on. Swarm does not designate OTU centroids *a priori*; instead the most abundant amplicon is picked as an OTU representative. In practice, these most abundant amplicons act similar to centroids by attracting less abundant amplicons in their vicinity, although the final results are

robust to changes of amplicon input order. Like sequence similarity networks, Swarm results in OTUs whose radii can be any value. Also like sequence similarity networks, the internal links among the amplicons in Swarm's OTUs can be plotted as networks and evaluated by graph theory

To compare how USEARCH, sequence similarity networks, and Swarm affect our ability to uncover novel diversity in protists, we used amplicon data derived from samples taken in European coastal marine habitats. To place the environmental amplicons into a taxonomic context of already known diversity, we obtained accessions from the curated Protist Ribosomal Reference database (PR^2) (Guillou et al., 2012). With this combination of unknown-environmental and taxonomically-identified amplicons, we asked: i) Do all three clustering approaches predict the same amount of novelty diversity? ii) Does graph theory discover additional novel diversity within OTUs that have underlying network topologies?

MATERIAL AND METHODS

Datasets

For environmental amplicons of benthic and planktonic eukaryotes, we used already published data from the BioMarKs Consortium (www.biomarks.eu) that sampled six near-shore marine sites in Norway, France, Spain, Italy and Bulgaria (*e.g.* Bittner et al., 2013; Dunthorn et al., 2014a; Logares et al., 2014; Massana et al., 2015). The sample design and sample processing, as well as Roche/454 GS FLX Titanium sequencing of the V4 region of 18S rDNA, is detailed in Massana et al. (2015). Raw reads were quality filtered and checked for chimeras with both UCHIME (Edgar, 2010) and ChimeraSlayer (Haas et al., 2011). The 1,476,249 cleaned V4 DNA and RNA reads were dereplicated into 312,503 strictly identical amplicons using a custom bash script. This and all other scripts can be found online in HTML format (Supplementary File 1).

For reference amplicons, we used PR^2 v203 taxonomic reference database (Guillou et al., 2012). From this database we extracted 109,021 taxonomically identified V4 amplicons. We then combined these reference amplicons with the environmental amplicons for all further downstream analyses.

Clustering

Three *de novo* clustering approaches were used to cluster the combined amplicons. First, USEARCH v8.0.1623 (Edgar, 2010), with a 97% global similarity value using options *-cluster smallmem* and *-sortedby size*. Second, basic network topology information was gathered by running a global pairwise alignment analysis in VSEARCH v1.1.3 (https://github.com/torognes/vsearch) using options *-allpairs_global* and *-iddef 1*. The resulting matrix contained 682,621,198 edges with

a weight of at least 97% global similarity value. Based on this matrix we created sequence similarity networks in R version 3.2.1 (http://r-project.org) using 'igraph' scripts (Csardi and Nepusz, 2006). Third, SWARM v2.1.1 (Mahé et al., 2015, 2014), with *-d* = 1 and *-f*. Singleton and doubleton OTUs were removed from the results of all three clustering approaches for downstream analyses.

Analyses

For each clustering approach we distinguished if an OTU consisted of: i) both environmental and reference amplicons, ii) exclusively reference amplicons, and iii) exclusively environmental amplicons. The number of reads in each OTU was also counted.

To compare the novel diversity uncovered by each clustering approach, we analyzed OTUs consisting of exclusively environmental amplicons. For each amplicon in exclusively environmental OTUs, we conducted global pairwise alignments of these amplicons with all PR^2 references in separate VSEARCH (using options -usearch_*global*, *-iddef 1* and *-id 0.70*) analyses, and gathered the highest global alignment score in % similarity to any reference sequence. This revealed how divergent the novel diversity of each clustering approach was with regard to taxonomically identified references. We also compared if the same environmental amplicons were classified as novel diversity among the different approaches.

To compare within OTUs, shortest path analyses were conducted within each sequence similarity network and within each Swarm OTU with 'igraph' scripts. The shortest path concept emerges from graph theory and exploits connections between nodes in a network (Newman, 2010). In this particular case we used shortest path analyses to find the minimal number of edges (*i.e.* links) that have to be crossed

within an OTU to move from each environmental node to its closest reference node. If an environmental node and a reference node were directly linked (*i.e.* direct neighbors separated by exactly one edge), they exhibited a distance of '1' to each other; we defined these environmental nodes as the part of diversity that is well represented by the PR^2 reference database. Environmental nodes that were not directly linked to reference nodes exhibited a distance of '≥2', and were thus indirectly linked. Environmental nodes in OTUs, which exclusively consisted of environmental amplicons, exhibited 'infinite' distances to all reference nodes since no shortest path existed. We defined all environmental nodes with distances '≥2' to reference nodes as novel variants of diversity that are currently not covered by the PR^2 database.

