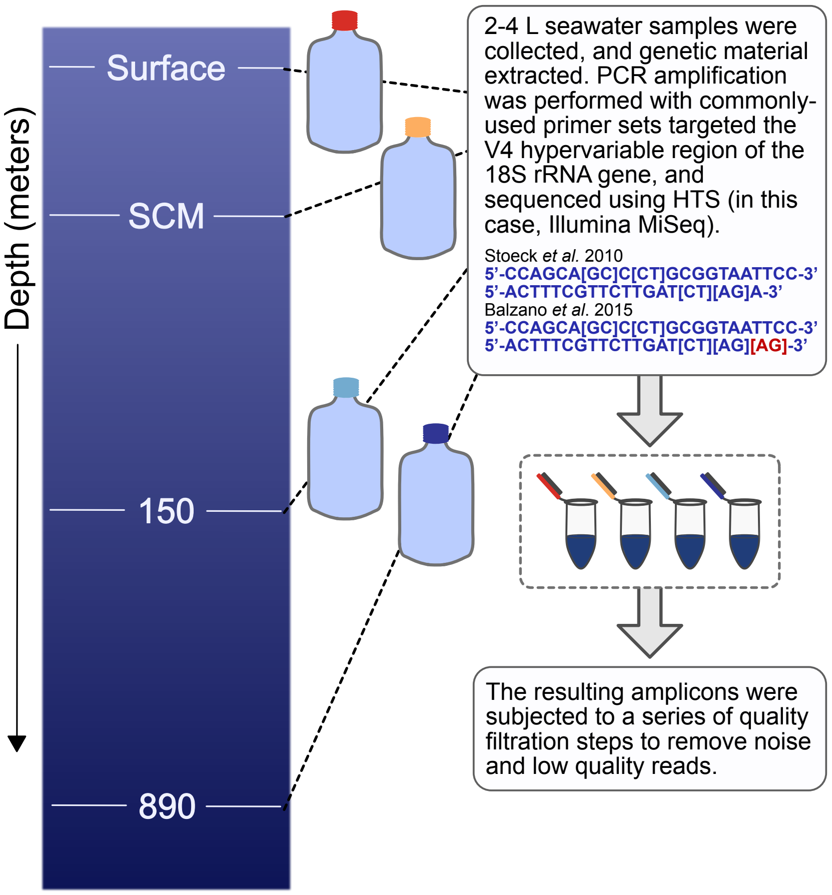
A. Sample collection & tag-sequencing



2-4 L seawater samples were collected, and genetic material extracted. PCR amplification was performed with commonly-used primer sets targeted the V4 hypervariable region of the 18S rRNA gene, and sequenced using HTS (in this case, Illumina MiSeq).

Stoeck *et al.* 2010
5'-CCAGCA[GC]C[CT]GCGGTAATTCC-3'
5'-ACTTTCGTTCTTGAT[CT][AG]A-3'
Balzano *et al.* 2015
5'-CCAGCA[GC]C[CT]GCGGTAATTCC-3'
5'-ACTTTCGTTCTTGAT[CT][AG][AG]-3'

The resulting amplicons were subjected to a series of quality filtration steps to remove noise and low quality reads.

B. Cluster sequences into approximate species-level designations

The number of species in a sample is estimated by clustering/collapsing sequences with high similarity into groups which serve to represent a species.

