Supplementary material SX

In order to control for potential bias in protist diversity due to the use of two different forward primers to amplify the V9 region of the SSU in the different datasets used in this study, previous datasets using the same two forward primers for amplifying DNA of the same samples were re-analysed. The first re-analysed dataset (Amaral-Zettler et al., 2009, secondary study accession: SRP000903) comprised eight surface water samples from the Palmer Station Long Term Ecological Research Site (Antartica) and one sample from Mount Hope Bay (Massachusetts, USA), with the latest sample not being re-analysed here due to the absence of replicates. The second re-analysed dataset (Stoeck et al., 2009, secondary study accession: SRP001212) comprised four freshwater surface samples from Framvaren Fjord (Norway) and four marine surface water samples from Cariaco Basin (Venezuela). Raw sequence data and their associated metadata were retrieved from the European Nucleotide Archive inside the same bioproject PRJNA109315. Th raw reads were download in SRA format and then converted back to the original sff format using SRA Toolkit v. 2.10.8 (https://github.com/ncbi/sra-tools). Raw sff reads were converted to fastq files using mothur v. 1.44.1 (Schloss et al., 2009). Each run library consist of pool of reads amplified with both V9 SSU forward primers 1380F and 1389F, and the reverse primer 1510R. The 5-nt barcode was stripped of the 5'-end of each library, then forward and reverse primers where detected using the linked adapter approach of cutadapt v. 2.10 (Martin, 2011) allowing until 6 mismatches on each primer and keeping only the sequence between the primers. For reads amplified with the 1380F primer, the 5 first nt at the 5'-end were additionally removed using OBITools v. 1.2.13 (Boyer et al., 2016) in order to keep only the common region of the V9 SSU amplified by both forward primers. All reads were then pooled, dereplicated, sorted by decreasing abundances and clustered into OTUs using SWARM v. 3.0.0 (Mahé et al., 2015). OTU representative sequences were checked for chimera using the *de-novo* algorithm of UCHIME as implemented in VSEARCH. After chimera removal, representative sequences were assign to taxonomy using the VSEARCH global pairwise alignment against all PR2 v. 4.12.0 reference sequences (Guillou et al., 2013). Consensus taxonomic assignment were created for representative sequences with multiple best matches using a consensus threshold of 60%. Sequences from duplicates libraries of the same sample in the dataset SRP000903 were pooled. Singleton to tripleton OTUs were removed and protist only OTUs (i.e. excluding Fungi, Metazoa, unidentified Opisthokonta, Streptophyta and unidentified Eukaryota) were kept for downstream analyses.

The dataset SRP001212 presented a significant higher number of protist reads for the samples amplified with the 1389F forward primer (Mann-Whitney test, p=0.0015, Figure SX1), while the dataset SRP000903 had significant linear increase of protist OTU richness along with the log number of reads (ANOVA, p=0.035, Figure SX2). Thus, in order to remove those two sequencing depth bias, diversity indices were compute on rarefied OTU matrices, with randomly sampling 526 reads for the SRP000903 dataset and 4 253 reads for the SRP001212 dataset for each combination of samples and forward primers (8 samples times 2 forward primers for 16 combinations in each dataset). Rarefaction were bootstrapped 100 times in order to avoid loss of data.

Alpha diversity indices (i.e. OTU richness, Shannon diversity, Simpson diversity, Chao1, Abundance Coverage Estimator) did not showed any significant differences between the two primers in both datasets (Mann-Whitney test, p-values corrected for multiple tests (Benjamini and Hochberg, 1995), p>0.05, Figure SX3).

At the gamma diversity level, the majority of OTUs were shared by both primers in both datasets (SRP000903: 63.5 % ± 1.5 of 448 OTUs ± 6 ; SRP001212: 58.9 % ± 0.6 of 1942 OTUs ± 7) with the large majority of reads belonging to shared OTUs (SRP000903: 96.3 % ± 0.2 of 8 416 reads, SRP001212: 92 % ± 0.4 of 68 048 reads).

