Amplicon sequence variant-based meiofaunal community composition revealed by DADA2 tool is compatible with species composition

Zoya Harbuzov (zoya.garbu@ocean.org.il)
University of Haifa
Valeria Farberova
University of Haifa
Moshe Tom
National Institute of Oceanography, Israel Oceanographic & Limnological Research
Alberto Pallavicini
University of Trieste
David Stanković
National institute of Biology
Tamar Lotan
University of Haifa
Hadas Lubinevsky
National Institute of Oceanography, Israel Oceanographic & Limnological Research

Article

Keywords: Eastern Mediterranean, meiofauna, molecular taxonomy, 18S-SSU-rRNA, metabarcoding, Amplicon Sequence Variant (ASV), Illumina

Posted Date: March 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1453156/v1

Abstract

The present study is aimed at implementing the amplicon sequence variant (ASV) concept to describe meiofaunal species composition independent of morphological identification, while strongly indicating reasonable compatibility with the underlying species. A primer pair was constructed and demonstrated to PCR amplify an 18S barcode from a variety of meiofaunal taxa, compatible with the high throughput sequencing (HTS) Illumina 300 X 2 bps platform. Sixteen 18S multi-species HTS assemblies were created from meiofaunal samples and merged to one assembly of ~2,150,000 contigs. Five quality scores (q=35, 30, 25, 20, 15) were implemented to filter five 18S barcode assemblies, which served as inputs to the DADA2 software, ending with 5 reference ASV libraries. Each of them was clustered, applying 3% dissimilarity threshold and revealed an average number of 1.38±0.078 ASVs/cluster which demonstrate high level of ASV uniqueness. The libraries which were based on $q \leq 25$ reached a near-asymptote number of ASVs which together with the low average number of ASVs/3% cluster, strongly indicate compatibility with the actual number of underlying species. Hence, the q=25 library was selected to be used as metabarcoding reference library. It contained 461 ASVs and 342 - 3% clusters with average number of 1.34±1.036 ASV/cluster and their BLASTN annotation elucidated the expected meiofaunal taxa. The 16 assemblies of sample-specific paired contigs were mapped to this reference library and sample ASV profiles, namely the list of ASVs and their proportional copy numbers were created and also normalized to the actual number of individuals in each sample.

Introduction

The long-term goal of our research is better understanding of the biotic and abiotic ecological interrelationships in marine soft substrate habitats. Meiofauna is an important size fraction in this habitat, containing a variety of invertebrate groups detailed in Giere (2009). There is no universally accepted meiofaunal size range and the 20-500 μm range was implemented by us. The present study suggests an efficient, accurate and practical approach for characterizing meiofaunal semi-quantitative species composition to be applied in ecological studies.

The species (or even sub-species populations) is the desired taxon for habitat's ecological analysis, assuming that unique set of biotic and abiotic factors affects roughly uniformly the survival of individuals belonging to one species, in contrary to different sets which affect other species. Therefore, the list of species in a studied habitat and their relative abundances, termed here species profile, is required to determine ecological characteristics of meiofaunal communities.

Morphologically identifying meiofauna to species level to be used for ecological analyses is extremely difficult from several reasons thoroughly discussed in the literature, and the implementation of molecular identification (barcoding) was suggested to overcome this obstacle (Carugati et al., 2015) replacing the term species with operational taxonomy unit (OTU) or amplicon sequence variant (ASV) for individuals identified by molecular barcodes. ASV is used when the barcode is assumed to represent a genuine unique biological sequence free of sequencing errors and assigned to one species (Callahan et al.,

2017). Partial sequences of the mitochondrial cytochrome oxidase subunit I (COI) (Hebert and Gregory, 2005) and ribosomal RNA genes (e.g., Aylagas et al., 2014; Hadziavdic et al. 2014) are widely used faunal barcodes.

The construction of an OTU/ASV profile from a multi-species environmental sample is termed metabarcoding, its scheme is presented in Fig. 1, and ideally, its resulting OTU/ASV profiles are aimed at including all the species in the sample in quantitative terms. This is a challenging task and its accomplishment could be biased at each stage of the metabarcoding process as recently reviewed (Alberdi et al., 2018; Pawlowski et al., 2018; van der Loos and Nijland, 2020). The major faults were: absence of some of the species from the sequence assembly due to too many primer mismatches and disproportionality of OTU copy numbers to actual individuals' counts. The postulated reasons are: variable number of copies per individual from different species and different PCR efficiencies caused by variable annealing stringency between the primers and the various species-specific barcode sequences.

Mere evaluation of community diversity and habitat's environmental status indices (Ayalagas et al., 2014, 2016; Lejzerowicz et al., 2015; Cordier et al., 2017; Fonseca et al., 2017; FrØslev et al., 2017) could be accomplished using only ASVs. Their annotation to morphological taxa, not necessarily at species level, could allow assignment of additional characteristics to each ASV, morphology-based feeding guilds and reproduction strategies, habitats of presence during developmental history, resistance to anoxia and many more abiotic ques, which would enable detailed ecological analysis of a studied marine sedimentary habitat and its biotic-abiotic interrelationships.

DNA for metabarcoding in sedimentary habitat is mainly extracted from two sources: 1) multi-species individuals sorted from a sediment sample. Sorting from large amounts of sediment allows statistically sufficient species representation even from an environment with poor faunal density, enables the counting of the individuals and even dividing them into taxa at achievable taxonomic levels (Fonseca et al., 2017; Lobo et al., 2017). 2) extraction from unsorted sediment, assumed to contain all the extra and intra-cellular environmental DNA, an approach which is implemented mainly for abundant organisms, because it is limited by the feasible amount of processed sediment (Guardiola et al., 2015; Lejzerowicz et al., 2015; Cordier et al., 2017).

The well-established COI barcode region has proved to be relatively difficult to amplify in free-living marine nematodes, the major meiofaunal group, because of "rampant gene rearrangement, hypervariation among haplotypes, and frequent recombination in mitochondrial genome" quoting Boufahja et al. (2015) and literature therein and see also Carugati et al. (2015). Therefore, the small sub-unit of ribosomal DNA gene (SSU-rDNA or 18S) was widely applied (Boufahja et al., 2015; Dell'anno et al., 2015; Carugati et al., 2015; Lee et al., 2017; Macheriotou et al., 2018). A series of primer pairs were proposed from a variety of regions along the 18S molecule for amplifying barcodes, and a wide range of barcode lengths were resulted, from 110 bps up to most of the 1700 bps of the entire molecule (Aylagas et al., 2014; Hadziavdic et al., 2014; Boufahja et al., 2015; Guardiola et al., 2015; Cordier et al., 2017; Fonseca et al., 2017). Longer barcodes from hypervariable parts of the 18S

molecule are assumed to better distinguish species, limited by the need to be compatible with the implemented HTS method.