Results and Discussion

Contrasting OTU results from three approaches

The number of resulting OTUs varied across the three clustering approaches (Table 1). The fewest OTUs in total were produced by sequence similarity networks (n= 8,202). Sequence similarity networks also produced the fewest OTUs containing both environmental and reference amplicons (n= 1,619), containing exclusively reference amplicons (n= 3,138), and containing exclusively environmental amplicons (n= 3,445). This approach was especially effective in linking environmental and reference amplicons: it had the most amplicons in OTUs containing both types (n= 253,965 environmental and n= 54,988 reference). On the other hand, this also led to fewer amplicons in exclusively environmental OTUs (n= 47,116), meaning that sequence similarity networks identified the least novel diversity in terms of both amplicons and OTUs.

USEARCH produced more OTUs in total (n= 12,427) and more OTUs (n= 5,342) that contained exclusively environmental amplicons (n= 71,337). The fraction of novel amplicons was therefore increased by one third in USEARCH. These differences in OTU numbers may be due in part to how the two methods use their global clustering values: while connected components in sequence similarity networks grow iteratively, OTUs in USEARCH are restricted to the radius of the centroid amplicons (Mahé et al., 2014). Amplicons whose sequences are more than 97% divergent from the centroid are consequently split from the OTU, although they might be less than 97% divergent from other amplicons of the OTU. This behavior of USEARCH and other closely-related methods results in an over-splitting of OTUs (Flynn et al., 2015; Mahé et al., 2014) compared to sequence similarity networks. Additionally, this behavior also causes OTU instability, meaning that a re-clustering

with USEARCH may result in slightly different OTU sizes and membership, especially if the input order of the amplicons is shuffled (He et al., 2015; Mahé et al., 2014). Since both factors are especially important for an accurate detection of novel diversity, we argue that the more conservative results of the sequence similarity networks are less prone to contain amplicons and OTUs that are spuriously classified as novel.

Although not tested here, previous studies have shown that all-vs.-all pairwise comparison clustering approaches such as sequence similarity networks generally produce more reliable and stable OTUs than heuristic clustering methods such as USEARCH (Schmidt et al., 2015; Sun et al., 2011). This higher reliability and stability of all-vs.-all pairwise comparisons comes at the cost of extensive computational time (Flynn et al., 2015; Sun et al., 2011), which increases with the square to the number of input sequences (Bik et al., 2012). By calculating a pairwise comparison matrix of the currently largest dataset of near-shore marine protists in Europe, we operated close to the limit of data that can be handled in all-vs.-all current approaches.

Compared to both approaches relying on global clustering values, Swarm, with its local clustering value, produced the most OTUs in total (n= 13,240). The Swarm approach also produced the most OTUs (n= 6,228) that contained exclusively environmental amplicons (n= 81,073). These higher numbers of OTUs in total and OTUs containing exclusively environmental amplicons may be due to Swarm's high clustering stringency that iteratively links amplicons with a small number of differences to each other. On the other hand, these high numbers may be due to missing intraspecific sequence variation in the PR^2 reference database, which usually contains only one accession per species. In natural communities, intraspecific genetic variation of microbial organisms may be much more diverse

than just a few base pair differences, especially in hypervariable gene regions (Brown et al., 2015; Decelle et al., 2014; Dunthorn et al., 2012; Pernice et al., 2013). But in Swarm, an environmental amplicon that differs by more than one base pair to reference accessions will be placed into a novel OTU, if there are no intermediate amplicons linking them. As long as reference databases are not designed to cover intraspecific sequence variation, it is a more effective strategy to compute Swarm OTUs from datasets consisting entirely of environmental amplicons, and perform a later taxonomical assignment; *e.g.* as in de Vargas et al. (2015), Filker et al. (2014) and Gimmler and Stoeck (2015).

Is novel diversity really novel?

