Consensus calling of MinION amplicon reads improves metabarcoding results

MSc Research Project Report, Biology, Leiden University

Heleen Joke Bouwer

Student number: S2506408

MSc Biology and Business

Institution: Naturalis, Darwinweg 2, Leiden

Supervisors: Arjen Speksnijder and Rutger Vos.

November 1, 2019 - August 21, 2020

Journal format: BMC Bioinformatics

Abstract

Background: Metabarcoding environmental DNA with high-throughput sequencing is a state-of-the-art method to assess biodiversity and to uncover dark taxa. MinION is the first handheld sequencer that can be taken into the field for on-site metabarcoding. This research aims to answer if bioinformatics can solve the issues that arise because of the higher error rate of MinION data.

Results: Biodiverse samples with a presumed large portion of dark taxa were selected from the Dutch Caribbean. The cytochrome oxidase 1 gene is used as a barcode for identification at the species level or higher levels. Generating a consensus sequence from closely related sequences resulted in minimized random errors and increased species identification of 175% compared to unclustered MinION data. Additional to the formulation of the workflow, an ecological analysis was conducted that revealed co-occurrence of species in similar habitats. The proportion of dark taxa in the sampled region is 81.87%.

Conclusion: Although the workflow did not attain the results that the existing Illumina workflows do, the potential is evident. The high proportion of dark taxa in the sampled region of Statia and the Saba Bank indicates the need for continued barcoding of species in the Dutch Caribbean to resolve database limitations.

Keywords: *MinION, eDNA, metabarcoding, consensus, Saba, St. Eustatius, Illumina*

Background

Biodiversity strengthens ecosystems, yet the current rate of species loss is estimated to be 1000 times faster than background extinction, as observed in fossil records [1]. Biodiversity increases the resilience of ecosystems, the efficiency of nutrient cycles, and the stability of the communities within the system [2]. Biodiversity loss is characterized by decreases in taxonomic and genetic diversity, and reduction of species abundance within ecosystems. Ecosystems weaken because of biodiversity loss, thereby threatening people's availability of food and quality of life [3].

The distribution and abundance of marine species are changing due to overfishing and climate change. Overfishing causes decreases in the biomass of key predators in the trophic system, allowing primary and secondary producers to increase in abundance. Consequences of this are trophic cascades and homogenization of habitats [4]. Climate change causes increased ocean temperature and acidification which has led to bleaching of corals, thereby reducing the diversity and abundance of coral-associated fish species [5]. There is some evidence that deep-sea ecosystems are affected by climate change, although this has not been thoroughly investigated [6, 7]. The pressures posed on marine ecosystems worldwide indicate the need for conservation efforts of marine ecosystems.

The marine environments of the Dutch Caribbean islands Saba and Sint Eustatius (henceforth "Statia") are species-rich and biodiverse, with 307 fish species observed in Statia coral reefs and 207 species on the Saba Bank [8]. The marine landscapes of the islands include the Saba Bank atoll, nearshore coral reefs, lava fingers, and deep seas. The Saba Bank is an oceanic atoll, situated within a protected area of 2679 square kilometer, making it the largest nature conservation area in the Kingdom of the Netherlands [9, 10]. The surrounding deep sea has depths of up to one kilometer.

Following the global trend, the Caribbean coral reefs are subjected to the effects of climate change and anthropogenic impacts [11–13]. The Saba Bank shows reduced numbers of herbivorous and commercial fish families between 2011-2015, a trend that is likely caused by overfishing and degradation of habitat [14]. Moreover, the Saba Bank coral cover has decreased from 40-60% in the 1990s to 8% since the 1990s [14].

Whereas biodiversity on the Saba Bank has been well documented, biodiversity in the deep sea around the Saba bank has remained relatively unexplored. Thus, the deep sea around the Saba Bank is presumed to contain a large number of dark taxa. Dark taxa are lineages that are only known from sequence data but that are not matched to a specimen nor have a taxonomic name [15]. However, sequences that cannot be assigned to a specimen, but do align to one or more barcode database entries, can be identified at the lowest common ancestor (LCA) of the database hits [16].

To understand the impacts of climate change and anthropogenic activity, a Netherlands Initiative Changing Oceans (NICO) expedition travelled to the Saba Bank in 2018, with researchers

from several Dutch research institutions. The aim of the expedition was to establish a baseline of genetic biodiversity within the deepest realm of the Dutch economic zone.

Traditional taxonomic assignment relies on physical identification of species. However, especially in habitats that are challenging to access, the traditional method has limitations. Taxonomic records are incomplete because many species have yet to be described [17, 18]. Moreover, species identification requires taxonomic expertise that few people have, and even for experienced taxonomists it is hard to recognize species in their early developmental stages [19, 20]. Practicability is an issue in locations that are difficult to access, such as the deep sea [20]. Identifying species in the deep sea currently relies on remotely operated vehicles or invasive capture methods, as diving is impossible at high depths [10].

Effective management of marine ecosystems requires accurate and rapid biodiversity assessment [21]. Molecular tools for species identification can overcome the limitations that the traditional taxonomic method has and provide accurate, rapid, non-invasive, and economical biodiversity assessment [20]. DNA-based techniques can be complementary to traditional biodiversity assessment and add value to ecosystem management and environmental research [22]. The marine baseline study in the Dutch Caribbean utilized DNA-based techniques for this reason.

Barcoding for species identification

Novel biodiversity assessment techniques use genetic markers called “barcodes” to identify species. Barcodes are sequences of DNA from a standardized DNA region that are unique in distinct species, hence can be used as taxonomic markers [19, 23]. DNA barcoding is a standardized approach for extracting a DNA barcoding sequence from a species and storing it in a database matched with a specimen. The barcode can then be used for identifying specimen by matching the sequence to the database of recorded barcodes.

Successful barcoding depends on the specificity of standardized barcodes. Whether a sequence is suitable as a barcode depends on two critical factors. Firstly, barcodes must be variable between species, yet conserved within species [24]. This implies that the molecular evolutionary rate of the marker must be high enough to distinguish recently diverged species. Secondly, to standardize the

method of retrieving the barcoding region from the specimen, the barcode must have a robust and well-established set of primers that allows amplification of the barcode from the group of interest [19].

The most widely adopted barcode for multicellular organisms is the 658 base pair region of the mitochondrial cytochrome oxidase 1 unit (CO1). CO1 is primarily accurate for identifying metazoans [19]. Plant barcoding studies use the plastid regions and the internal transcribed spacer region (ITS) of nuclear ribosomal DNA [25, 26]. Fungal barcoding studies use the ITS region [26, 27]. The 16S region is used for barcoding bacteria [28].

Another factor for successful barcoding is the completeness of barcode databases. The Barcode of Life Data Systems (BOLD) is the primary public database for validated CO1 barcodes, barcode index numbers (BINs), and primers [29]. BOLD contains a curated library of over 7 million barcode sequences belonging to over 212,000 multicellular species. Moreover, BOLD provides BINs belonging to putative species and clusters of closely related individuals. BINs can be used to document biodiversity even when taxonomic information is not available for that barcode. BOLD is an initiative of The International Barcode of Life (iBOL). iBOL is a research alliance of universities, natural history museums and research centers that are dedicated to developing a global and accessible DNA-based system for identification of species [26].

Metabarcoding of environmental DNA for biodiversity assessment

Massive parallel sequencing of mixed amplicons has been termed “DNA metabarcoding” (henceforth “metabarcoding”) [30]. Metabarcoding is a high-throughput technique as it uses next-generation sequencing platforms to generate data. The difference between metabarcoding and barcoding is that metabarcoding is the process of identifying species in a mixed sample. In contrast, barcoding is sequencing a genomic region from a specimen from a natural history collection.