A variety of approaches were suggested and used for the construction of a comprehensive barcode reference library which would encompass all inhabiting species in a studied soft substrate habitat with the main obstacle of partial knowledge of all the inhabiting species and their species-specific barcodes. Reference libraries were created using comprehensive publicly available sequences (Dell'anno et al., 2015; Lejzerowicz et al., 2015; Klunder et al., 2019; Leite et al., 2020), local barcodes (Aylagas et al., 2014; Lobo et al., 2017) or fully species-independent OTU /ASV libraries. The later were initially constructed using clustering methods coupled to implementation of semi-arbitrary percentage dissimilarity threshold which does not distinguish ASVs but only OTUs (Guardiola et al., 2016; Fonseca et al., 2017). Others used the ASV-related approach by introducing denoising methods which specifically reduce sequencing errors (Gaspar and Thomas, 2015; FrØslev et al., 2017; Callahan et al., 2016, 2017). Metabarcoding utility was tested in meiofaunal communities by several studies (Guardiola et al, 2016; Lobo et al., 2017; Macheriotou et al., 2018; Schenk et al., 2020).

Based on the strongly indicated assumption that there is no biotic or even faunal universal primer pair which could PCR amplify species, indicated by the plethora of taxa-specific primers (e.g., Haye et al 2004; Mikkelsen et al 2006; Miya et al., 2015; Layton et al., 2016; Schroeder et al., 2021 and many others), the demonstrated metabarcoding approach here is meiofauna-specific and aimed only at improving the characterization of meiofaunal communities. The construction of the reference library is based on ASVs derived only from species of the studied area using the DADA2 denoising protocol (Callahan et al., 2016), initially, without assignment of taxa names to ASVs. This ASV source would avoid false positives which may emerge by using wider library from public sources.

Therefore, the study is divided into three parts: 1) designing a universal primer pair followed by bio-informatic testing of its suitability for both wide range of meiofaunal taxa and the required length for the utilized Illumina sequencing platform, 2) constructing the reference library, and indicating that the number of ASVs is roughly compatible with the actual number of species, 3) metabarcoding of HTS assemblies of 16 multi-species sample DNAs, demonstrating the benefit of using absolute individual counts to normalize the proportionate copy numbers mapped to various ASVs.

Results

3.1 Design of universal primer pair

The selected universal primer pair was: Forward - 5'-GAGGTAGTGACGAAAAATAAC-3'; Reverse - 5'-CGTTCTTGATTAATGAAAACATTC-3', and the resulted PCR amplicon is located roughly between base pairs 400-900, regions V3-V4 of the ~1700 bp 18S molecule. The compatibility of these primers with 18S sequences from a variety of meiofaunal groups, those which were present in our samples, was tested and the results are presented in Table 1, demonstrating an average level of primer mismatch <1. It has to be noted that part of the mined sequences missed the area of one of the primers. The results strongly

indicate that the selected primer pair is suitable to serve as universal PCR primer pair of meiofauna. This suitability was farther confirmed by examining the annotations of the obtained reference library sequences (see below). This general primer compatibility testing could be done for any other primer and target taxa.

Table 1 – Testing the similarity between the universal primers and 18S sequences of a variety of meiofaunal groups

Taxon	# examined sequences	# of sequences aligned by the forward primer	Average mismatch of the forward primer [bps]	# of sequences aligned by the forward primer	Average mismatch of the reverse primer [bps]
Polychaeta	490	472	0.44 ± 0.8	433	0.52 ± 0.8
Nematoda	203	203	0.09 ± 0.3	203	0.08 ± 0.4
Harpacticoida	69	69	0.22 ± 0.4	69	0.03 ± 0.02
Ostracoda	37	37	0.08 ± 0.4	16	0.9 ± 0.6
Isopoda	17	17	0.12 ± 0.5	17	0.12 ± 0.5
Cumacea	9	8	0.75 ± 0.9	8	0.75 ± 1.4

3.2 Selection of appropriate reference library

Five quality score-dependent DADA2 reference libraries were constructed. An input of 2,143,575 reads was used for the construction of each library, gradually decreasing linearily with increasing **q** value (Fig. 2). The number of created ASVs for each quality score-dependent library is presented in Fig 3. Unlike the number of input HTSs, both the number of ASVs and stronger, the number of 3% clustered ASVs, nearly approached an asymptote. The average number of ASV/3% clustered reference library in relation to the quality score-dependent input is presented in Fig. 4, averaging 1.38±0.078 ASVs/cluster across libraries, with no statistical difference between the various values (Kruskal-Wallis rank sum test; P=0.32). The near-asymptote shape of the ASV and clustered ASV curves in Fig. 3 and the low average number of sequences per 3% clustered ASV (Fig. 4) led to the conclusion that a total of 340-380 species-specific barcodes, the range of the asymptotic part of the clustered ASV curve in Fig 3, were PCR amplified from our meiofaunal samples. The distribution pattern of the ASVs per cluster elucidated an overwhelming majority of clusters represented by one ASV and few with higher numbers (presented only for q=25, Fig. 5), indicated also by the relatively high standard deviations (Fig. 4). The small average number of ASVs per cluster indicates also that within the limits of the tested quality scores, the DADA2 protocol efficiently but not perfectly distinguish unique ASVs. Consequently, reference library which emerged from the contig assembly created by the q=25 CUTADAPT was preferred to be used by the metabarcoding process due to the asymptotic number of ASVs on one hand and the low probability of erroneous bases of q=25 (0.33%) on the other hand.

OTU reference libraries were also constructed using the five CUTADAPT filtered barcode assemblies and using 3% dissimilarity threshold. The number of OTUs resulted from each assembly is presented in Fig. 6 elucidating lack of OTU asymptote with increasing number of OTUs with decreasing q values.

The results of BLASTNing the selected reference library to the GenBank nucleotide collection (nt/nr) is presented in Table 2.

Table 2 – Annotation of the sequences of the reference library. # of families are in parentheses.

Taxonomic group	# of ASVs
Nematoda (27)	197
Annelida (17)	56
Copepoda (13)	35
Ostracoda (4)	11
Others	43

Annotation using public databases are not an exact identification, providing only a general view of the existing biota, because on global public database scale, the species specificity or more generally the variability of the 18S barcode used here is not known, and in addition, not all taxonomic groups are equally represented in the public database. However, the general view is compatible with the taxa present in marine meiofauna and it is also compatible with the individuals that were sampled and sorted by us in autumn of 2018 in the studied site (Harbuzov, Per. Info.). Nematoda from a variety of families dominate the annotations, followed by annelida, almost all of them polychaeta, copepoda, mainly benthic families and a variety of other meiobenthic groups: ostracoda, nemertea, rotifera, bryozoa, mollusca, platyhelminthes, hemichordata, echinodermata, cnidaria, phoronida, gnathostomulida, xenacoelomorpha, kinorhyncha and Brachiopoda.