After the identification of novel variants of OTUs and amplicons, the next step in the discovery of novel diversity is normally the design of specific primers and probes for the targeted recovery of organisms from environmental samples (Edgcomb et al., 2011b; Gimmler and Stoeck, 2015; Hartikainen et al., 2014; Orsi et al., 2012; Seenivasan et al., 2013). However, this process is time-, cost-, and labor-intensive. An accurate initial classification of novel diversity by clustering approaches is therefore crucial.

There were 29,059 environmental amplicons that were classified as novel by all three clustering approaches (Figure 1). However, the number of environmental amplicons classified as novel exclusively by one approach differed dramatically: 1,232 in sequence similarity networks, 13,777 in USEARCH, and 40,132 in Swarm. Most environmental amplicons which shared less than 97% sequence similarity with references in PR^2 were congruently classified as novel by all three approaches. But both USEARCH and Swarm classified as novel numerous amplicons that were more

than 97% similar to PR^2 references (Figure 2, Supplemental Figure S1). Even though clustering in USEARCH was performed at 97% similarity to delineate novel environmental amplicons from amplicons representing previously described diversity, we found 15,438 amplicons in exclusively environmental OTUs with more than 97% similarity to PR^2 references; for Swarm this fraction amounted to 47,007 amplicons. The even larger overestimation of novel diversity by Swarm is caused by a combination of the approach's high clustering stringency and missing intraspecific variation in the PR^2 database. On the other hand, sequence similarity networks classified no environmental amplicon inadvertently as novel, thereby supporting our argument of more accurate novel diversity detection in the latter approach. Furthermore, this supports the assumption that during network construction from global pairwise alignment scores, pairwise connections below 97% sequence similarity to PR^2 species references were successfully excluded. We conclude that the conservative results of sequence similarity networks most closely match our definition of how we delineated novel diversity from previously described diversity at a given global clustering threshold value.

Beyond that, 97% of the novel diversity amplicons in sequence similarity networks were identified as novel by at least one of the other two clustering approaches (Figure 1). On the other hand, the 1,232 amplicons exclusively identified as novel by sequence similarity networks clustered into singletons or doubletons in USEARCH and Swarm and were thus excluded from downstream analyses. The novel diversity uncovered by sequence similarity networks therefore comes closest to a common denominator of amplicons that are truly less than 97% similar to references in PR^2. Furthermore, we strongly advise to perform an additionally taxonomic assignment step in Swarm and USEARCH to validate if potential novel

diversity is indeed highly diverse from deposited references. At the same time, though, we are aware that even amplicons which are highly similar to entries in reference databases may represent novel genetic variants. Such hidden diversity is unlikely to be unveiled by approaches solely relying on global similarity values. Instead, more stringent approaches that trace local substitutions or methods which explore internal OTU structure stand a higher chance of revealing novel genetic variants, since they provide a higher resolution of genetic diversity.

Graph theory allows a more detailed evaluation of HTS datasets

Beyond just being able to relay the number of OTUs, sequence similarity networks and Swarm provided additional underlying information for each of their OTUs in the form of network topologies. As pointed out by Forster et al. (2015), these network topologies can reveal additional within-OTU connections among environmental and reference amplicons by using shortest path analyses.

With sequence similarity networks and OTUs containing both types of amplicons, 239,472 of the 253,965 environmental amplicons were directly connected to reference amplicons (Figure 1). The other 14,493 environmental amplicons were indirectly connected to reference amplicons, and represent novel genetic variation on top of the 47,116 amplicons already placed into exclusively environmental OTUs. With Swarm and OTUs containing both types of amplicons, only 5,757 of the 142,946 environmental amplicons were directly connected to reference amplicons. The 137,189 environmental amplicons with indirect connections also represent novel genetic variation along side of the 81,073 amplicons in exclusively environmental OTUs. This large number of indirectly connected amplicons in Swarm may be an overestimation because current reference databases do not yet cover intraspecific

sequence variation (see above). However, our analyses are a first indication that shortest path analyses are a promising way to explore Swarm OTUs. By analyzing paths within an OTU one could, for example, investigate whether amplicons from the same sampling site are more often directly connected to each other than to amplicons from another site. Thus screening for genetic variation related to regional populations or species.

Nevertheless, shortest path analyses are just one way to explore genetic variance and novel diversity within OTUs with network topologies. Graph theory can be used to ask numerous questions in microbial ecology (Junker and Schreiber, 2011; Newman, 2010; Proulx et al., 2005). For instance, analyses of assortativity can indicate if environmental sequences from a certain habitat more preferentially connect with reference sequences than environmental sequences from another habitat (Forster et al., 2015), thus revealing which habitat's microbial community is less adequately covered by reference databases.