Seawater and ocean sediment are environmental samples that can be used to assess the biodiversity of marine life. Environmental samples contain intact DNA from single cells organisms and degraded environmental DNA (eDNA) [31]. Oceanic eDNA can originate from shed skin, scales, feces, reproductive cells, and other organic sources. DNA metabarcoding is a non-invasive and rapid technique, making it a convenient tool for biodiversity assessment.

Several studies have shown that eDNA metabarcoding to estimate fish biodiversity in marine environments is congruent with visual observations [20, 32]. Moreover, the detection probability of fish with eDNA metabarcoding is equal to or higher than traditional survey methods such as hand netting [22].

Using eDNA for metabarcoding benefits from spatial, temporal, and PCR replicates. Increasing the number of PCR replicates improves detection probabilities, especially for rare species [33]. Temporal replicates assure optimal recovery of total biodiversity and can shed light on seasonal community changes between and within seasons [34]. Likewise, spatial replicates are needed to recover the overall biodiversity of a sampling area.

Moreover, the results obtained from metabarcoding studies depend on the number of barcoded species in the area of interest. The coverage of barcoded species varies between regions. For example, North Sea species may have a higher coverage in barcode databases than oceanic deep-sea species. Low assignment rates of sequences to reference barcodes illustrates the incompleteness of barcode databases [32].

Small, portable, affordable sequencer: MinION

Labs commonly outsource next-generation sequencing (NGS) with Illumina to specialized companies. Outsourcing happens because acquiring an Illumina sequencer is a substantial investment. Consequently, when experiments are conducted in countries without NGS sequencer available, samples have to be transported abroad, adding time and costs to projects [35]. Generating direct data would accelerate the speed at which metabarcoding experiments can be conducted.

Oxford Nanopore Technologies (ONT) has developed the MinION, a handheld sequencer that generates direct data. The portability of the MinION allows it to be used from any location. MinION sequencing requires a simple library preparation and can be used with a laptop. All equipment needed fits into a single backpack allowing on-site use for generating barcodes and metabarcoding of eDNA [35, 36].

The MinION is portable and user friendly, making it a theoretically useful device for on-site biodiversity assessment. The MinION flow cell harbors nanopores set in an electrically resistant

polymer membrane through which a current passes [37]. When a DNA strand runs through a nanopore, each passing nucleotide disrupts the current. A sensor measures the current disruption, and the disruption is recorded as raw data. The raw data can be converted into a sequence of nucleotides; this process is referred to as “base calling”. MinION provides tools to perform either real-time or local base calling. Base calling, in this case, uses neural network algorithms, meaning the software optimizes decision making by providing it with control data. Neural networks allow for optimal base calling tailored to genetic regions and species. ONT offers users the option to train their own neural networks and has pre-trained models available [37].

Because ONT develops improved flow cells and updates and releases base calling software continually, the sequence quality generated by nanopore sequencing is increasing. The accuracy updates on the ONT community webpage show these improvements every month [37].

Nanopore sequencing data differs from second-generation sequencing experiments for two main reasons: errors and sequence read length. The error rate in nanopore sequencers is relatively high, namely 5-15%, compared to less than 1% error with the Illumina sequencing platform. Errors in nanopore sequencing data are both systematic and unsystematic. Systematic errors mainly occurring in homopolymers of more than five bases and cannot be eliminated when creating a consensus sequence [38]. The systematic errors originate from a shortcoming of the current technology to accurately record current disruptions caused by repeated passing of the same nucleotide. The highest number of systematic errors is within the first 10-15 base pairs that run through the nanopore [39]. Unsystematic errors occurring at random are correctable by generating consensus sequences, with barcode accuracies of 98-100% having been achieved [35, 36].

However, a limited number of studies have successfully used metabarcoding eDNA with MinION for biodiversity assessment on species level. While several studies have proposed bioinformatics pipelines, no standard workflow has been formulated for it [35, 36, 40]. The primary goal of this research was to set up a pipeline for metabarcoding with MinION that has equivalent or better performance as compared to the existing Illumina pipeline. With this in mind the main research question was:

- Can bioinformatics improve the quality of MinION amplicon sequencing data?

Practically, this translates to the following questions:

- How to base call, demultiplex, cluster, and BLAST MinION pooled sequencing data for metabarcoding given the high error rate and long sequences that nanopore data contain?
- Do high-quality reads give the same identifications as low-quality reads?

Assuming that random errors are a cause for the higher error rate in nanopore data, we hypothesize that part of the errors can be corrected by generating consensus sequences from a cluster. Supplying these sequences to BLAST should result in a higher identification rate of species as compared to using non-clustered data or centroid sequences of clusters.

The Illumina dataset can be used to answer questions concerning ecosystems dynamics:

- Does exchange between species take place across the Saba Bank and between the Saba Bank and St. Eustatius?
- What is the proportion of dark taxa within the sampled region?

Because the Saba Bank is an atoll, the hypothesis is that exchange between taxa across the Saba Bank is more likely than the exchange between taxa between the Saba Bank and St. Eustatius. The proportion of dark taxa is presumed to be high, as the biodiversity in the Dutch Caribbean waters, and especially the Saba deep sea, has not been extensively mapped.

Impact

The application of metabarcoding eDNA with MinION to assess biodiversity extends beyond the Dutch Caribbean and marine systems. A validated MinION approach would be a valuable tool for monitoring biodiversity, especially in the face of population growth, climate change and the anticipated continuation of decreasing biodiversity.

Answering the research questions as formulated above gives insight into the distribution of marine species in the Dutch Caribbean waters. Biodiversity assessment is essential for preparing local ecosystem management and fishery policies. Furthermore, a robust eDNA metabarcoding workflow will help to investigate biotic change after hurricanes, coral bleaching, and anthropogenic impacts.

Methods

Sampling Locations

Data from expeditions to the Saba Bank and Saint Eustatius was analyzed. The data was generated from sequencing DNA present in environmental samples taken during expeditions to Saba and St. Eustatius (Figure 1).

St. Eustatius In February 2015 an expedition to St. Eustatius to set up a baseline for marine biodiversity was organized. During this expedition, 21 environmental samples were collected from sediment and water for eDNA analysis with MinION. The rover-diving technique biodiversity surveys was carried out around the island at 35 locations of biodiversity hotspots like reef, boulders, and wrecks. Presence and absence of species was recorded during an observation time of sixty minutes at each spot. Moreover, photographs were taken and voucher specimen for DNA barcoding were collected.

Saba Bank In February 2018 a ten-day expedition to the Saba Bank was organized. During the expedition, 23 locations around the Saba Bank were sampled. A total of 64 environmental samples were taken, of which seven were sediment samples and 57 water column samples at different depths.

Sediment sampling was done using a boxcore. Water column samples were taken at different depths using a CTD instrument that simultaneously records water properties (such as temperature, depth, conductivity, pressure, salinity, density, oxygen). The CTD sampled the water column at different depths (5m above bottom, 50m above bottom, 100m above bottom, chlorophyll max layer, mixed layer). Before filling up the storage bottles with the samples, the bottles were cleaned out with chlorine and filled with water. This water was filtered and used as a negative control to detect contamination of the storage bottles. Then, the bottles were filled up with three replicate samples consisting of three equal parts water from three water samplers on the CTD (Appendix/Links/1). The

water samples were then filtered onboard of the ship (Appendix/Links/2). The filter, which contained organic material and eDNA, was put in a tube for transportation to the Naturalis Biodiversity Center Laboratories. The filtering material was cleaned with a chlorine solution before moving on to the next samples.

Additionally, specimens were collected using a dragnet, after which they were photographed and barcoded (Appendix/Links/3). Remote control video was taken to visually identify species (Appendix/Links/4).