Table 3 – sample data including number of contigs that were filtered using q=25, their percentage mapping to the reference library with mismatch threshold of 3% and the # of 3% clustered ASVs in each sample

[1]Sample name	Total # of paired contigs (q=25)	% aligned to the reference library	# of aligned ASVs
HS122A-1	44,922	83.6	59
HS122A-2	59,206	75.6	62
HS122C-1	52,398	92.3	43
HS122C-2	50,580	80.2	52
HS122C-3	71,261	86.6	37
HS394-1	32,794	90.8	116
HS394-2	25,961	88.9	161
HS394-3	11,720	83.6	156
HS394-4	28,962	83.9	149
TA46-1	85,727	91.2	143
TA46-2	118,352	83.8	118
TA46-3	18,691	91.1	126
TA46-4	30,456	86.7	73
TA46-5	272,903	93.7	105
TA46-6	15,914	90.2	152
TA76-1	41,122	79.8	86
Average	60,061±63,229	86.4±5.2	

[1] The designations of the sample names were composed of the perpendicular to the coast transects, HS – Haifa transect and TA – Tel Aviv transect, the bottom depth in meters and the serial number of replicate PCRs done with each sample DNA as template.

Metabarcoding of samples

Sixteen samples were used as technical samples for demonstrating our metabarcoding approach. However, they were sampled from five natural sites and their replicate assemblies were created by several PCR runs using the five DNA templates. Contigs from each sample that were filtered by CUTADAPT with q=25 were paired and mapped to the reference library. An average of 86.4±5.2% of the paired and filtered contigs from each sample were mapped (Table 2) to 337 out of 342 clustered ASVs of the reference library and the number of mapped contigs was quite variable among samples (Table 3). The clustering of the various sample profiles using both quantitative values, the copy number and the normalized counts is presented in Fig. 7. The improvement of the clustering according to sampling sites after normalization to real densities is demonstrated by the better clustering according to sampling sites with higher similarity percentages.

Discussion

The study presents a molecular taxonomy-related approach for the construction of semi-quantitative species profiles of meiofaunal communities and for their comparisons, one of the analyses required for better understanding the ecology of many natural habitats. The presented procedure relies only on the sampled individuals in the studied habitat, it is morphological identification-free but strongly indicated to fairly represent the actual inhabiting species. The approach responds to the difficulty to practically use meiofaunal morphological taxonomy in the context of an ecological analysis of species composition.

It was assumed by us and elsewhere that the concept of universal primer pair, being able to amplify a specific barcode in all eukaryotic species is not achievable. Hence, limited innovative aspect of this study is the design of a primer pair demonstrated to be compatible with both major meiofaunal taxa and the Illumina 300 x 2 bps sequencing platform.

The novel DADA2 software was used to construct an ASV reference library built from sequences of the amplicons elucidated in the 16 sample assemblies used for metabarcoding, initially not using morphological identification. The resulting library revealed annotation to 18S of meiofaunl taxa and was indicated to be fairly compatible with the real number of species inhabiting the studied habitat. DADA2 developers (Callahan et al., 2016, 2017) claimed that their analysis is able to distinguish real biological sequences (ASVs) from an OTU assembly by using a machine learning process of the Illumina error rate for each processed assembly. However, they mentioned that distinction of particular ASV depends on a minimum of OTU copies emerging from this ASV and assuming that the majority of these OTUs have the real biological sequence with no sequencing error. Indeed, the numbers of ASVs increased with increasing number of amplicons (Fig. 2), because assumedly more ASVs reached the minimum OTUs number required for their distinction. However, at a certain stage, a near-asymptote ASV number was formed (Fig. 3) assumedly representing the barcodes of most PCR-amplified species. Clustering of the reference library at 3% dissimilarity threshold demonstrated mostly one ASV/cluster (Figs 4 and 5), further indicating the validity of one-ASV-one-species concept. The number of amplicons participating the DADA2 analysis increased by gradually reducing the sequencing quality of the participating amplicons, enabled the selection of the asymptote-approaching reference library, which concurrently used a reasonable sequencing quality score. Another way to gradually increasing the number of DADA2 input contigs, which was not applied by us yet, is the construction of a very broad OTU assembly using reasonable quality score and gradually increasing the number of DADA2 participating contigs until reaching an asymptote. Concerning the DADA2 approach, the still slow increase of ASVs even after reaching near-asymptote status (Fig. 3) is assumedly a result of the elucidation of very rare species and a full asymptote may require more participating contigs than the 2.15 million used here.

The 3% dissimilarity threshold among barcodes is widely accepted as a species-distinguishing dissimilarity, although not always accurate. It was used here to generally respond to natural variability of the barcode sequence within certain species and also to reduce potential over-splitting of species to ASVs performed by DADA2, which may statistically occur in few instances. Hence, the 3% threshold was used

at the stage of the reference libraries examination to demonstrate the efficiency of the DADA2 procedure in fulfilling the one ASV-one-species concept, and during the map-to-reference process it was used to partially adjust within-species natural variability.

Modifying of the reference library due to spatially and temporally broadening of the sampling effort may be accomplishing by repeating the present process from the beginning, but also by repeating the process only for new samples and mapping newly resulted ASVs to the previous library.

Mere clustering of the five assemblies with the gradually increasing amplicon numbers and applying a 3% dissimilarity threshold, revealed no asymptote (Fig. 6), assumed to be a result of the gradually increasing number of sequencing errors with no relation to the real number of species.

More definite but much slower demonstration of the compatibility between ASVs and species could be further proved in the future by gradual morphological identification of single individuals followed by their barcode amplification, Sanger sequencing and comparison to the reference library. Studying the morphological-barcoding relationships would better demonstrate the within- and between-species variability of our 18S barcode. Although apparently, 342 meiofaunal 3% clustered ASVs indicate quite variable 18S region. The annotation of the reference library to public databases using family level, provides a way to assign additional biological features to the ASVs such as feeding guilds based on established classifications like those of polychaeta (Jumars et al., 2015) or of nematoda (Jensen, 1987) and similarly to other biological traits.

The proportional number of counted individuals of the various inhabiting species does not fit one-to-one with the proportional number of amplicon copies because of slightly different efficiency of the universal primers for each of the species during PCR amplification and also from the different number of 18S copies per individual in each of the species present in the sample (van der Loos and Nijland, 2020). Therefore, ecological use of metabarcoding introduces an inherent error. An additional error may occur when comparing ASV profiles of different samples because of the relativity of ASV copy numbers, not related to real individual counts. The sorting and counting of individuals before DNA extraction may be beneficial for reducing this source of error as shown here (Fig. 7). The normalization made here, which was based on the total number of individuals/sample was quite simple. However, identification of the sample to higher-than-species taxa and higher-than-species annotation of the reference library, would enable to separate normalization of each higher taxonomy group (e.g. nematoda, polychaeta, etc.), increasing the accuracy, and leaving the bias only to within-group level.