At the beta diversity level, no significant changes in community composition were recorded in bootstrap rarefied OTU matrices (Bray-Curtis dissimilarity, PERMANOVA test, R<0.05, p>0.05) as well as in the non-rarefied OTU matrices (Hellinger transformation and Bray-Curtis dissimilarity, PERMANOVA test, p>0.05; Figure SX4). Bray-Curtis dissimilarity values did not show any significant changes dues to primer in both datasets (Mann-Whitney test, p-values corrected for multiple tests, p>0.05; Figure SX5).

Taxonomic compositions, as characterized at the phylum level, did not show significant difference in the number of OTUs between primers in any of the bootstrap rarefaction (Mann-Whitney test, p-values corrected for multiple tests, p>0.05; Figure SX6). Only three taxa in the SRP001212 dataset (Dinoflagellata, Opalozoa and Haptophyta) displayed significant higher number of reads when using the primer 1380F, with significant p-values in 58, 57 and 99 of the 100 bootstraps (Mann-Whitney test, p-values corrected for multiple tests).

In conclusion, there is no identifiable bias of using either the forward primer 1380F or 1389F associated with the reverse primer 1510R to describe protist alpha, beta and gamma diversity, as well as the diversity and relative abundance of the main protist taxa. The apparent higher alpha and gamma diversity described in the two previous studies datasets re-analysed here (Amaral-Zettler et al., 2009; Stoeck et al., 2009) are most likely due to a bias in the sequencing depth which is driven by the primer choice. Indeed, each library sequenced were equimolar pools of PCR products from the same sample produced with either the primer pair 1380F + 1510R or 1389F + 1510R. However, the reads amplified with the 1380F primer are expected to be 14-nt longer than the reads produced by the 1389F primer, which represent approximately an increase of 10 % in sequence length. Thus, an equimolar pooling would have result in pooling less 1380F reads than 1389F reads and obviously resulted in in raw read libraries with more 1389F reads than 1380F reads. Here we show that the bootstrapped rarefaction approach efficiently removed the sequencing depth bias and no differences in diversity between the two primers could be observed after this rarefaction. Both 1380F and 1389F forward primer are equivalent to describe protist diversity in environmental sample.

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–e6372. https://doi.org/10.1371/journal.pone.0006372

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate - a practical and powerful approach. J. R. Stat. Soc. Ser. B-Methodol. 57, 289–300. https://doi.org/10.2307/2346101

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., Coissac, E., 2016. obitools: a unix-inspired software package for DNA metabarcoding. Mol. Ecol. Resour. 16, 176–182. https://doi.org/10.1111/1755-0998.12428

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahe, 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., 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41, D597–D604. https://doi.org/10.1093/nar/gks1160

Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M., 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3, e1420. https://doi.org/10.7717/peerj.1420

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200

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., Van Horn, D.J., 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. https://doi.org/10.1128/AEM.01541-09

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, 72. https://doi.org/10.1186/1741-7007-7-72