DNA amplification and sequencing

Before the start of this project, library preparation and sequencing were performed at the laboratories of Naturalis Biodiversity Center.

From the water samples that were taken eDNA was extracted. Three PCR replicates were constructed. Each sample was amplified separately. Using IlluminaMETH030 MiSeq barcode amplification with PCR was conducted (Appendix/Links/5). The primers used were the universal CO1 primer BF2-il1 (forward primer) and BR2-il1 (reverse primer) with sequences 5'–GCHCCHGAYATRGCHTTYCC–3' and 5'–TCDGGRTGNCCRAARAAYCA–3', respectively [34]. Illumina index sequences were added to allow pooling of samples (Appendix/File 1). Each PCR well plate contained two control wells. One contained the negative control that was filtered water instead of DNA. The other contained positive controls with *Muntiacus reevesi* (Muntjac) to check for contamination during PCR.

Illumina sequencing

For this research Illumina data is the recognized exemplar of correctness, or "the gold standard". Before sequencing, the amplicons were extended with Illumina flow cell adapter sequences. The pooled control sample was sequenced with the Illumina MiSeq at BaseClear (Appendix/Link/6). Demultiplexing was done simultaneously with base calling. The output files were organized per sample.

The forward and reverse reads were merged using FLASH v1.2.11 (minimum overlap 50, max overlap 300, mismatch ratio 0.2, outies allowed) (Appendix/Commands/1) [41]. The remainder of data processing followed the same steps as MinION from adapter trimming and onwards.

Minion sequencing

From the prepared replicates, replicate B was sequenced with MinION. MinION library preparation required an additional step to the library preparation as described for Illumina. The ONT sequencing kit (SQK-LSK109k, version ACDE_9064_v109_revD_23May2018) for 1D sequencing was applied, without fragmenting the DNA (Appendix/Links/7). Per the MinION protocol, DNA CS lambda was added as a positive control. This is the 3.5kb region of the *Escherichia coli* lambda genome. The library was loaded on the SpotON flow cell version 9.4.1. Sequencing was carried out using a DELL XPS 15 portable laptop with Intel 9 core, 2TB SSD, and 32GB RAM.

Minion Base Calling

Three command line base callers were tested to base call the fast5 files generated by sequencing with MinION: Guppy, Bonito and Rerio. Base calling was done one the same DELL computer that the MinION sequencing was done with. Only the files in that were placed in the pass folder were base called.

Guppy The Guppy GPU version 3.5.2 with configuration file dna_r9.4.1._450bps_hac.cfg was used to base call the fast5 pass files [42]. The chunk size and chunks per runner were set to 1500, and four runners per device were set (Appendix/Commands/2).

Bonito The pass fast5 files were base called with the research release base caller Bonito version 0.1.0 [43]. The configuration file dna_r9.4.1 was used (Appendix/Commands/3). The output format of the data was in FASTA format.

Rerio Another round of base calling was done with a research release of the Rerio models for Guppy [44]. The configuration file used was res_dna_r941_min_flipflop_v001.cfg (Appendix/Commands/4).

Error estimation of reads using CS DNA

The positive control reads of the *E. coli* lambda genome were used to estimate the error rates of the base callers. First, a custom BLAST database of the lambda genome was created using the makeblastdb program [45] (Appendix/Commands/5). As the *E. coli* lambda DNA should be the only source of long reads, sequences with a length of 1500-4000 bases were selected using NanoFilt and retained for BLAST [46](Appendix/Commands/6). The retained reads were, for each base call method separately, BLASTed to the custom lambda database (Appendix/Commands/7). The error rate of the alignment with the lambda genome was visualized in a density plot using ggplot2 in R (Appendix/Rmarkdown/1). Mean, mode, median, of the error rate were computed in R, along with an ANOVA to test whether the differences between the error rates of the base callers were significant.

Quality scores of reads

At the time of this research, the Bonito base caller did not generate FASTQ scores. Thus, only the FASTQ scores of Guppy and Rerio reads could be investigated. The quality scores of Rerio and Guppy were computed using Nanoplot [46] (Appendix/Commands/8).

Metabarcoding workflow

Removing lambda reads

The reads belonging to the control *E. coli* Lambda DNA were removed using NanoLyse (Appendix/Commands/9). The reference database used was the same custom lambda DNA reference database that was constructed for determining the error rates [47].

Demultiplexing

The output of base calling with Rerio was used for testing the metabarcoding workflow. First, reads were demultiplexed with Minibar, which allows for demultiplexing with custom tags [48]. The samples had to be separated based on the eight-base dual Illumina tags. First, the tags were structured in a text file according to the input requirements of Minibar (Appendix/Command/10, Appendix/File 2). Because the edit distance between the Illumina tags was as low as 2, the percentage identity was set

at 0.8. This will only match a read to a sample if it has six matching bases. Minibar searched for tags within 90 bases at the tails of the sequence and trimmed the adapter and anything occurring before it (Appendix/Commands/11).

Adapter trimming

The priming adapters and all nucleotides preceding the adapter were removed using two rounds of cutadapt v2.10 [49](Appendix/Commands/12). For both rounds the max error rate was set at 0.2. Because double-stranded DNA is loaded on the Nanopore flowcell, the reads are in both forward and reverse orientations. The first round looked for the front adapter on either side of the read, and the reverse complemented read. Reads that did not contain a front adapter were discarded. The next round of cutadapt was only done for the reads where the front adapter was found. The next round of cutadapt trimmed the reverse adapter. During the first round all reads were automatically oriented in the same orientation, so the second round does not require a reverse complement search. Length filtering was set to minimum 400 bases to maximum 440 bases to allow for natural variation in sequence length. The reads of which both adapters were found, trimmed, and that passed the length filter were saved for clustering.

Chimera check, clustering, and consensus

The reads were clustered into operational taxonomic units (OTUs) with VSEARCH [50]. A chimera check was done with VSEARCH (Appendix/Commands/13). Several rounds were done using clustering percentages of 80, 85, 90, 95, and 97 using the CD-HIT definition for clustering (Appendix/Commands/14). To accommodate sequence alignment for the higher error rate of Minion reads, the gap open and the gap extension penalties were set at 60 and 4, respectively. Input sequences were supplied to VSEARCH sorted on descending order of length, meaning longer sequences were preferred as a centroid. VSEARCH computed a consensus sequence from the resulting OTUs, based on residue frequency per nucleotide position. For comparison, the Illumina sequences of replicate B was clustered at 97%.

BLASTN

The non-clustered data, and the resulting centroid and consensus sequence for each clustering percentage were identified using BLAST [51] (Appendix/Commands/7). A local version of the BOLD database, containing only species and no duplicates, was used to assign reads to a database entry (sequences downloaded February 6, 2020). Both the read and the reverse complement of the read were aligned to the database to find the best alignment. For each query, the top hundred hits above both 70% query coverage and 70% hit identity were retained for LCA determination.

Lowest common ancestor

LCA analysis can assign taxonomy for reads that cannot be assigned a species-level identification. To prepare the BLAST output for LCA analysis, the taxonomy of the species was added to the BLAST file (Appendix/Commands/15). To this file, a column containing a dummy variable was added to prepare the result for the next step (Appendix/Commands/16). To identity the lowest common ancestor of the top 100 BLAST hits, an LCA script was applied [33] (Appendix/Commands/17). The LCA script removed any hit that contained the string "environmental" in its taxonomy. Then, the OTUs had exceeded both the top identity threshold of 98 percent and the top query coverage threshold of 99 percent were assigned at the species level of the hit. All unique top hits above 98 percent identity and 99 percent query coverage were recorded as a species-level identification for that OTU.