Conclusions

Each of the three clustering approaches provided different perspectives on microbial diversity, while also showing individual weaknesses. Despite these weaknesses, we argue that the combination of high stringency clustering methods and sequence similarity networks, and the implementation of further tools from graph theory will be beneficial for the evaluation of HTS datasets. Such tools will uncover underlying patterns from microbial HTS data, which hold important information about environmental microbial communities.

References

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., Huse, S.M., 2009. A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. PLoS ONE 4, e6372. doi:10.1371/journal.pone.0006372

Berney, C., Romac, S., Mahé, F., Santini, S., Siano, R., Bass, D., 2013. Vampires in the oceans: predatory cercozoan amoebae in marine habitats. ISME J. 7, 2387–2399. doi:10.1038/ismej.2013.116

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R., Thomas, W.K., 2012. Sequencing our way towards understanding global eukaryotic biodiversity. Trends Ecol. Evol. 27, 233–243. doi:10.1016/j.tree.2011.11.010

Bittner, L., Gobet, A., Audic, S., Romac, S., Egge, E.S., Santini, S., Ogata, H., Probert, I., Edvardsen, B., de Vargas, C., 2013. Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. Mol. Ecol. 22, 87–101. doi:10.1111/mec.12108

Brown, E.A., Chain, F.J.J., Crease, T.J., MacIsaac, H.J., Cristescu, M.E., 2015. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecol. Evol. 5, 2234–2251. doi:10.1002/ece3.1485

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D., Dennett, M.R., Moran, D.M., Jones, A.C., 2009. Defining DNA-Based Operational Taxonomic Units for Microbial-Eukaryote Ecology. Appl. Environ. Microbiol. 75, 5797–5808. doi:10.1128/AEM.00298-09

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal Complex Syst. 1695, 1–9.

Decelle, J., Romac, S., Sasaki, E., Not, F., Mahé, F., 2014. Intracellular Diversity of the V4 and V9 Regions of the 18S rRNA in Marine Protists (Radiolarians) Assessed by High-Throughput Sequencing. PLoS ONE 9, e104297. doi:10.1371/journal.pone.0104297

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Bescot, N.L., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E.G., Sardet, C., Sullivan, M.B., Velayoudon, D., 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605. doi:10.1126/science.1261605

Dunthorn, M., Klier, J., Bunge, J., 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. doi:10.1111/j.1550-7408.2011.00602.x

Dunthorn, M., Otto, J., Berger, S.A., Stamatakis, A., Mahé, F., Romac, S., Vargas, C. de, Audic, S., Consortium, B., Stock, A., Kauff, F., Stoeck, T., 2014a. Placing Environmental Next-Generation Sequencing Amplicons from Microbial Eukaryotes into a Phylogenetic Context. Mol. Biol. Evol. 31, 993–1009. doi:10.1093/molbev/msu055

Dunthorn, M., Stoeck, T., Clamp, J., Warren, A., Mahé, F., 2014b. Ciliates and the Rare Biosphere: A Review. J. Eukaryot. Microbiol. 61, 404–409. doi:10.1111/jeu.12121

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. doi:10.1093/bioinformatics/btq461

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., Holder, M., Taylor, G.T., Suarez, P., Varela, R., Epstein, S., 2011. Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. ISME J. 5, 1344–1356. doi:10.1038/ismej.2011.6

Edgcomb, V.P., Orsi, W., Breiner, H.-W., Stock, A., Filker, S., Yakimov, M.M., Stoeck, T., 2011. Novel active kinetoplastids associated with hypersaline anoxic basins in the Eastern Mediterranean deep-sea. Deep Sea Res. Part Oceanogr. Res. Pap. 58, 1040–1048. doi:10.1016/j.dsr.2011.07.003

Egge, E.S., Johannessen, T.V., Andersen, T., Eikrem, W., Bittner, L., Larsen, A., Sandaa, R.-A., Edvardsen, B., 2015. Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing. Mol. Ecol. n/a–n/a. doi:10.1111/mec.13160

Filker, S., Gimmler, A., Dunthorn, M., Mahé, F., Stoeck, T., 2014. Deep sequencing uncovers protistan plankton diversity in the Portuguese Ria Formosa solar saltern ponds. Extremophiles 19, 283–295. doi:10.1007/s00792-014-0713-2