It is our opinion that the ease and actually the possibility to construct sample profile at roughly species-like level using metabarcoding with its inherent error is superior to the practically non-realistic rely on morphological identification. In addition, the repeatability of the generic barcode-dependent sample profile creation, which rely on stable ASV sequences, bypass both poor taxonomy knowledge of certain groups and disputes among taxonomists causing changes of taxonomic status of species with time, and complicating broad temporal comparisons.

Both the construction of the reference library and the metabarcoding process relies on the same set of HTSs. However, 5 ASVs of the reference library out of 342 were aligned to no sample contig. The explanation to this phenomenon may be looked for in the exact ASV selection process of DADA2 with which we are not familiar and is out of the scope of this study.

The sample HTS used for pairing was filtered by q=25, similar to the quality score used by the DADA2 analysis. However, this is not obligatory. The pairing of single-end HTSs is done in the region of relatively poor base calls, the 3’ end of the sequence. However, the pairing improves the base identification at that poor region, and lower q could be used, increasing the number of applied contigs during the mapping process and fortifying the statistical significance of the metabarcoding.

A fast and efficient morphological identification-free approach for community ecological analysis of marine meiofauna which still keeps the relationship of the ASVs with the actual number of species is presented here. A novel primer pair was constructed, suitable for meiofaunal communities and a normalization method by real individual counts was suggested. Morphological identifications and their accompanied barcodes could be gradually added later from public databases or from in-house Sanger sequencing, not delaying essential ecological analysis.

Materials And Methods

Sampling, sorting and counting

Meiofauna was sub-sampled from a 0.25m^2 box corer (BX-650, Ocean instruments, San Diego, CA) using a 9.4 cm diameter plexiglass core pushed down to the 17 cm horizon of the sediment. The core was horizontally sliced to allow convenient preservation on board by mild suspension in 99% ethanol using a wing mixer to allow efficient ethanol penetration into the sticky mud. In the laboratory, the samples were sonicated for 30 seconds to reduce sediment aggregation and sieved through 500 and 20 µm sieves, the 20 µm retained individuals were sorted from the remaining sediment by gradient density centrifugation through colloidal silica suspension (Ludox HS-30, Sigma-Aldrich Cat. No. – 420824, density 1.18 g*cm^{-3}) according to Heip et al. (1985) and Danovaro (2010) and each entire sample was used for further processing and analysis. Five sampling sites within the depth range of 45-360 m which were sampled in October 2017 along the coast of Israel were selected to establish and demonstrate the present approach for constructing sample ASV profiles.

DNA extraction

Homogenization of samples was performed by the FastPrep® homogenizer and lysing matrix A beads (MP Biomedicals). DNA from whole samples was extracted using the E.Z.N.A.®Mollusc DNA Kit (Omega bio-tek, Cat-D3373). DNA levels were measured by fluorometric evaluation (QFX fluorometer, Denovix).

Design and testing of universal primers

The required barcode has to be compatible with the 300 X 2 bps Illumina Sequencing platform (Illumina, San Diego, CA; ~10^5 HTSs/sample), permitting the formation of paired-end sequences of 18S barcode and the designed PCR primer pair has to be universal for the target meiofaunal groups, enabling the formation of amplicons from all the relevant species-specific DNAs. Hence, sequences of 18S of free-living nematoda, harparticoid copepoda, polychaeta, isopoda, ostracoda, and cumacea were mined from the public sequence databases (supplementary table 1) and were aligned by the Geneious prime software (Biomatters LTD.). Iterative visual inspection of relatively uniform sequence regions led to the design of potential primer pairs compatible with the sequencing platform requirements. The average mismatch level of each of the selected primers with the mined sequences was calculated using Geneious prime through its menus: “Primers – test with saved primers” and its exported summarizing table. The primer pair with the lowest mismatch level was selected, producing a ~450-510 bp barcode, suitable to be formed by a 300X2 bp Illumina platform. The Illumina platform was selected as it is compatible with the DADA2 protocol used downstream.

Construction of an 18S ASV library

The PCR amplification conditions of the target 18S barcode were tested using a range of annealing temperatures and number of amplification cycles. The primers used for the amplifications contained the CS1 and CS2 Illumina tails which are required for HTS. The applied PCR conditions were: 95°C – 2 min; 30-40 cycles (94°C – 30 sec, 46-59°C – 30 sec, 72°C – 30 sec); 72°C – 3 min. The results of this and later optimization efforts are not the topic of this study. However, the 16 amplicon assemblies resulted from the optimization process were used here as technical samples for demonstrating the suggested metabarcoding process. The amplicons for HTS were cleaned up from 1% agarose gels (NucleoSpin® Gel and PCR Clean-up kit, Macherey-Nagel, Germany) and were HTSed on Illumina 300 x 2 bps platform by a service laboratory.

The 32 resulting FASTQ sequence files (16 forward and 16 reverse) were submitted to GenBank as a Sequence Read Archive (SRA) BioProject PRJNA791542 and merged into two files, forward and reverse on a LINUX platform, using the CAT command. The merged files were filtered using the CUTADAPT software (Martin, 2011) and the filtration included truncation of the poor 3’ side of each sequence using five alternate quality score values, q=15, 20, 25, 30 and 35, primer removal, and eliminating both short resulted sequences (<250 bps) and sequences with more than 3 Ns in a row. The five resulting pairs of forward and reverse FASTAQ files were used as inputs for the DADA2 analytical process (Callahan et al., 2016), applied through the Qiime2 software plugin (Bolyen et al., 2019). DADA2 created a list of apparently unique ASVs from each of the input FASTQ files and each of the q-related ASV assemblies was considered a potential metabarcoding reference library. Too short sequences were manually removed from each of these reference libraries, realized to be erroneously aligned during pairing. Annotation of the remaining contigs by BLASTN against the nucleotide Genbank standalone database to identify and eliminate non-18S sequences and contaminating mammalian ones, resulted with ~450-510 bp assemblies. P-distance resemblance table of each of the five reference libraries was prepared by the MEGA-X software (Kumar et al., 2018) and served as input for clustering process of the sequences using

the PRIMER-v7 software (Clarke et al. 2014; Clarke and Gorley., 2015) through its group average clustering protocol. ASV Clusters were determined using the widely accepted < 3% dissimilarity threshold among cluster members and for each assembly, the number of ASV clusters and the average number of ASVs per cluster were calculated.

OTU assemblies were also constructed from the five, quality score-related paired barcode assemblies, using the VSEARCH software (Rognes et al., 2016) applying dissimilarity threshold of 3%. The following VSEARCH commands were used through the Qiime2 platform: join-pairs - dereplicate-sequences - cluster-features-de-novo. These reference libraries were constructed for comparison with the DADA2 ones to demonstrate the differences between the two strategies for preparing reference library.