The LCA analysis continued for the remaining OTUs of which no species-level identification was found. Hits that had a bitscore within eight percent of the highest bitscore and had identity percentages of 70 percent and query coverage of 70 percent were retained for LCA. For example, if an OTU sequence has a top BLAST-hit with a bitscore of 200, the bitscore threshold for that OTU is set at 184. Any other BLAST hits of that OTU that have a bitscore higher than 184, and an identity percentage and query coverage of 70 percent are retained for LCA analysis. The LCA was found by aggregating the BLAST-hits per taxonomic rank on genus level and higher. If on genus level there is a unique genus listed one or more times, this will be set as that genus. However, if there are two genera listed, the family level is checked, etcetera (Table 1). If no unique taxon is listed at any rank, "no

identification" was assigned to the read. To see the difference in the results, the LCA analysis was also conducted with the bitscore set at twelve percent.

Comparison of methods

Hit counts The performance of the clustering methods compared to each other and the Illumina gold standard was evaluated based on the lowest common ancestor. The results were accumulated and visualized in R using a custom script using the ggplot package (Appendix/Command/18). The number of unique hits on every taxonomic level was counted for the consensus and centroid of each clustering percentage. The same method was followed for Illumina data. The counts from all methods were plotted in a bar plot altogether, allowing for comparisons of clustering percentages their consensus and centroid.

Overlap The overlap between the identifications on each taxonomic level for each method was computed with a custom scrip in R. Venn diagrams per taxonomic level were created using ggplots in R to allow for visualization of overlap (Appendix/Command/19).

Ecological Analysis with Illumina Data

The demultiplexed Illumina sequencing data, consisting of three PCR replicates, was used for the ecological analysis. All three replicates were merged with flash as described in the previous section (Library preparation>Illumina sequencing>merging). The replicate files were then combined using the additive strategy, meaning the replicates were combined per sample to recover as many taxa as possible. The combining was done using a custom script (Appendix/Commands/20). The samples were trimmed as described before (Metabarcoding workflow/adapter trimming). Reads of length 400-440 were retained.

Chimeras were checked per sample by VSEARCH *de novo* option for chimera checking. Duplicate sequences were removed from the data by doing a round of VSEARCH clustering at 100%. The dereplicated FASTA files contained the original abundance information of the sequences so read numbers could be computed for the OTU table.

For every sequence in the sample, a prefix was added to the FASTA header to indicate the sample it was from (Appendix/Commands/21). All sequences were then combined into one FASTA file

(Appendix/Command/22). This FASTA file was clustered at 98% with VSEARCH with the CD-HIT definition of identity (Appendix/Command/14). If sequences from different locations appeared in the same cluster at 98% identity, it was assumed they belonged to the same species and thus to the same OTU. The clusters that contained reads from more than one samples can indicate that OTU of one species appeared in both samples.

Control filtering

In R, the positive, negative, and storage bottle control were filtered out using abundance filtering. The abundance filter was set at 0.0061%, which was the contamination rate of the positive control. First, low abundance OTUs were filtered out if they occurred at 0.0061% of the sample total. This means that if a sample has 16000 reads, any OTU containing two reads or less will be disregarded for further analysis. Thereafter, all remaining singletons were filtered out, and all samples containing less than 2000 reads.

An additional check on the storage control bottle was done on the Saba samples. Per OTU belonging to any Saba sample, if a sample contained less than twice the number of reads that appeared in a control, the reads belonging to that sample were filtered out. This ensured that any contamination that occurred either during sampling on the boat during sampling was removed from further analysis.

The number of OTUs per sample were computed and plotted per sample type (sample, bottle control, PCR positive control, PCR negative control). The reads of the storage bottles were BLASTed to investigate the major sources of contamination that occurred during sampling.

BLAST

After removing the control samples, the remaining OTUs were BLASTed twice. First, the OTUs were BLASTed to the full Genbank database (sequences downloaded 5 Mar 2020). An LCA analysis was done on the Genbank BLAST results to remove any LCAs classified as prokaryotes from the OTU table in order to reduce noise in the data. The remaining OTUs were BLASTed with the CO1 BOLD database (sequences downloaded 6 Feb 2020). The curated status of the BOLD database is another way to reduce noise and produce accurate results.

Sampling location patterns

The sampling location data of Saba and Statia were combined in R. The longitude latitude altitude data were selected from both frames. A tag was added to indicate where the sample was taken (Saba Bank South, Saba Bank North, or Statia). The number of OTUs found that were unique to one region, OTUs that are shared between regions, and OTUs that occur between all regions were visualized in a plot. A chi-square test of independence was conducted on observed and expected exchange between regions.

A 3D plot was constructed using the plotly package in R. The bathymetry data was downloaded for free from The General Bathymetric Chart of the Oceans in GeoTiff format. The sampling locations were plotted on this map to visualize the spatial distribution of sampling points. To this plot, lines were added between locations that shared eDNA.

A principal component analysis (PCA) was done on all sample locations, and binned location data. The OTU values for habitats within a region were summed, thereby reducing the dimensionality of the data. PCA was done for presence/absence of OTUs. The mean number of OTUs found per habitat data were plotted with ggplot and tested for statistical differences. The Statia data was binned into "Statia Mixed". Another PCA was done containing all samples. All code is publicly available (Appendix/Rmarkdown/2).

Maps and networks

The OTU table was used to construct a network analysis. The OTU table contains all the OTUs as rows, and sampling locations as columns. The cells contain how many reads belonging to that OTU have been found at the corresponding location. The creation of the table listing all exchanges is described in R. From the exchange table, a nodes and edges list was created by taking all sampling location that had at least on exchange with another location. These combinations were used to plot lines between sampling locations on the 3D map to indicate similarity.

Additionally, an interactive 2D map was created with Leaflet. All sampling points were added to the map. When clicking on an item, information is displayed about the location found, along with an

image of the species. The images were harvested from the BOLD, Naturalis, and Atlas of Living. The code and a map are available on GitHub (Appendix/Rmarkdown/3).

Results

Sequencing Statistics

Minion sequencing ran for 20 hours. The run generated 23,4 GB of raw data, equal to 5,314,990 reads. Of these reads, 1,096,176 passed the pass filter that is inherent to the Minion software.

Base calling took three hours with standard Guppy, five hours with Rerio model, and over twelve hours with Bonito. The number sequences base called with each method was 1,096,176. The number of bases base called differed per method. Guppy base called 608,404,021 bases, Bonito base called 561,920,554 bases, and Rerio base called 610,182,198 bases.

The plot of the error percentage from the lambda reads that were extracted from each of the three base calling procedures suggested that the distributions of the errors are not identical (Figure 2). Summary statistics show that Rerio has the lowest error rate, and additionally the lowest standard deviation in error rate. A one-way ANOVA was conducted to compare the means of the distributions of error rate of alignment between base calling with Guppy, Bonito, and Rerio. There were significant differences between the base callers [$F(2, 106917) = 1921$, $p < 0.001$]. A Tukey's Honestly Significant Difference post hoc test shows that there are significant differences between all the combinations ($p<0.001$).

Metabarcoding with Minion

The Rerio data was used for testing the workflow as it had the lowest error rate. The total reads had an average quality score of 13.0. After the first round of cutadapt 421,525 reads that had no front adapter thus were discarded. The remaining reads had a read quality of 13.3. After the second round of cutadapt, another 240,150 reads were discarded. After two rounds of cutadapt, 434,501 reads were saved. The reads saved had a mean quality score of 13.7. 289,903 reads that passed the filter of length

400 to 440. These reads had a mean quality of 15.2, which corresponds to a mean error rate of 3.02 percent.

Clustering at 80, 85, 95, and 97% resulted in clusters with respective singletons proportions of 65.5%, 73,8%, 81.1%, 90.2% and 95%. At all clustering percentages, the hit counts for MinION were higher for consensus sequences than centroid sequences (Figure 3). Consensus sequences at 85% clustering percentage have the highest diversity in identified species. For Illumina, generating a consensus sequence from clusters did not result in a higher number of species identifications.