Flynn, J.M., Brown, E.A., Chain, F.J.J., MacIsaac, H.J., 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. doi:10.1002/ece3.1497

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., Lopez, P., Stoeck, T., Bapteste, E., 2015. Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol. 13, 16. doi:10.1186/s12915-015-0125-5

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. doi:10.1093/bioinformatics/bts565

Ghodsi, M., Liu, B., Pop, M., 2011. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics 12, 271. doi:10.1186/1471-2105-12-271

Gimmler, A., Stoeck, T., 2015. Mining environmental high-throughput sequence data sets to identify divergent amplicon clusters for phylogenetic reconstruction and morphotype visualization. Environ. Microbiol. Rep. 7, 679–686.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., Vargas, C. de, Decelle, J., del Campo, J., Dolan, J.R., Dunthorn,

M., Edvardsen, B., Holzmann, M., Kooistra, W.H.C.F., Lara, E., Bescot, N.L., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., Christen, R., 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. gks1160. doi:10.1093/nar/gks1160

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., Methé, B., DeSantis, T.Z., Consortium, T.H.M., Petrosino, J.F., Knight, R., Birren, B.W., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 21, 494–504. doi:10.1101/gr.112730.110

Hartikainen, H., Ashford, O.S., Berney, C., Okamura, B., Feist, S.W., Baker-Austin, C., Stentiford, G.D., Bass, D., 2014. Lineage-specific molecular probing reveals novel diversity and ecological partitioning of haplosporidians. ISME J. 8, 177–186. doi:10.1038/ismej.2013.136

He, Y., Caporaso, J.G., Jiang, X.-T., Sheng, H.-F., Huse, S.M., Rideout, J.R., Edgar, R.C., Kopylova, E., Walters, W.A., Knight, R., others, 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome 3, 20.

Junker, B.H., Schreiber, F., 2011. Analysis of Biological Networks. John Wiley & Sons.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Gobet, A., Kooistra, W.H.C.F., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Romac, S., Shalchian-Tabrizi, K., Simon, N., Stoeck, T., Santini, S., Siano, R., Wincker, P., Zingone, A., Richards, T.A., de Vargas, C., Massana, R., 2014. Patterns of Rare and Abundant Marine Microbial Eukaryotes. Curr. Biol. 24, 813–821. doi:10.1016/j.cub.2014.02.050

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

Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M., 2014. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2, e593. doi:10.7717/peerj.593

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W.H.C.F., Logares, R., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Probert, I., Romac, S., Richards, T., Santini, S., Shalchian-Tabrizi, K., Siano, R., Simon, N., Stoeck, T., Vaulot, D., Zingone, A., Vargas, C., 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ. Microbiol. doi:10.1111/1462-2920.12955

Nebel, M., Pfabel, C., Stock, A., Dunthorn, M., Stoeck, T., 2011. Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. Environ. Microbiol. Rep. 3, 154–158. doi:10.1111/j.1758-2229.2010.00200.x

Newman, M., 2010. Networks: an introduction. Oxford University Press.

Orsi, W., Edgcomb, V., Faria, J., Foissner, W., Fowle, W.H., Hohmann, T., Suarez, P., Taylor, C., Taylor, G.T., Vd'ačný, P., Epstein, S.S., 2012. Class Cariacotrichea, a novel ciliate taxon from the anoxic Cariaco Basin, Venezuela. Int. J. Syst. Evol. Microbiol. 62, 1425–1433. doi:10.1099/ijs.0.034710-0

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S.S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A.M., Gile, G.H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P.J., Kostka, M., Kudryavtsev, A., Lara, E., Lukeš, J., Mann, D.G., Mitchell, E.A.D., Nitsche, F., Romeralo, M., Saunders, G.W., Simpson, A.G.B., Smirnov, A.V., Spouge, J.L., Stern, R.F., Stoeck, T., Zimmermann, J., Schindel, D., de Vargas, C., 2012. CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLoS Biol 10, e1001419. doi:10.1371/journal.pbio.1001419

Pernice, M.C., Logares, R., Guillou, L., Massana, R., 2013. General Patterns of Diversity in Major Marine Microeukaryote Lineages. PLoS ONE 8, e57170. doi:10.1371/journal.pone.0057170