Construction of ASV sample profiles

HTSs from each of the 16 samples, resulted from the CUTADAPT assembly of q=25, were paired by the VSEARCH software, using the VSEARCH Join-pairs plugin of Qiime2.
Metabarcoding was performed by Geneious Prime through its menus: “Elign/assemble - map to reference”, using 3% allowed contig dissimilarity and 5% gaps of a maximum of 3 bp per gap, resulted with ASV profiles of each sample.

Similarity among sample profiles using copy numbers was examined by their clustering, using the PRIMER-v7 software. The applied clustering parameters were: square root conversion of the ASV copy numbers, the Bray-Curtis similarity index and the group average clustering protocol. However, the number of contig copies mapped to each ASV is proportional, not providing real density which affects the comparison among sample profiles. Therefore, a normalization of the proportionate copy numbers in each sample by the real numbers of individuals in the sample was accomplished, accompanied with repeated clustering.

Declarations

Acknowledgments

The crew of R/V ”Bat Galim” are deeply thanked for the mud sampling. Dr. Rafi Gassel from Syntezza bioscience Ltd. is appreciated for the HTS services. The study was partially supported by grants 216-17-011 and 220-17-004 from the Israeli ministry of energy to HL and MT.

Competing interests

Each of the contributing authors declare to have no competing interests in relation to this article.

Data availability statement

The original HTS sequences used in this article were deposited as GenBank SRA BioProject PRJNA791542. The accession numbers of the mined 18S sequences used for primer examination are

provided as Supplemental table 1 of this article.

Author Contributions

ZH is the leading researcher and the study is part of her PhD thesis, ZH and VF carried out the experimental work, data analysis was performed by ZH, VF and MT and was assisted by AP and DS. The writing was done by ZH and was reviewed and corrected by AP and DS, and by HL , MT and TL, ZH PhD supervisors.

References

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Meth. Ecol. Evol.* **9**, 134–147 (2018).

Aylagas, E., Borja, Á., & Rodriguez-Ezpeleta N. Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). *PLoS ONE* ***9***, e90529; doi.org/10.1371/journal.pone.0090529 (2014).

Aylagas, E., Borja, Á., Irigoien, X., & Rodriguez-Ezpeleta N. Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Front. Mar. Sci.* ***3***, Article 96; doi.org/10.3389/fmars.2016.00096 (2016).

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., et al. & Caporaso, J. G. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37 852–857 (2019).

Boufahja, F., Semprucci, F., Beyrem, H., & Bhadury, P. Marine nematode taxonomy in Africa: promising prospects against scarcity of information. *J. Nematol.* **47**, 198–206 (2015).

Callahan, B. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Meth.* **13**, 581-583 (2016).

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

Carugati, L., Corinaldesi, C., Dell'Anno, A., & Danovaro, R. Metagenetic tools for the census of marine meiofaunal biodiversity: An overview. *Mar. Genom.* **24**, 11–20 (2015).

Clarke, K. R., Gorley, R. N., Somerfield, P. J., & Warwick, R. M. Change in marine communities: an approach to statistical analysis and interpretation. Primer-E Ltd. (2014).

Clarke, K. R., & Gorley, R. N. Getting started with PRIMER v7. PRIMER-E: Plymouth, Plymouth Marine Laboratory, **20**, (2015).

Cordier, T. et al. Predicting the ecological quality status of marine environments from e-DNA metabarcoding data using supervised machine learning. *Environ. Sci. Technol.* **51**, 9118–9126 (2017).

Danovaro R. Methods for the Study of Deep-Sea Sediments, Their functioning and Biodiversity (ed Danovaro R.). (CRC Press, Taylor and Francis Group, 2010).

Dell'Anno, A., Carugati, L., Corinaldesi, C., Riccioni, G., & Danovaro. R. Unveiling the Biodiversity of Deep-Sea Nematodes through Metabarcoding: Are We Ready to Bypass the Classical Taxonomy? *PLoS ONE 1-18*; DOI:10.1371/journal.pone.0144928 (2015).

Fonseca, V. G. et al. Revealing higher than expected meiofaunal diversity in Antarctic sediments: a metabarcoding approach. *Sci. Rep.* ***7***: 6094; DOI:10.1038/s41598-017-06687-x (2017).

Frøslev, T. G. et al. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Comm.* ***8***: 1188; DOI: 10.1038/s41467-017-01312-x (2017).

Gaspar J. M., & Thomas, W. K. FlowClus: efficiently filtering and denoising pyrosequenced amplicons. *BMC Bioinform.* ***16***: 105; DOI 10.1186/s12859-015-0532-1 (2015).

Giere, O. *Meiobenthology - The Microscopic Motile Fauna of Aquatic Sediments.* (Springer-Verlag, 2nd revised and extended edition, 2009).

Guardiola, M. et al. Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ* ***4***:e2807; DOI 10.7717/peerj.2807 (2016).

Guardiola, M. et al. Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from Sediments of Marine Canyons. *PLoS ONE* ***10(10)***: e0139633; DOI:10.1371/journal.pone.0139633 (2015).

Hadziavdic, K. et al. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS ONE* ***9(2)***: e87624; doi:10.1371/journal.pone.0087624 (2014).

Hebert, P. D. N., & Gregory T.R. The promise of DNA barcoding for taxonomy. *System. Biol.* **54**, 852–859. (2005).

Heip, C., Vincx M., & Vranken, G. The ecology of marine nematodes. Oceanogr. *Mar. Biol. Ann. Rev.* **23**, 399-489 (1985).

Haye, P.A., Kornfield, I., & Watling, L. Molecular insights into Cumacean family relationships (Crustacea, Cumacea). *Molec. Phylogenet. Evol.* **30**, 798–809 (2004).

Jensen, P. Feeding ecology of free-living aquatic nematodes. *Mar. Ecol. Prog. Ser.* **35**: 187-196. (1987).

Jumars P. A., Dorgan, K. M., & Lindsay, S. M. Diet of Worms Emended: An Update of Polychaete Feeding Guilds. *Ann. Rev. Mar. Sci.* ***7***, 497–520 (2015).

Klunder, L. et al. Diversity of Wadden Sea macrofauna and meiofauna communities highest in DNA from extractions preceded by cell lysis. *J. Sea Res.* **152**, [101764]; 10.1016/j.seares.2019.101764 (2019).

Kumar S., Stecher G., Li M., Knyaz C., & Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molec. Biol. Evol.* **35**, 1547-1549 (2018).

Layton, K. K. S., Corstorphine, E. A., & Hebert, P. D. N. Exploring Canadian Echinoderm Diversity through DNA Barcodes. *PLoS ONE **11**(11)*: e0166118; DOI:10.1371/journal.pone.0166118 (2016).

Lee, M. R. et al. The identification of sympatric cryptic free- living nematode species in the Antarctic intertidal. PLoS ONE ***12***(10): e0186140; doi.org/10.1371/ journal.pone.0186140 (2017).