Overlap between identifications

BLASTing the consensus sequence of the MinION data clustered at 85% resulted in a 175% increase in overlapping identifications as compared to unclustered data (Figure 4). The consensus sequence at 85% identified 22 out of 41 species that Illumina also identified. Additionally, it identified ten species that Illumina did not identify at species level. Two of these reads were identified on the genus level in Illumina. The other eight reads were either of low quality or had other high scoring matches indicating that these species may be false positives.

The species that both unclustered and clustered MinION data find is likely a true hit as it has a high read quality, and there are several reads in the data that have high identity with this read.

Ecological analysis with Illumina data

The full Illumina sequencing run with three PCR replicates generated 4,164,985 read pairs that belonged to the samples. All replicates were used for the ecological analysis. After merging the forward and reverse read and merging the replicates, 4,063,608 reads were retained for adapter trimming. On 97.3 percent of the reads (4,052,211), both adapters were found and trimmed. Of these reads 62.0 percent passed the length filter, meaning 2,512,237 reads were retained for chimera check. The mean read number of the samples and controls was 25,122 (sd = 12,498.7), with six samples that had read numbers less than 4000 reads.

Chimera checking with VSEARCH chimera did not identify any chimeras. Clustering at 98% resulted in a total of 343,801 clusters, of which 236,155 were singletons. The average cluster size per sample was 3438.0 (sd = 2800.9).

The positive controls containing the *M. reevesi* OTU with average 22,368 reads, shows one-read contamination in three samples, corresponding to a contamination rate of 0.0067%. The negative control had a contamination rate of 0.061%.

After filtering on low abundance of 0.0067%, the number of OTUs that were retained was 52,663. This means 82.26% of OTUs was discarded. Singleton removal discarded 12,254 OTUs. The additional storage bottle control for the Saba samples removed another 33 OTUs. Any sample with less than 2000 reads in the total of OTUs was discarded. The 40,376 OTUs were aligned against Genbank for identification of prokaryotes. After removing prokaryotic OTUs, the remaining 17,948 OTUs were aligned to the BOLD database.

Reads and OTUs per sample

The mean number of OTUs per sample between regions, after filtering on abundance and controls, was not significantly different as determined by a one-way ANOVA ($F(2,86) = 0.098$, $p = .383$).

Per habitat there were significant differences between mean number of OTUs per sample found as determined by a one-way ANOVA ($F(5,82) = 5.78$, $p < .001$) (Figure 5). The Tukey post hoc revealed significant differences ($p > .05$) between the habitats 0.05 m above bottom ($n = 7$, $m = 207.6$, $sd = 131.4$), and 50 meter above bottom, ($n = 14$, $m = 1262.79$, $sd = 510.5$, $p < .05$), the chlorophyll max layer ($n = 12$, $m = 1750.7$, $sd = 523.3$, $p < .001$), the mixed layer ($n=15$, $m = 1745.0$, $sd = 1282.7$, $p < .001$), and Statia mixed layer ($n = 25$, $m = 1494.6$, $sd = 627.1$, $p < .01$).

The mean number of OTUs per sample type had statistically significant differences as determined by a one way ANOVA ($F(3,96) = 7.827$, $p < .001$) (Figure 6). A Tukey post hoc test revealed that the number of OTUs in the samples ($n = 89$, $m = 1361.3$, $sd = 812,9$) was significantly higher than the storage bottle control ($n = 7$, $m = 23.3$, $sd = 10.9$, $p < .001$). No statistically significant differences

were found between the samples and the PCR negative control (n = 2, mean = 5.0, sd = 2.9, p = .07) and the PCR positive control (n = 2, m = 4.0, sd = 1.4, p = 0.07).

A BLAST of the sequences in the storage bottle controls revealed that there were several genera were sources of contamination of which *Ephinephelus* was the most abundant, having 75% of the total identifications at the genus level (Supplementary Figure 1). At the species level, *Ephinephelus guttatus* (common name: Red Hind) represented 95% of total identifications. A BLAST of the positive control revealed that the genus Muntiacus, which is a member of the deer family, was overrepresented in terms of read numbers.

Post BLAST

From the 17,948 OTUs aligned to BOLD with BLAST, 3,848 had at least one BLAST hit with an identity percentage of 70 and coverage of 70. The LCA analysis resulted in 3,249 identifications at the kingdom level or lower (Figure 7). The remaining 599 reads were classified as "no identification" either because they contained a filter string ("unknown" or "environmental"), or because no consensus could be established on any taxonomic level. The proportion of molecularly dark taxa in the sampled region is 81.87%.

Community composition

The analysis of all samples shows that habitats are separated in 2D space along axis based on presence/absence explaining 25% of the variance (Figure 8). The PCAs per habitat shows that habitats are separated in 2D space along axis based on presence/absence explaining 56.1% of the variance (Figure 9). For both PCAs, the two dimensions that represent the axis have eigenvalues larger than 1. The PCA shows patterns of habitats along the water column cluster together, with the largest distances between similar habitats observed between 50 meter above bottom of Saba North and Saba South (SN 50 mab and SS 50 mab) and the chlorophyll max layers of Saba North and Saba South. The sediment samples explained least of the variance in both PCAs.

When analyzing the number of OTUs that are shared between the three regions, Saba North and Saba South share more species than Statia and the Saba regions (Figure 10). A chi-square test of

513 independence confirmed that the observed exchange frequencies were dependent (X^2 (2, N = 3609) =
514 2929.2, p < .001).

515 Discussion

516 This study described a new metabarcoding workflow with the MinION sequencer. The
517 differences between the error rates of the Guppy, Bonito, and Rerio base callers show how, in a short
518 span of time, the data quality of MinION sequencing has improved. Updates in base calling software,
519 and sequencing platforms from Oxford Nanopore Technologies enhance the quality of the data.

520 The samples used in this study were tagged with eight-base tags that were designed for
521 demultiplexing samples that were sequenced by Illumina. The insufficient demultiplexing results of
522 these tags were in line with another study that used barcodes with low edit distances for demultiplexing
523 MinION data [48]. Longer tags in combination with tags that have higher edit distances are essential to
524 improve the recovery of reads in future MinION experiments. Another potential solution could be
525 rolling-circle amplification, which is a new method of DNA amplification that exploits the long-read
526 sequencing ability of the MinION [38]. However, this technique is presently applicable for sequences
527 of approximately 100 base pairs. When rolling-circle amplification becomes available for amplicons of
528 more than 400 base pairs, the current eight-base tags could work. Presuming that increases in data
529 quality and innovations in amplification techniques will continue, accurate metabarcoding with
530 MinION will become more feasible in the future.

531 The main finding of this study is that using consensus sequences from MinION data improves
532 the quality of the data and leads to improved taxonomic assignment of OTUs on the species level.
533 Constructing consensus sequences of MinION read clusters at 85% improved the overlap of
534 identifications with the Illumina method by 175%, compared to unclustered data. These findings
535 confirm the hypothesis that better quality sequences lead to better identifications. Moreover, the results
536 presented in this study suggest that future metabarcoding studies with MinION will benefit from using
537 consensus sequences for taxonomic identifications.

However, the MinION workflow found 53.7% of species that the Illumina method found. This result deviates from results from other studies, where MinION identifications were on par with Illumina metabarcoding [48, 52]. It is important to note that these studies used mock samples, using the 18S region for arthropods [48] and CO1 region for fish [52]. In contrast, the samples used in the current study were highly biodiverse, with many closely related species. Closely related species are more likely to cluster together when clustering at 85%, which may be one explanation for the low identification rate. This issue has already been highlighted by other researchers [48]. Future studies could investigate how clustering percentages affect results in mock communities with closely related species. The true success of a metabarcoding workflow can only be established when the content of the mixed sample is known [23].