Proulx, S.R., Promislow, D.E., Phillips, P.C., 2005. Network thinking in ecology and evolution. Trends Ecol. Evol. 20, 345–353.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Horn, D.J.V., Weber, C.F., 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. 75, 7537–7541. doi:10.1128/AEM.01541-09

Schmidt, T.S.B., Matias Rodrigues, J.F., von Mering, C., 2015. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. Environ. Microbiol. 17, 1689–1706. doi:10.1111/1462-2920.12610

Seenivasan, R., Sausen, N., Medlin, L.K., Melkonian, M., 2013. Picomonas judraskeda Gen. Et Sp. Nov.: The First Identified Member of the Picozoa Phylum Nov., a Widespread Group of Picoeukaryotes, Formerly Known as “Picobiliphytes.” PLoS ONE 8, e59565. doi:10.1371/journal.pone.0059565

Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M.J., Chistoserdov, A., Orsi, W., Edgcomb, V.P., 2009. Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. BMC Biol. 7, 1–20. doi:10.1186/1741-7007-7-72

Sun, Y., Cai, Y., Huse, S.M., Knight, R., Farmerie, W.G., Wang, X., Mai, V., 2011. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Brief. Bioinform. bbr009. doi:10.1093/bib/bbr009

ACKNOWLEDGEMENTS

We would like to thank the computational resources at the Regional Computing Center at the University of Kaiserslautern, and the BioMarKs consortium for the data analyzed in this study.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

DF was supported by a graduate scholarship of Stipdenstiftung Rheinland-Pfalz. MD and FM were supported by the Deutsche Forschungsgemeinschaft (grant #DU1319/1-1). TS was supported by the Deutsche Forschungsgemeinschaft (grant #STO414/11-1).

Grant Disclosures

The following grant information was disclosed by the authors:

Deutsche Forschungsgemeinschaft: DU1319/1-1.

Deutsche Forschungsgemeinschaft: STO414/11-1.

Tables and Figures

Table 1 Sequence clustering results of the three tested approaches. Indicated is the amount of OTUs and the amount (and type) of amplicons within these OTUs for each class of OTUs defined in our analyses.

	USEARCH	Sequence similarity networks	Swarm
OTUs	12427	8202	13240
OTUs containing environmental and reference amplicons	2527	1619	1993
Environmental amplicons	223735	253965	142946
Reference amplicons	33386	54988	18774
OTUs containing exclusively reference amplicons	4558	3138	5019
Reference amplicons	59368	46255	49147
OTUs containing exclusively environmental amplicons	5342	3445	6228
Environmental amplicons	71337	47116	81073

Figure 1 Venn-Diagram of the number of amplicons in exclusively environmental OTUs. The area of each clustering approach was proportionally adjusted to the amount of amplicons in environmental OTUs detected in that approach. Overlapping areas reflect amplicons detected in each of the respective approaches. Numbers indicate how many amplicons are represented by each area whereas each area's size is proportional to the number of amplicons included.