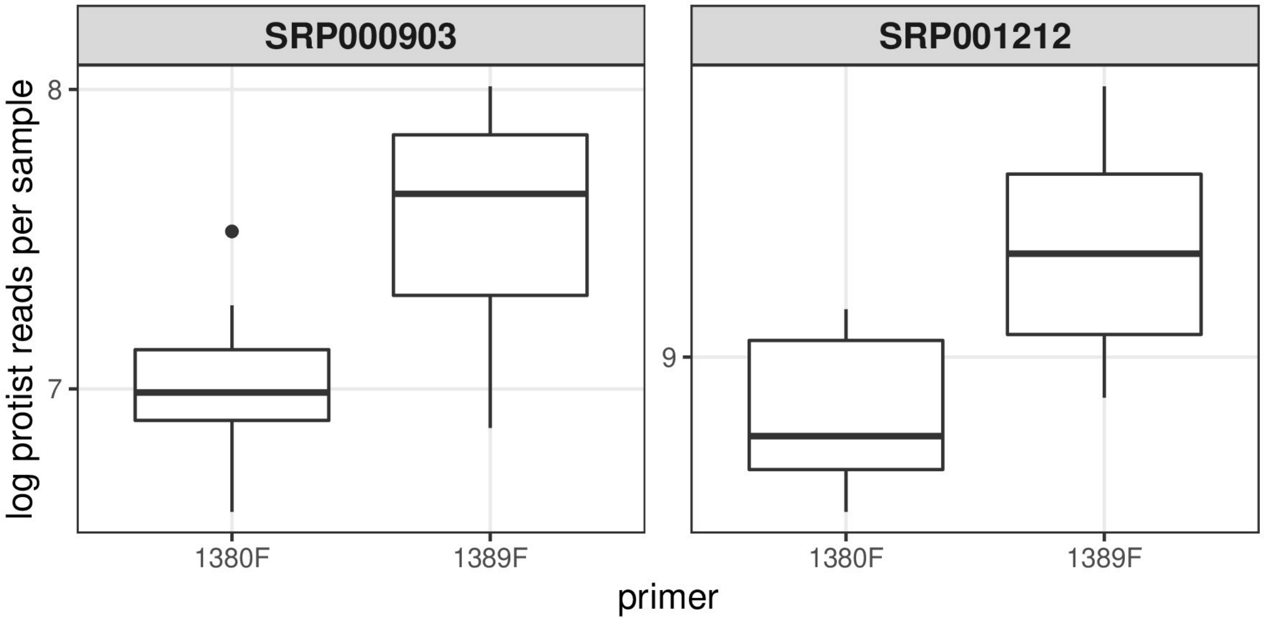

Figure SX1 Primers influence on protist read counts. The difference in log transformed read counts of protists OTUs between the two primers was only significant for the SRP001212 dataset (Mann-Whitney test, p-value<0.05).