The CO1 region used in this study is the main region for barcoding animalia. The CO1 marker performs well on environmental samples that contain species that are evolutionary distant from each other, in combination with a high-quality database. In reality, environmental samples often contain closely related species, whose CO1 regions have high homology because of their recent evolutionary divergence. Species may be assigned to a wrong database entry that belongs to a closely related relative, an effect that is strengthened by low database coverage. The LCA analysis could correct these misassignments by assigning the OTU at a lower taxonomic level.

Because the quality of the MinION reads is relatively low, the LCA analysis does not perform as well as it does for the higher quality Illumina data. A potential solution could be barcoding with longer regions to compensate for the higher error rate and increase taxonomic resolution. For example, the 18S region or a region spanning multiple genes [48, 53]. However, with longer barcoding regions, variation in region length causes PCR bias, with shorter sequences preferentially amplified over longer sequences [48].

The analysis of community patterns within the Caribbean region was done with Illumina sequencing data. These samples could be accurately demultiplexed, allowing for analysis of the contamination that arose during sampling and PCR. Between the CTD sampling device and the filtering step, the sampled water was stored in bottles. Although these storage bottles were cleaned out with chlorine before filling them with the CTD and box core samples, the storage bottles contained sources

566 of contamination. It demonstrates that even when diligently sterilizing bottles, contamination occurs
567 while relocating sample material from the sampling device to a storage bottle. For example, *E. guttatus*
568 was consumed for lunch on one of the sampling days, explaining its high abundance in the control
569 samples.

570 While contamination is close to inevitable with sequencing experiments, it may be reduced by
571 optimizing decontamination of storage bottles and increasing sterility in the area where the sampling
572 devices are emptied. The contamination risk could technically be reduced even further by omitting the
573 storage bottle step altogether and performing all laboratory work on a cleanroom aboard the ship.
574 Technically, this would be possible when sequencing with a device like MinION. The practical
575 feasibility of onboard sequencing and its effect on contamination rates, could be tested in future on-site
576 metabarcoding experiments.

577 As previously described for lower quality MinION data, the percentage at which reads are
578 clustered influences the results of metabarcoding studies. Likewise, for Illumina data, the clustering
579 percentages should not be too low, as this will result in closely related species ending up in the same
580 cluster, resulting in an underestimation of diversity. However, setting the clustering percentages too
581 high results in multiple clusters for the same species, which may result in an overestimation of diversity.
582 This is not problematic when dealing with a sample for which the content has been well documented in
583 barcode databases, because this can be corrected for after matching multiple OTUs to the same database
584 entry. However, especially when dealing with samples with a high proportion of dark taxa, such as the
585 samples used in this study, the proportion of dark taxa may be overestimated when clustering high
586 percentages.

587 The clustering percentage chosen in this study, 98%, is equal to the community accepted cutoff
588 for alignment identity with a barcode [23, 33, 48]. A clustering percentage cutoff at 98% allows for
589 some natural variation in sequences and gives some space for potential sequencing errors. A BLAST
590 identification cutoff at 98% prevents multiple high matches with species in a database. Another
591 consequence of clustering at high percentage is the high portion of singletons. The consensus is that
592 singletons should be considered as sequences, as they are highly likely to contain errors [20, 54]. Thus,
593 setting the clustering percentage higher than necessary results in a loss of data.

The LCA script used in this study works well for samples of which the database coverage is high. However, in this study the LCA script results in partly biased identifications due to the low database coverage of the sampled region. More specifically, the LCA analysis has a bias toward genus-level identification if the number of reads that fall within the set bitscore range of that OTU is low. As touched upon before, this effect was even higher with Nanopore reads, probably because the quality of sequences was lower. For example, if an OTU has one read that falls within the top bit score threshold, it is automatically set at the genus-level, regardless of whether it is a 70% identity hit or a 97% hit. In reality, a sequence that has 70% identity with a database entry likely does not share a genus with this entry, but with a higher rank identification such as phylum. As demonstrated in this study, setting a higher bit score range moved the bias away from genus level but has as drawback that it leads to higher numbers of unidentified OTUs. The LCA script could be adapted to accommodate for more representative assignments when the top blast hits belong to one species but have low identity.

The 81.87% of dark taxa confirms the suggestion that much of the Caribbean biodiversity remains undocumented in barcode databases [8]. There are important biases to consider when dealing with metabarcoding in a region with low species coverage in barcode database. Firstly, the top hits may belong to taxa that are not present in the sampling area, because the correct species is not present in the database. For the Saba Bank and Statia region, the full biodiversity nor genetic biodiversity has been mapped, making it challenging to confirm whether the found taxa are actually present in the area. For example, for the Saba Bank, none of the species that the Illumina metabarcoding method that were found had previously been observed on the Saba Bank (Appendix/Links/8). This emphasizes the need for continued biodiversity research and barcoding efforts around the Saba Bank.

The results from this study suggest that there are OTUs that occur in both Statia and the north and south side of the Saba Bank. The north and south side of the Saba Bank show a higher number of shared OTUs than either side of the Saba Bank and Statia. Assuming that an OTU represents a species, our initial hypothesis that the atoll feature of the Saba bank enables exchange more than the deep sea between the Saba Bank and Statia can be confirmed. However, as the PCA indicated, the habitats on both sides of the Saba Bank are more alike than the habitats of Statia, which may explain the higher number of shared OTUs.

The PCA results suggest that habitats account for more variation in OTU presence than regions do. However, within a habitat, distinctions can be observed between regions, suggestion that there is a combined effect of region and habitat on the presence of species. Future studies could apply further ecological analysis to the existing data that may reveal additional community patterns.

In both PCAs the sediment samples (SN & SS 0.05 mab) have least weight on the principal components shown in the graph. This indicates that the presence of OTUs in these samples are least useful to predict what habitat the sample was from. One explanation could be that the OTU diversity was lowest in the sediment samples. Another plausible explanation for this observation could be that the eDNA from the water column settles on the sediment, leading to higher concentration of eDNA in the sediment, with a composition that resembles the aquatic samples. Evidence has been presented that the concentrations of fish eDNA are higher in aquatic sediments than in surface waters of fresh water bodies [55]. The results from the current study suggest that this may also be the case in deep sea.

In conclusion, creating consensus sequences from clusters at 85% when sequencing with MinION, improved the results of metabarcoding marine water samples in the Dutch Caribbean. Moreover, there is exchange between species across the Saba Bank and between Statia and the Saba Bank. The Saba Bank is the largest nature conservation area in the Kingdom of the Netherlands, and likely the most unexplored region. The proportion of dark taxa in the Saba Bank and Statia region is high, indicating the need for further barcoding studies before taxonomic biodiversity can be estimated accurately.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publications:

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the Zenodo repository. DOI:10.5281/zenodo.3992310

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Competing interest

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

Arjen Speksnijder has contributed to collecting water samples during the expeditions, and acquisition of data.

Elza Duijm has executed the laboratory work that was necessary for acquiring the sequencing data. Arjen Speksnijder and Rutger Vos have contributed to the design of the study and to the revision of the work.

Acknowledgements

The authors thank Dick Groenenberg (Naturalis) for the contribution of Galaxy support and the support with coding.

Author information

Heleen Bouwer is a Master student in Biology at Leiden University, Netherlands. This article was written as part of an internship at Naturalis Biodiversity Center

Arjen Speksnijder is a researcher and the head of laboratories at Naturalis Biodiversity Center.

Rutger Vos is a bioinformatician and researcher at Naturalis Biodiversity Center, and a lecturer at the Institute of Biology of Leiden University, Netherlands.