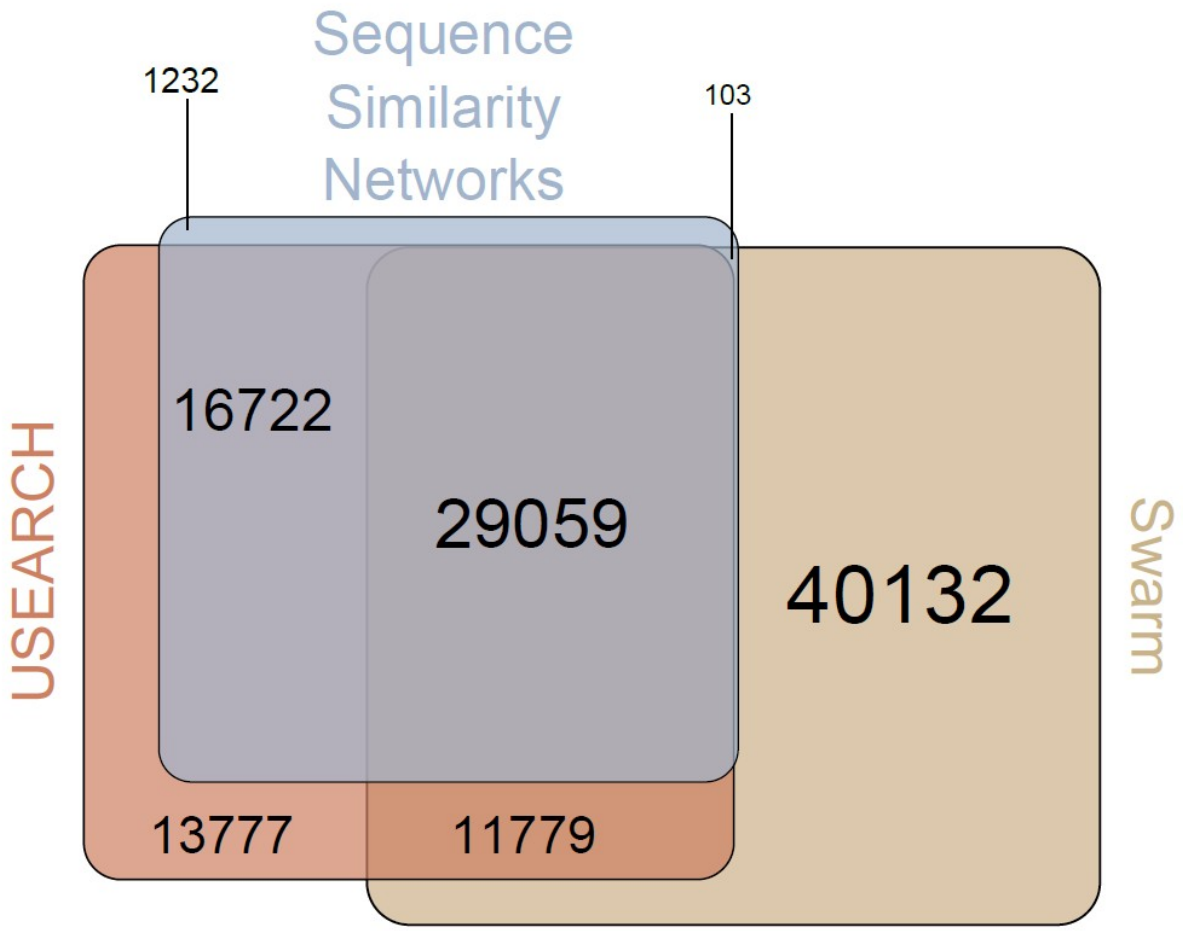


Figure 2 Genetic divergence of amplicons in exclusively environmental OTUs to PR^2 references by clustering approach. Each point represents one amplicon clustered into an exclusively environmental OTU by the respective clustering approach. Position on the x-axis gives the abundance of each amplicon in the initial dataset before dereplication. The y-axis gives the highest pairwise sequence similarity score of an amplicon to any entry in the PR^2 database as calculated by VSEARCH.