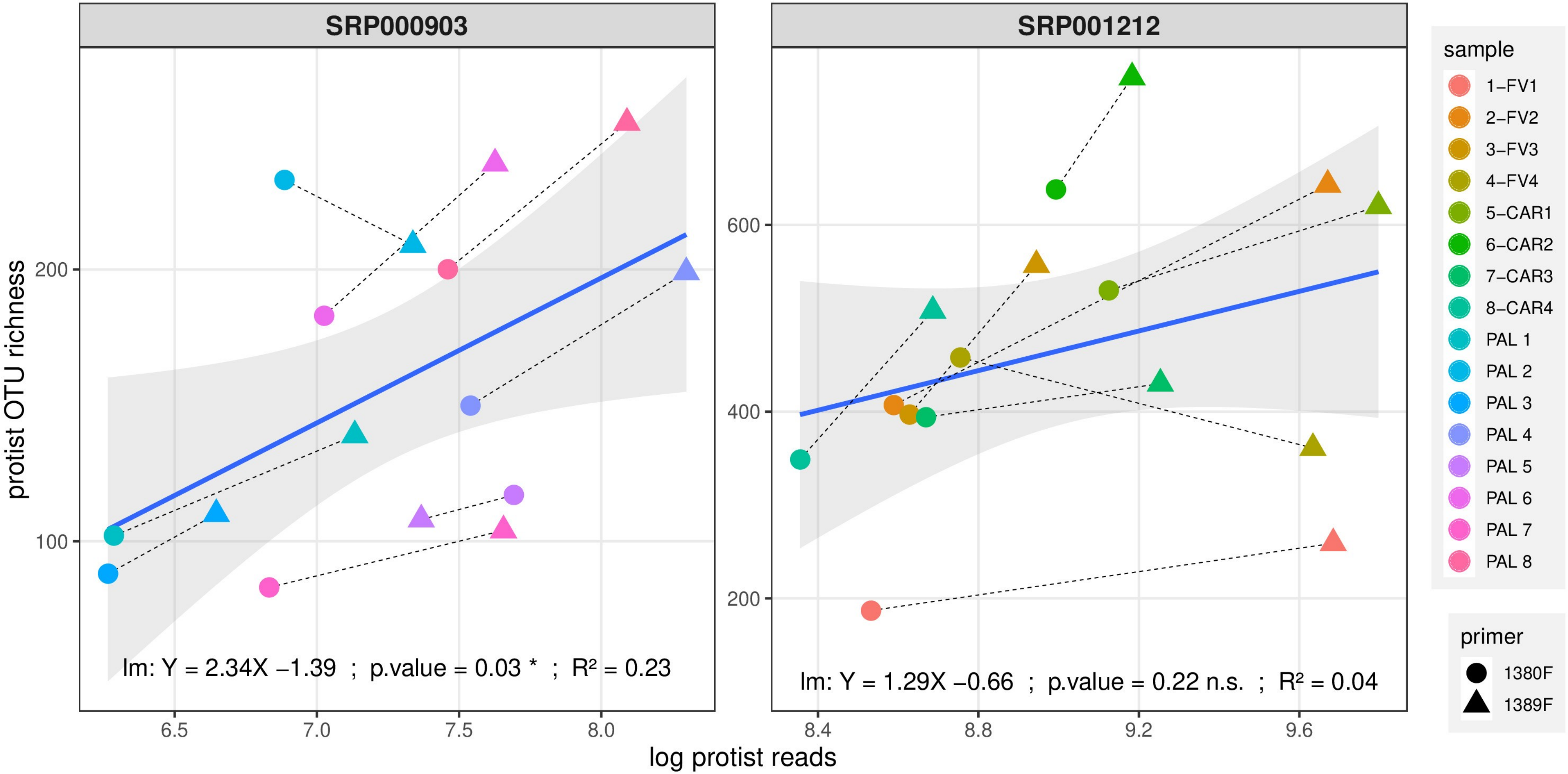

Figure SX2 Linear model between log read counts and OTU richness of protist OTUs.