References

1. Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, et al. The biodiversity of species and their rates of extinction, distribution, and protection. Science (80-). 2014;344.
2. Tilman D, Isbell F, Cowles JM. Biodiversity and Ecosystem Functioning. Annu Rev Ecol Evol Syst. 2014;45:471–93. doi:10.1146/annurev-ecolsys-120213-091917.
3. Rounsevell M, Fischer M, Torre-Marin Rando A, Mader A. The regional assessment report on biodiversity and ecosystem services for Europe and Central Asia.
4. Kirby RR, Beaugrand G, Lindley JA. Synergistic effects of climate and fishing in a marine ecosystem. Ecosystems. 2009;12:548–61.
5. Hoegh-Guldberg O, Bruno JF. The Impact of Climate Change on the World's Marine Ecosystems. Source Sci New Ser. 2010;328:1523–8.
6. Levin LA, Bris NL. The deep ocean under climate change. Science. 2015;350:766–8.
7. Yasuhara M, Danovaro R. Temperature impacts on deep-sea biodiversity. Biol Rev. 2016;91:275–87.
8. Davies MR, Piontek S. The marine fishes of St. Eustatius, Dutch Caribbean. Mar Biodivers. 2017;47:27–35.
9. Williams JT, Carpenter KE, van Tassell JL, Hoetjes P, Toller W, Etnoyer P, et al. Biodiversity assessment of the fishes of Saba Bank atoll, Netherlands antilles. PLoS One. 2010;5.
10. National SB, Netherlands P. United Nations Environment Program. Ecological criteria Cultural and socio-economic criteria Representativeness Conservation value Rarity Naturalness Critical habitats Diversity Connectivity/coherence Productivity Cultural and traditional use Socio-economic benefits.
11. Cardinale BJ, Srivastava DS, Duffy JE, Wright JP, Downing AL, Sankaran M, et al. Effects of biodiversity on the functioning of trophic groups and ecosystems. Nature. 2006;443:989–92.
12. Carpenter KE, Abrar M, Aeby G, Aronson RB, Banks S, Bruckner A, et al. One-third of reef-building corals face elevated extinction risk from climate change and local impacts. Science (80-). 2008;321:560–3.

13. Bakker DM, Duyl FC, Perry CT, Meesters EH. Extreme spatial heterogeneity in carbonate accretion potential on a Caribbean fringing reef linked to local human disturbance gradients. Glob Chang Biol. 2019;25:4092–104. doi:10.1111/gcb.14800.
14. Becking LE, Meesters E. A_4_3_3 V04 DATE. 2017. www.wur.nl/marine-research.
15. Ryberg M, Nilsson RH. New light on names and naming of dark taxa. 2018. doi:10.3897/mycokeys.30.24376.
16. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17:377–86. doi:10.1101/gr.5969107.
17. Costello MJ, Wilson S, Houlding B. Predicting total global species richness using rates of species description and estimates of taxonomic effort. Syst Biol. 2012;61:871–83.
18. Eschmeyer WN, Fricke R, Fong JD, Polack DA. Marine fish diversity: History of knowledge and discovery (Pisces). Zootaxa. 2010;:19–50.
19. Hebert PDN, Cywinska A, Ball SL, Dewaard JR. Biological identifications through DNA barcodes.
20. Yamamoto S, Masuda R, Sato Y, Sado T, Araki H, Kondoh M, et al. Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. Sci Rep. 2017;7.
21. Day J. The need and practice of monitoring, evaluating and adapting marine planning and management-lessons from the Great Barrier Reef. Mar Policy. 2008.
22. Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, et al. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Mol Ecol. 2016.
23. Günther B, Knebelsberger T, Neumann H, Laakmann S, Martínez Arbizu P. Metabarcoding of marine environmental DNA based on mitochondrial and nuclear genes. Sci Rep. 2018;8:14822. doi:10.1038/s41598-018-32917-x.
24. Pentinsaari M, Salmela H, Mutanen M, Roslin T. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life OPEN. Nat Publ Gr. 2016. doi:10.1038/srep35275.
25. CBOL Plant Working Group CPW, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, et al. A DNA barcode for land plants. Proc Natl Acad Sci U S A.

725 2009;106:12794–7. doi:10.1073/pnas.0905845106.
726 26. Illuminate Biodiversity - International Barcode of Life. https://ibol.org/. Accessed 7 Nov 2019.
727 27. Horton TR, Bruns TD. The molecular revolution in ectomycorrhizal ecology: Peeking into the
728 black-box. 2001.
729 28. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal
730 catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551:457–63.
731 doi:10.1038/nature24621.
732 29. Ratnasingham S, Hebert PDN. A DNA-Based Registry for All Animal Species: The Barcode Index
733 Number (BIN) System. PLoS One. 2013;8.
734 30. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation
735 biodiversity assessment using DNA metabarcoding. Mol Ecol. 2012;21:2045–50.
736 31. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. Environmental DNA. Mol Ecol.
737 2012;21:1789–93.
738 32. Leray M, Knowlton N. DNA barcoding and metabarcoding of standardized samples reveal patterns
739 of marine benthic diversity. Proc Natl Acad Sci. 2015;112:2076–81.
740 doi:10.1073/PNAS.1424997112.
741 33. Beentjes KK, Speksnijder AGCL, Schilthuizen M, Hoogeveen M, Pastoor R, van der Hoorn BB.
742 Increased performance of DNA metabarcoding of macroinvertebrates by taxonomic sorting. PLoS
743 One. 2019;14:e0226527. doi:10.1371/journal.pone.0226527.
744 34. Sigsgaard EE, Nielsen IB, Carl H, Krag MA, Knudsen SW, Xing Y, et al. Seawater environmental
745 DNA reflects seasonality of a coastal fish community. Mar Biol. 2017.
746 35. Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, et al. Real-time
747 DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity
748 assessments and local capacity building. Gigascience. 2018;7:1–14.
749 36. Maestri, Cosentino, Paterno, Freitag, Garces, Marcolungo, et al. A Rapid and Accurate MinION-
750 Based Workflow for Tracking Species Biodiversity in the Field. Genes (Basel). 2019;10:468.
751 37. Oxford Nanopore Technologies. https://nanoporetech.com/. Accessed 24 Dec 2019.
752 38. Wilson BD, Eisenstein M, Soh HT. High-Fidelity Nanopore Sequencing of Ultra-Short DNA

753 Targets. Cite This Anal Chem. 2019;91:6789. doi:10.1021/acs.analchem.9b00856.
754 39. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A
755 comprehensive examination of Nanopore native RNA sequencing for characterization of complex
756 transcriptomes. Nat Commun. 2019;10:1–14. doi:10.1038/s41467-019-11272-z.
757 40. Ramgren AC, Newhall HS, James KE. DNA barcoding and metabarcoding with the Oxford
758 Nanopore MinION. Genome. 2015;58:268.
759 41. Magoč T, Magoč M, Salzberg SL. FLASH: fast length adjustment of short reads to improve
760 genome assemblies. 2011;27:2957–63. doi:10.1093/bioinformatics/btr507.
761 42. Community - Protocol - Guppy. https://community.nanoporetech.com/protocols/Guppy-
762 protocol/v/GPB_2003_v1_revS_14Dec2018. Accessed 6 Jul 2020.
763 43. GitHub - nanoporetech/bonito: Convolutional Basecaller for Oxford Nanopore Reads.
764 https://github.com/nanoporetech/bonito. Accessed 6 Jul 2020.
765 44. GitHub - nanoporetech/rerio: Research release basecalling models and configurations.
766 https://github.com/nanoporetech/rerio. Accessed 6 Jul 2020.
767 45. Madden T, Coulouris G. BLAST Command Line Applications User Manual.
768 46. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and
769 processing long-read sequencing data. Bioinformatics. 2018;34:2666–9.
770 doi:10.1093/bioinformatics/bty149.
771 47. GitHub - wdecoster/nanolyse: Remove lambda phage reads from a fastq file.
772 https://github.com/wdecoster/nanolyse. Accessed 7 Jul 2020.
773 48. Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, et al. Nanopore
774 sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity
775 assessments with high phylogenetic resolution across broad taxonomic scale. Gigascience. 2019;8.
776 49. naturalis/galaxy-tool-add-metadata-otutable. https://github.com/naturalis/galaxy-tool-add-
777 metadata-otutable. Accessed 24 Dec 2019.
778 50. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: A versatile open source tool for
779 metagenomics. PeerJ. 2016;2016.
780 51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol