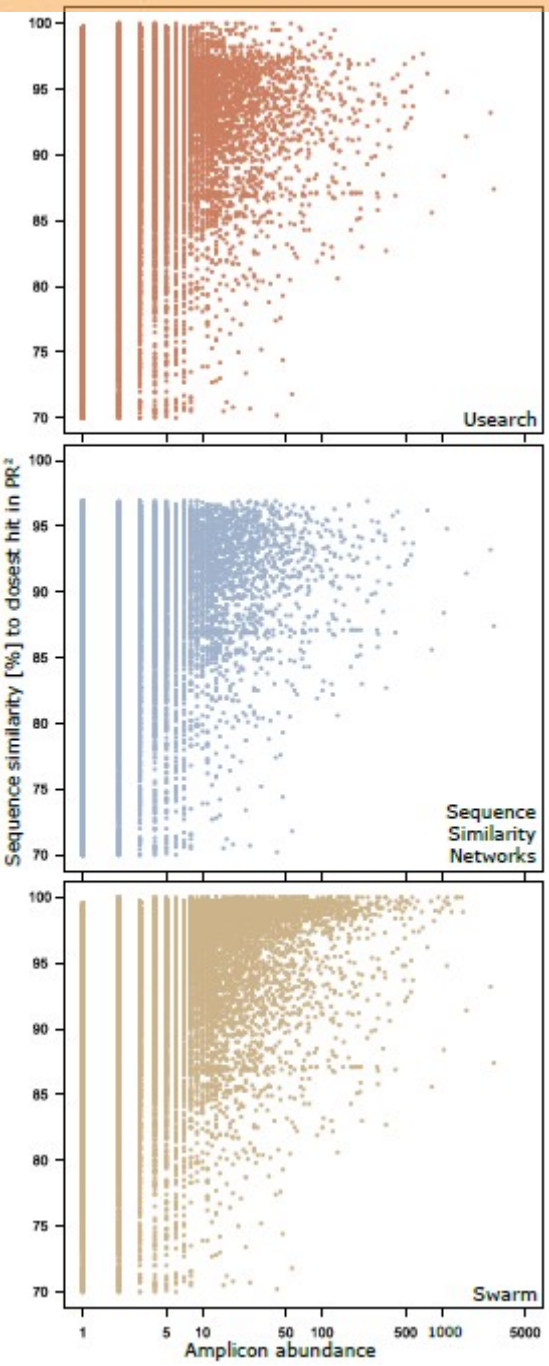
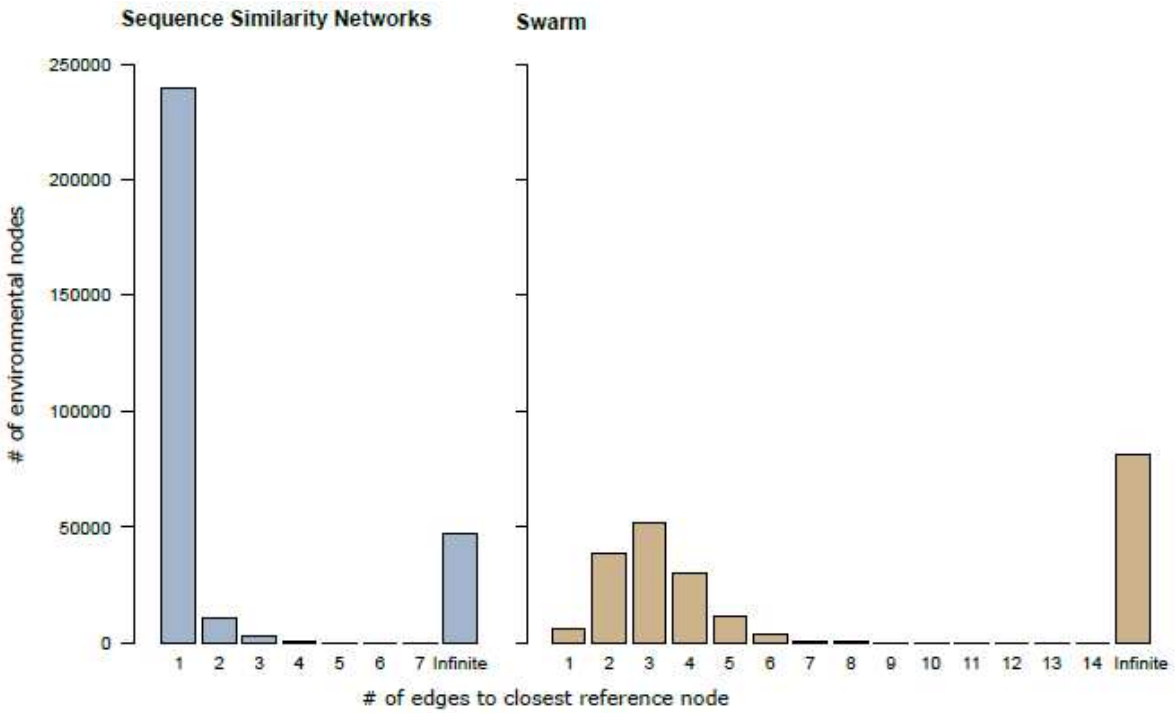

Sequence similarity [%] to closest hit in PR²
100
95
90
85
80
75
70
Usearch
Sequence Similarity Networks
Swarm
1
5
10
50
100
500
1000
5000
Amplicon abundance

Figure 3 Shortest path analyses of CCs and swarms. The barplots illustrate how many edges separated each environmental amplicon from its closest reference amplicon in sequence similarity networks and Swarm. A distance of '1' edge means that the environmental amplicon was directly connected to a reference (i.e. at least 97% similarity for Sequence Networks, and at least 99.7% similarity for Swarm given the average amplicon length). 'Infinite' means that the environmental amplicon was placed into an exclusively environmental OTU (see also Table 1) and did not exhibit any connection to a reference amplicon.



Supplemental Figure S1 Genetic divergence of amplicons in exclusively environmental OTUs exclusively detected by one clustering approach. The figure represents a subset of the data shown in Figure 2. Instead of showing all points, we specifically highlighted the fraction of amplicons which were clustered into exclusively environmental OTUs by exclusively one approach.