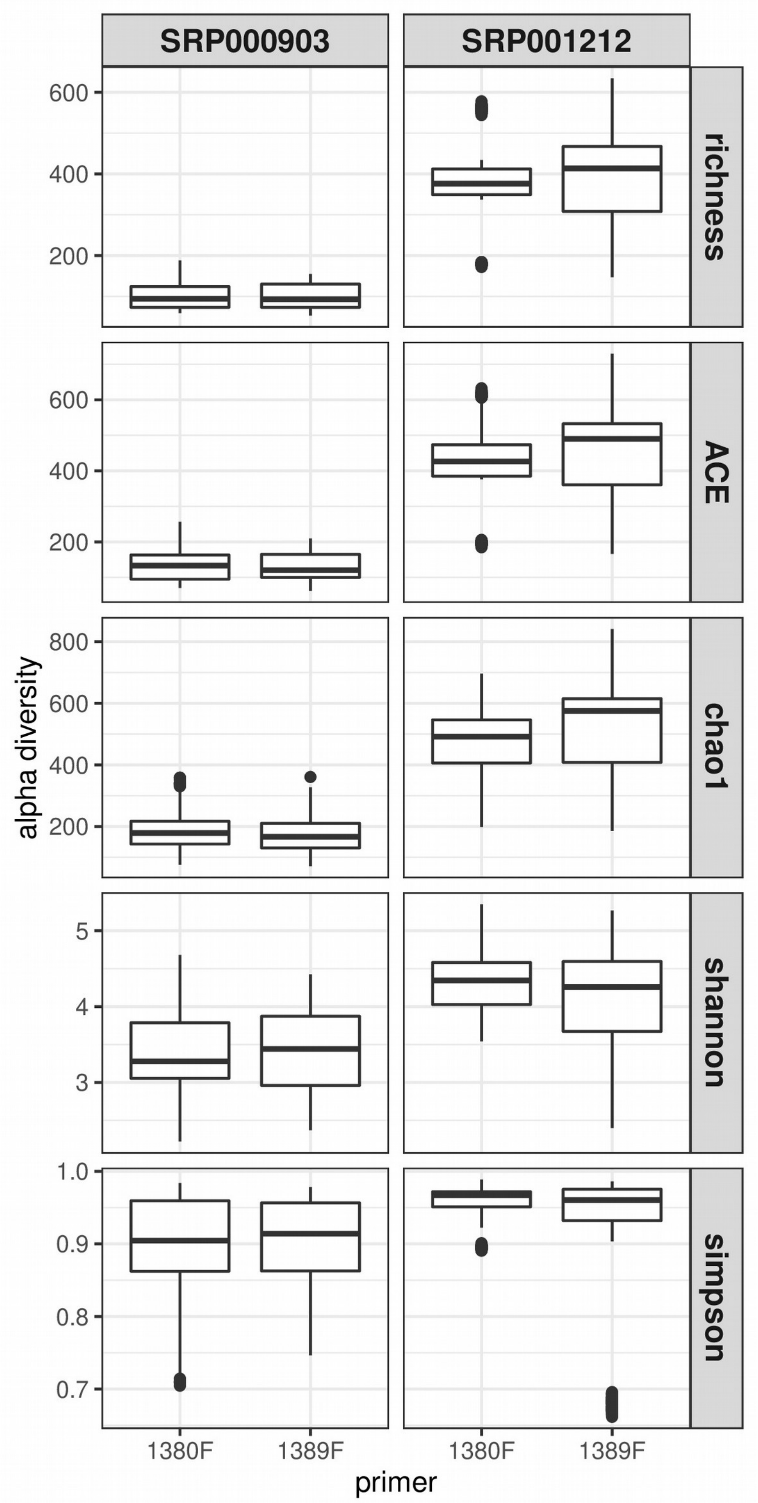

Figure SX3 Protist alpha diversity from all bootstraped rarefaction in function of the forward primer used.