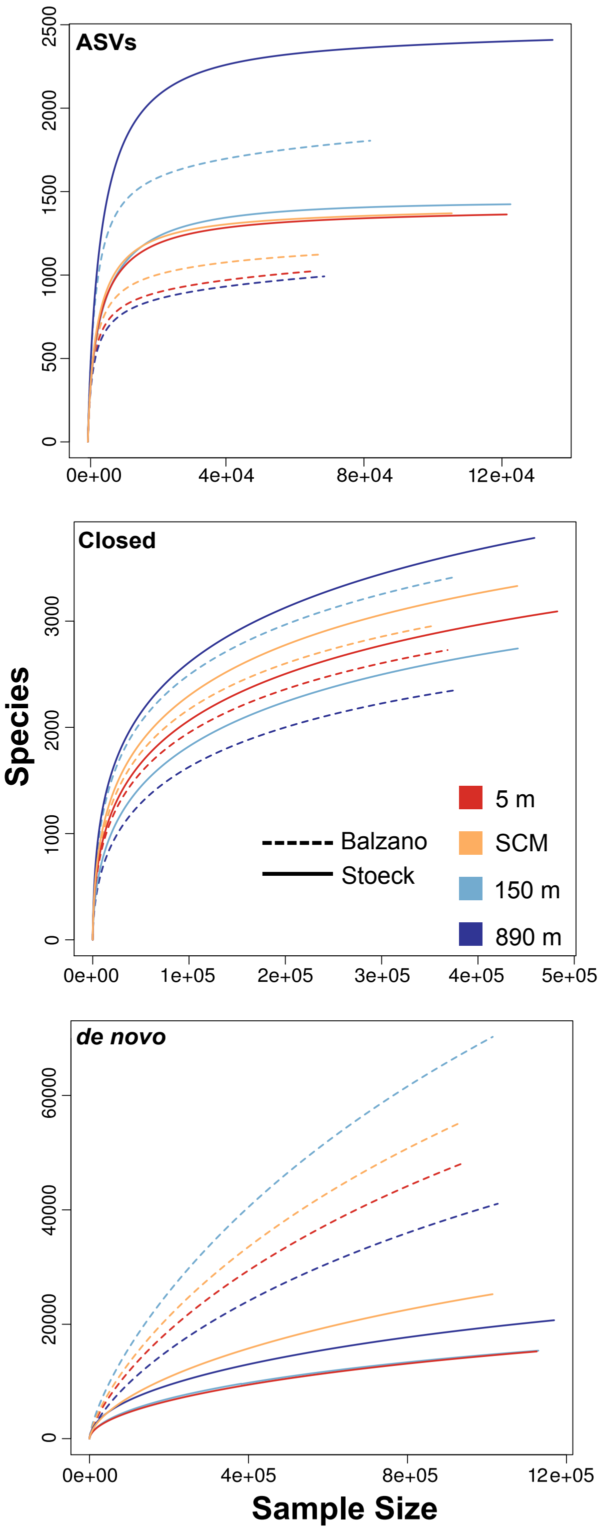
The most common approach involves clustering sequences into **Operational Taxonomic Units** (**OTUs**). It relies on setting a similarity threshold (i.e. 97%), whereby sequences are grouped together if they have no more than 3% base pair differences. OTU methods include: clustering all the sequences into OTUs blindly (***de novo***), employing a reference database of known sequences to guide sequence assignments to an OTU (**closed-reference**), and **open-reference OTU clustering**, which combines the *de novo* and closed-reference methods.

One recent approach for sequence clustering identifies **Amplicon Sequence Variants** (**ASVs**), which may vary by only a single base pair difference. Sequences are de-noised prior to ASV determination in an effort to more accurately identify the biological variability between sequences.

OTUs or ASVs	Clustering method
5,112	ASVs (QIIME 2)
5,363	Closed reference (QIIME 2)
32,867	*de novo* (QIIME 2)
2,048	Open-reference (QIIME 1)

C. Rarefaction curves

Results from OTU clustering, or more recently ASVs, often have been analyzed using rarefaction to estimate total protistan species richness in a sample, based on repeated sampling from the same dataset. The degree to which all diversity has been assessed in the sample is indicated by the asymptotes of the resulting curves.



Legend: Datasets from four environmental samples were generated and analyzed using several commonly applied approaches and parameters to yield estimates of protistan species richness. The different predictions indicate the inconsistency of results presently obtained from field surveys of protistan genetic diversity, even when starting with the same material. (**A**). Seawater was collected from 4 depths (surface, subsurface chlorophyll maximum (SCM), 150 m, and 890 m) from a time-series station situated off the coast of southern California. Extracted genetic material was PCR amplified (in this example using two primer sets (Stoeck et al. 2010 and Balzano et al. 2015) to isolate the V4 hypervariable region of the 18S rRNA gene. (**B**). The remaining quality sequences were grouped into Operational Taxonomic Units or Amplicon Sequence Variants that, optimally, approximate species-level distinctions. (**C**). Three sets of rarefaction curves demonstrate how applying two primer sets and three clustering protocols on the same samples can generate different results. The asymptotes of rarefaction curves have often been misinterpreted as indicators of the total richness in the environment. In actuality, the asymptotes provide an indication of the thoroughness of the sampling, and the effectiveness of the method employed to assess diversity. Different methods typically yield different results, as shown here by the different rarefaction curves.