Lejzerowicz, F. et al. High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Sci. Rep.* ***5***:13932; DOI: 10.1038/srep13932 (2015).

Lobo, J., Shokralla, S., Costa, M. H., Hajibabaei, M., & Costa, F. O. DNA metabarcoding for high- throughput monitoring of estuarine macrobenthic communities. *Sci. Rep.* ***7***: 15618; DOI:10.1038/s41598-017-15823-6 (2017).

Macheriotou, L. et al. Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecol. Evol.* **9**, 1211–1226 (2019).

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**(1), 10-12 (2011).

Mikkelsen, P. M., Bieler R., Kappner, I., & Rawlings, T. A. Phylogeny of Veneroidea (Mollusca: Bivalvia) based on morphology and molecules. *Zool. J. Linnean Soc.* **148**, 439-521 (2006).

Miya, M. et al. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *R. Soc. open sci.* ***2***: 150088; dx.doi.org/10.1098/rsos.150088 (2015).

Pawlowski, J. et al. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* **637–638**, 1295–1310 (2018).

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* ***4***:e2584; DOI10.7717/peerj.2584 (2016).

Schenk, J. et al. Nematodes as bioindicators of polluted sediments using metabarcoding and microscopic taxonomy. *Environ. Int.* ***143***: 105922; doi.org/10.1016/j.envint.2020.105922 (2020).

Schroeder, A., Pallavicini, A., Edomi P., Pansera, M., & Camatti E. Suitability of a dual COI marker for marine zooplankton DNA metabarcoding. *Mar. Environ. Res.* **170**: 105444; doi.org/10.1016/j.marenvres.2021.105444 (2021).

Steyaert, M. et al. Advances in metabarcoding techniques bring us closer to reliable monitoring of the marine benthos. *J. Appl. Ecol.* **57**(11), 2234–2245 (2020).

van der Loos, L. M. & Nijland R. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molec. Ecol.* **30**(13), 3270-3288 (2020).

Figures

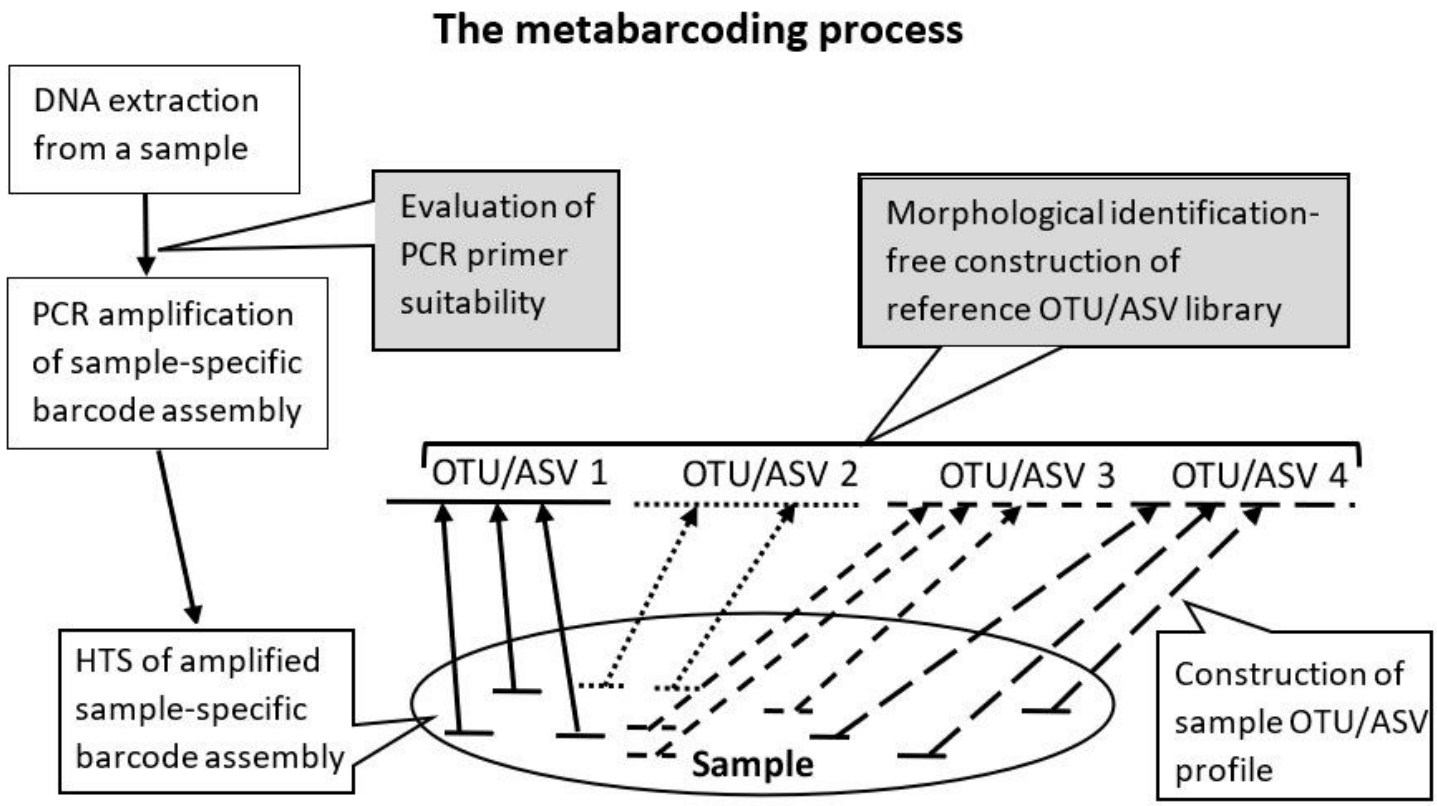

Figure 1

The metabarcoding process. OTU/ASV 1-4 - reference barcode library, **Sample** – high throughput sequences (HTS) assembly of the amplified barcode contigs from a sample, representing the species composition expressed as proportional copy numbers. Each sample-specific contig is mapped to its identical barcode in the reference library and the list of species-specific mapped copy numbers to each reference barcode determines the OUT/ASV profile of the sample. Gray-highlighted boxes mark contributions of this study.