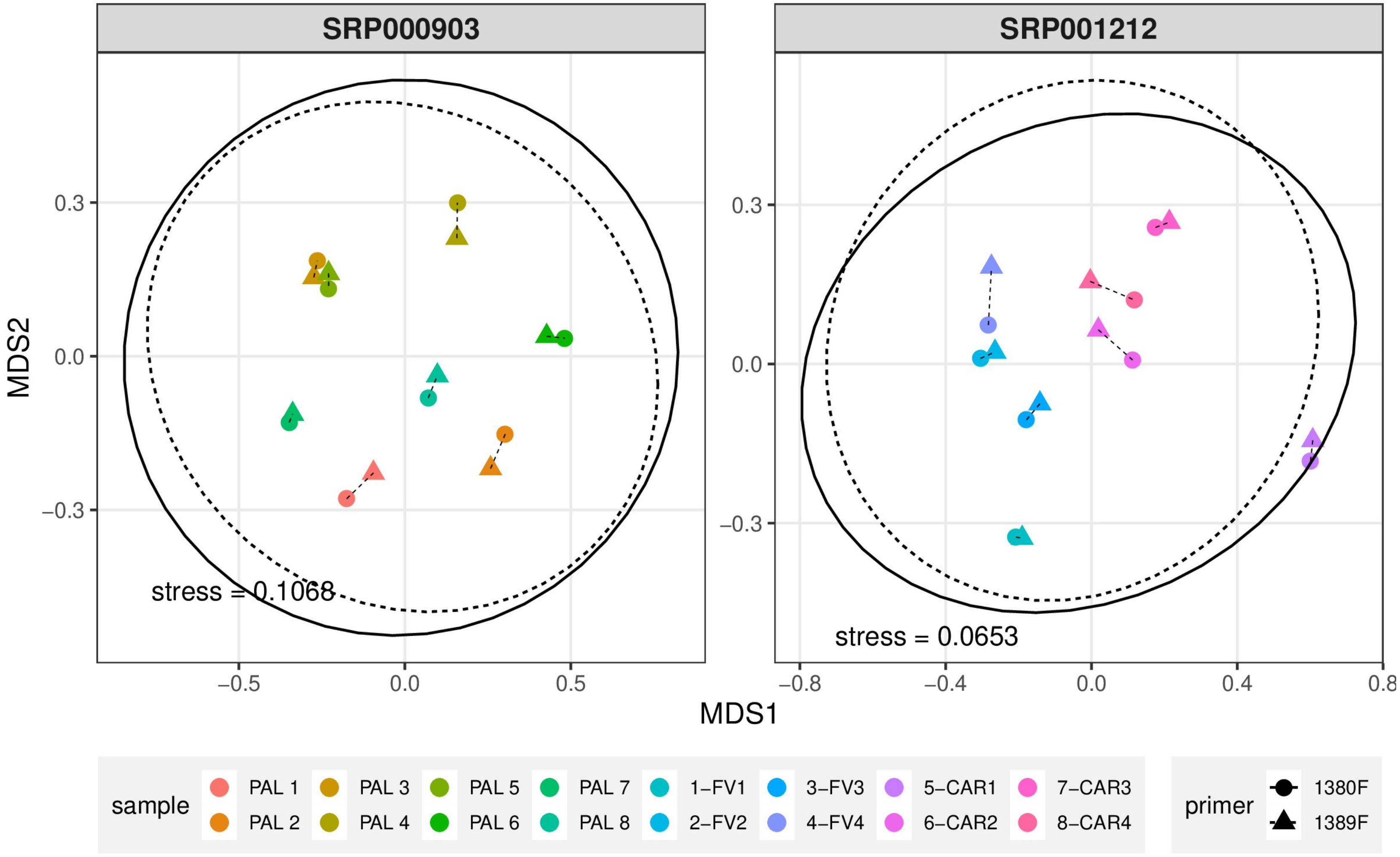

Figure SX4 NMDS based on Bray-Curtis dissimilarity values from Hellinger transformed (no bootstrap) OTU matrices. Plain and dotted lines ellipses represent 95 % confidence interval for samples produced either with the primer 1380F or 1389F, respectivelly.

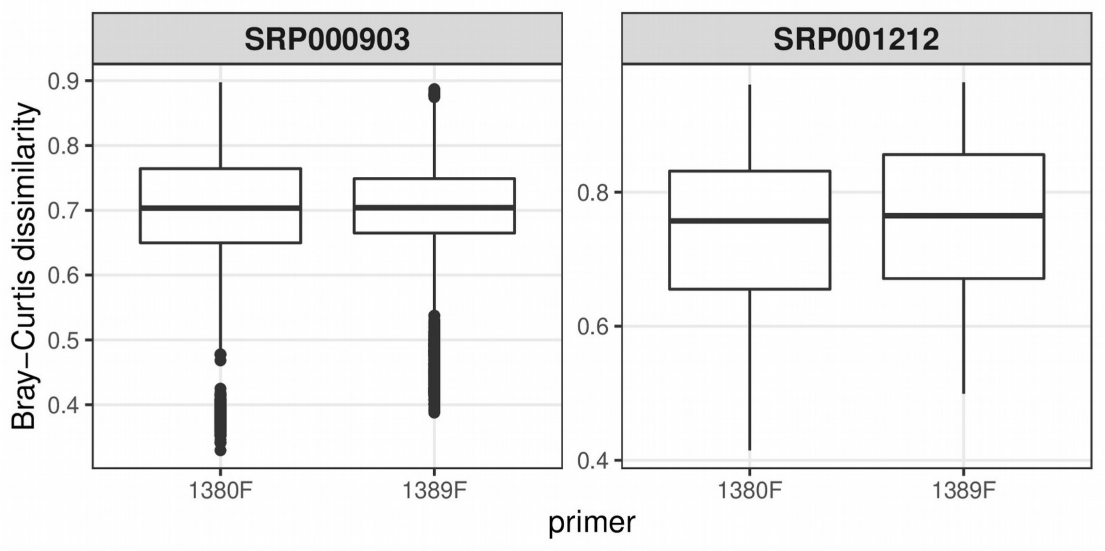

Figure SX5 Distribution of Bray-Curtis dissimilarity values between sample's libraries produced with the same primer pair.

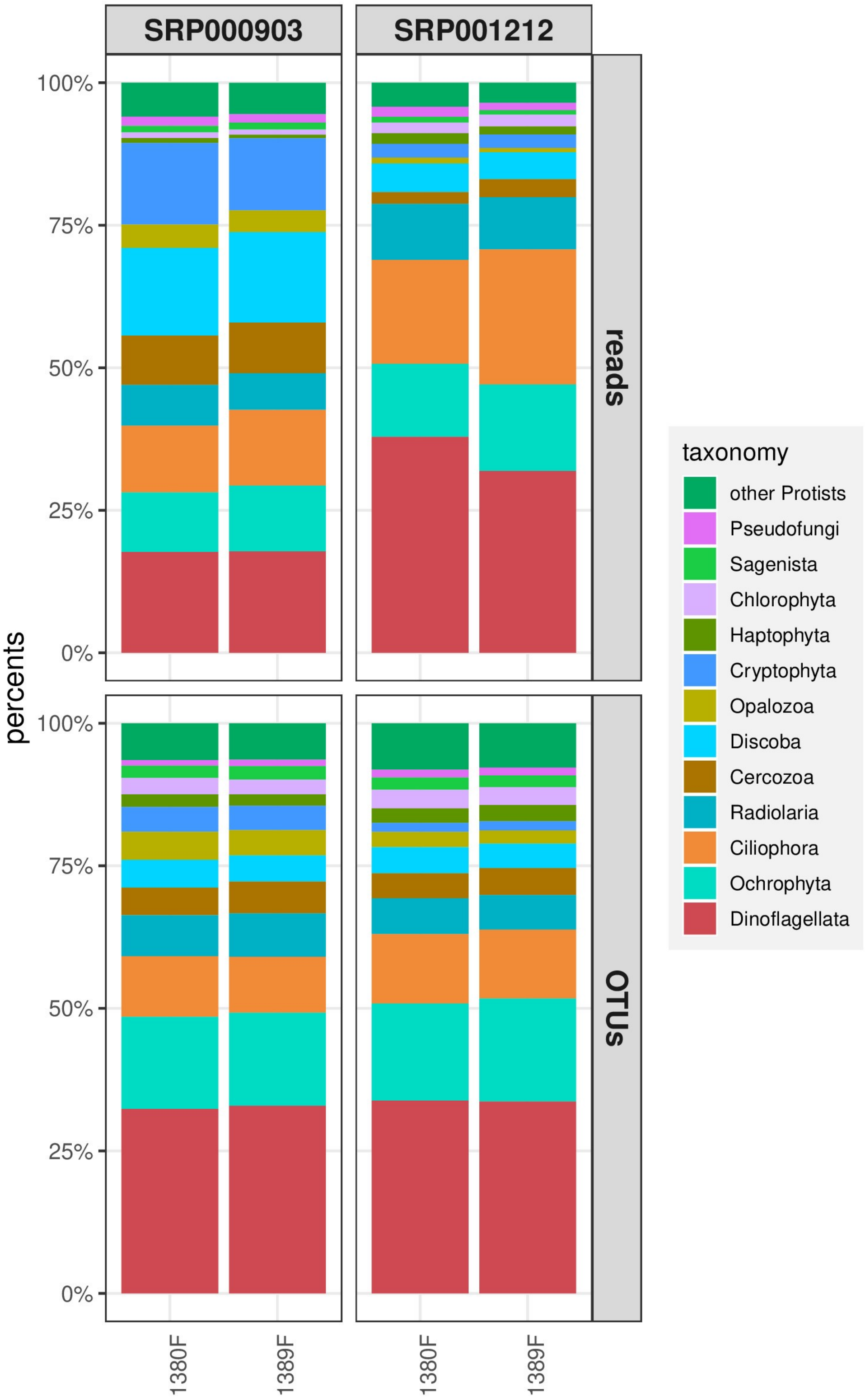

Figure SX6 Average percentages of reads and OTU richness of the main protist phyla over 100 bootstrapped rarefaction.