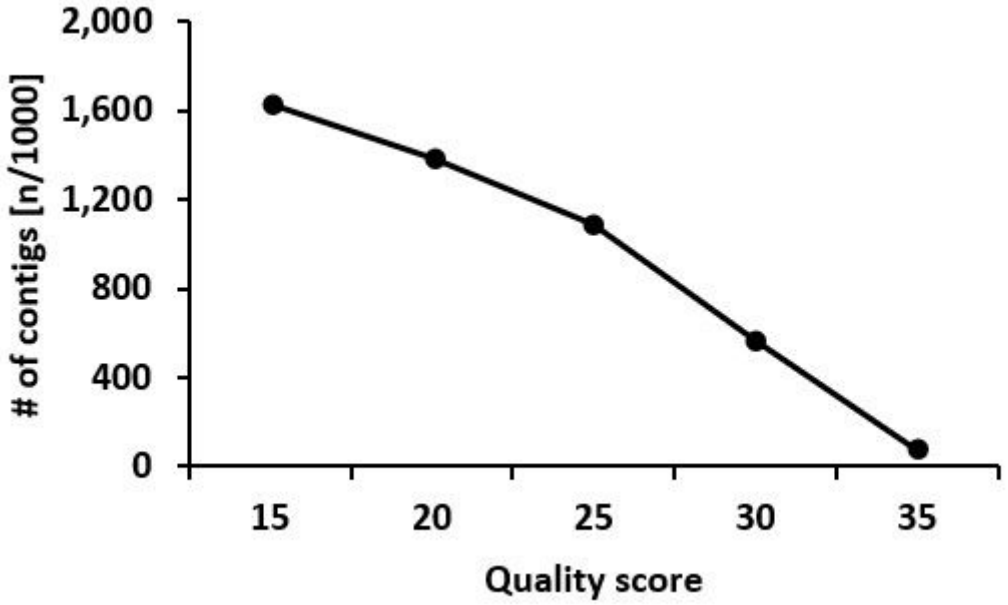

Figure 2

The effect of the the applied quality score during the CUTADAPT filtration procedure on the number of contigs which passed the filtration thresholds. The original assembly included 2,143,575 reads.

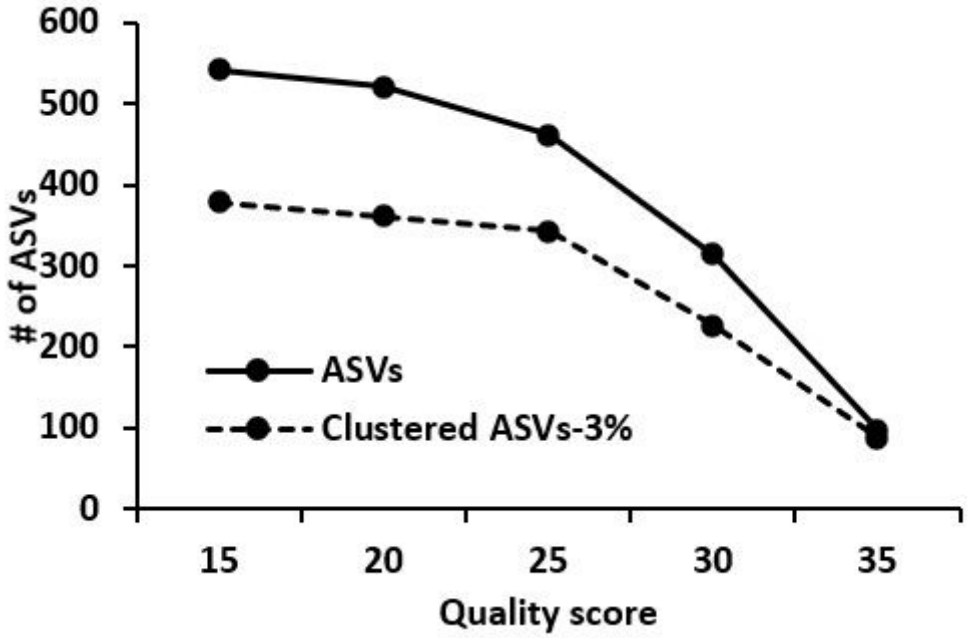

Figure 3

The effect of the the applied quality score and its underlying number of contigs (Fig. 2) during the CUTADAPT filtration procedure on the number of ASVs and clustered ASVs (<3% dissimilarity) resulted from the DADA2 procedure.

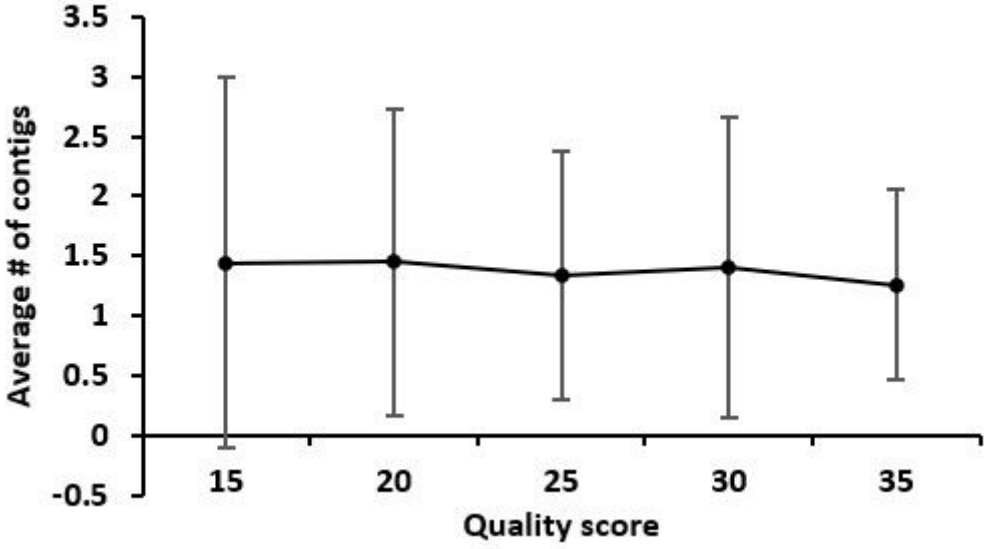

Figure 4

The effect of the the applied quality score during the CUTADAPT filtration procedure on the average number of unique ASVs / ASV cluster (<3% dissimilarity).

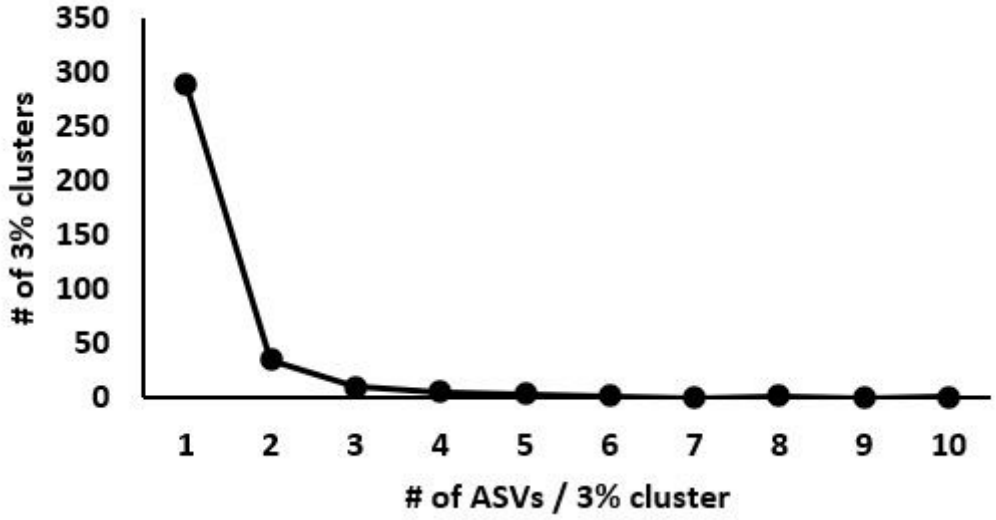

Figure 5

The distribution of the number of ASVs / 3% cluster in the q=25 reference library.

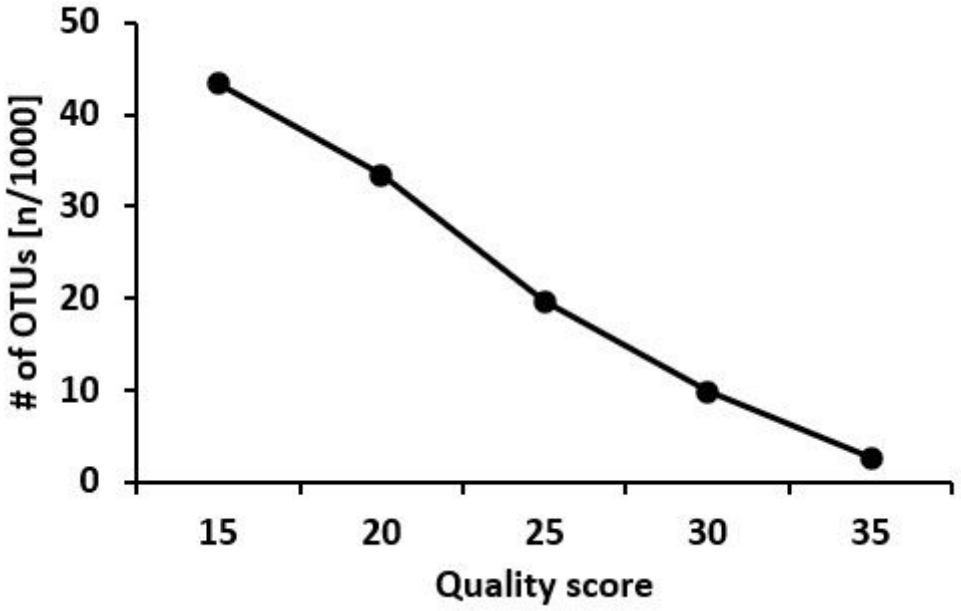

Figure 6

The effect of the the applied quality score during the CUTADAPT filtration procedure on the number of OTUs resulting from clustering procedure applying 3% dissimilarity threshold and performed using the VSEARCH software.

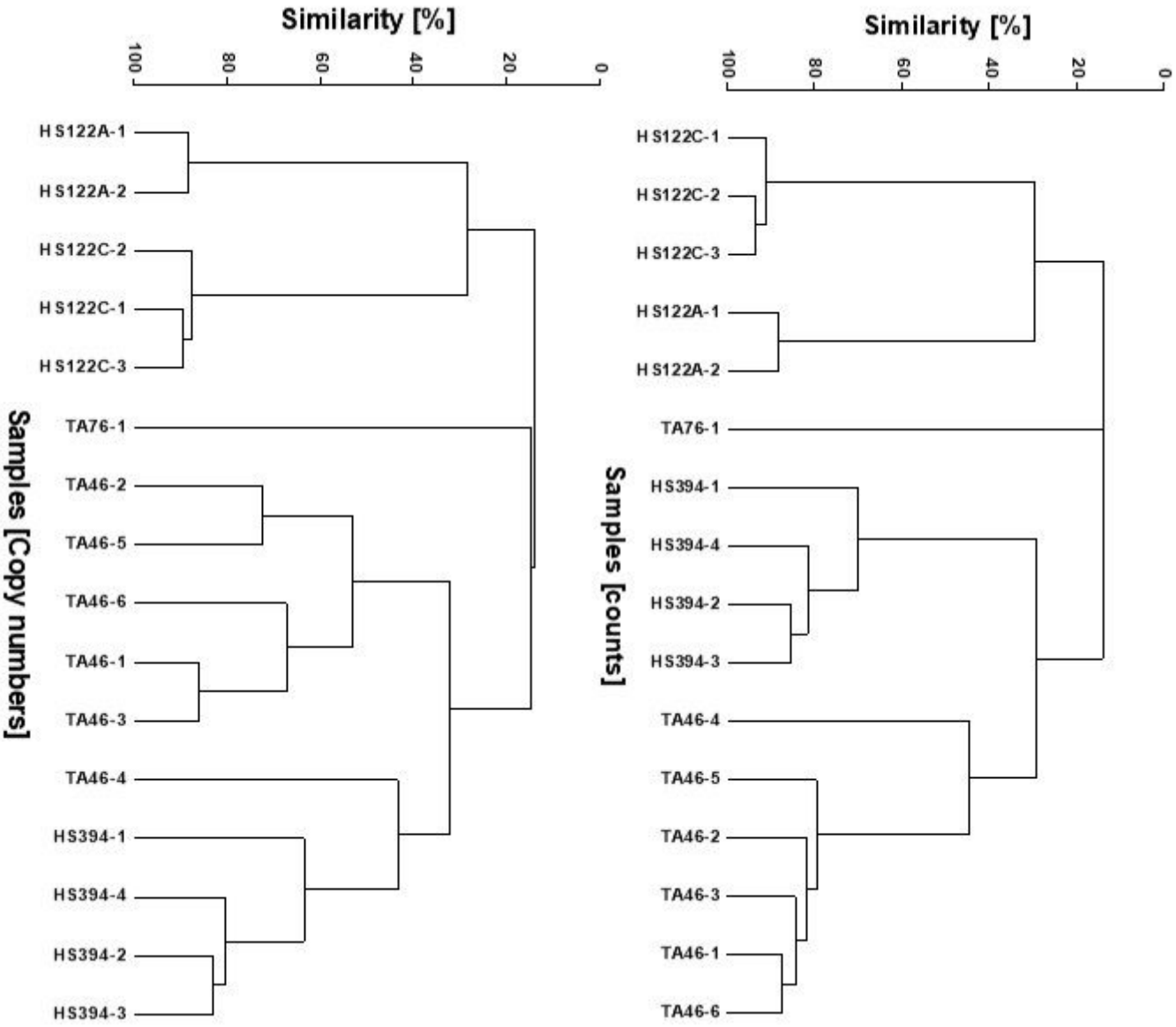

Figure 7

Comparison among sample profiles using the proportional copy numbers (left) and normalizations to real counts (right). The designations of the sample names were composed of the perpendicular to the coast transects, HS – Haifa transect and TA – Tel Aviv transect, the bottom depth in meters and the serial number of replicate PCRs done with each sample DNA as template

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable1.docx