781 Biol. 1990;215:403–10.
782 52. Voorhuijzen-Harink MM, Hagelaar R, Van Dijk JP, Prins TW, Kok EJ, Staats M. Toward on-site
783 food authentication using nanopore sequencing. 2019. doi:10.1016/j.fochx.2019.100035.
784 53. Holman LE, de Bruyn M, Creer S, Carvalho G, Robidart J, Rius M. Detection of introduced and
785 resident marine species using environmental DNA metabarcoding of sediment and water. Sci Rep.
786 2019.
787 54. Elbrecht V, Leese F. Validation and development of COI metabarcoding primers for freshwater
788 macroinvertebrate bioassessment. Front Environ Sci. 2017;5 APR.
789 55. Turner CR, Uy KL, Everhart RC. Fish environmental DNA is more concentrated in aquatic
790 sediments than surface water. Biol Conserv. 2015;183:93–102.
791

Appendix

The appendix is available on GitHub with the following link:

https://github.com/naturalis/minion-pipeline-tooling/tree/master/appendix

The datasets are available on Zenodo under doi:10.5281/zenodo.3992310

Figures

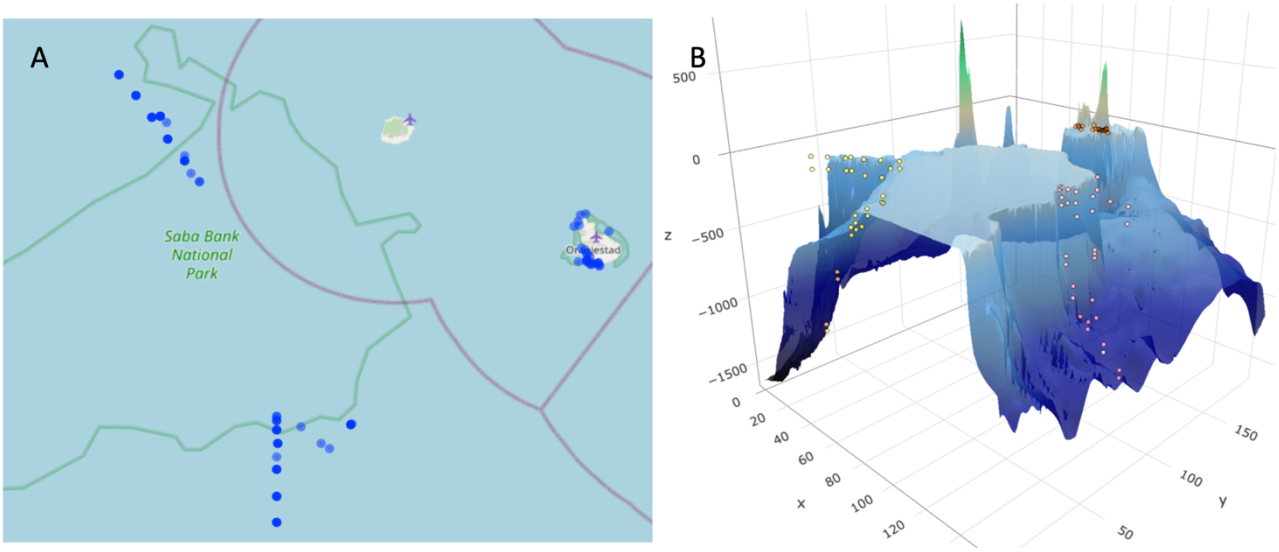
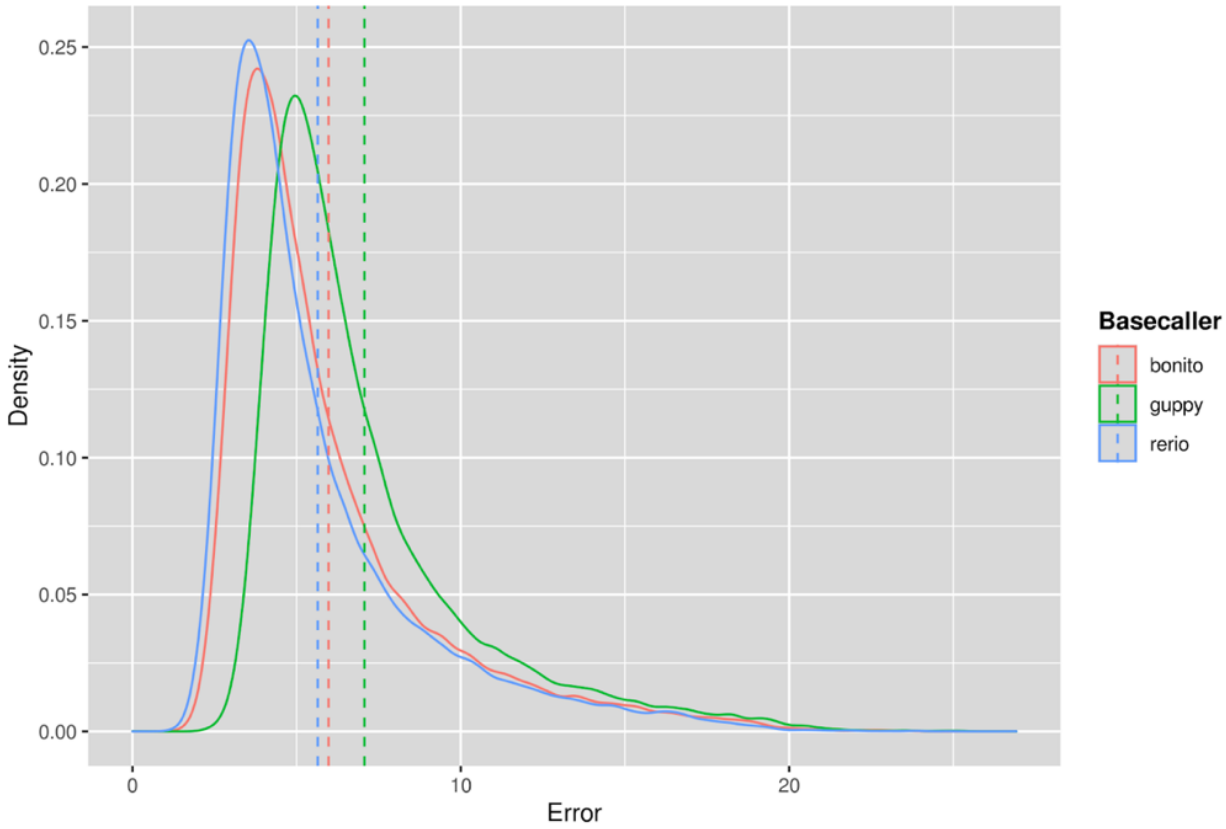


Figure 1. Sampling points around the Saba Bank and St. Eustatius. A: 2D map of sampling points. B: Bathymetric map showing a 3D representation of sampling points.



799

800 **Figure 2**. Comparison of error rates of alignment to the E. coli lambda genome of three base

801 callers, with the dotted line the mean error rate of the corresponding base caller.

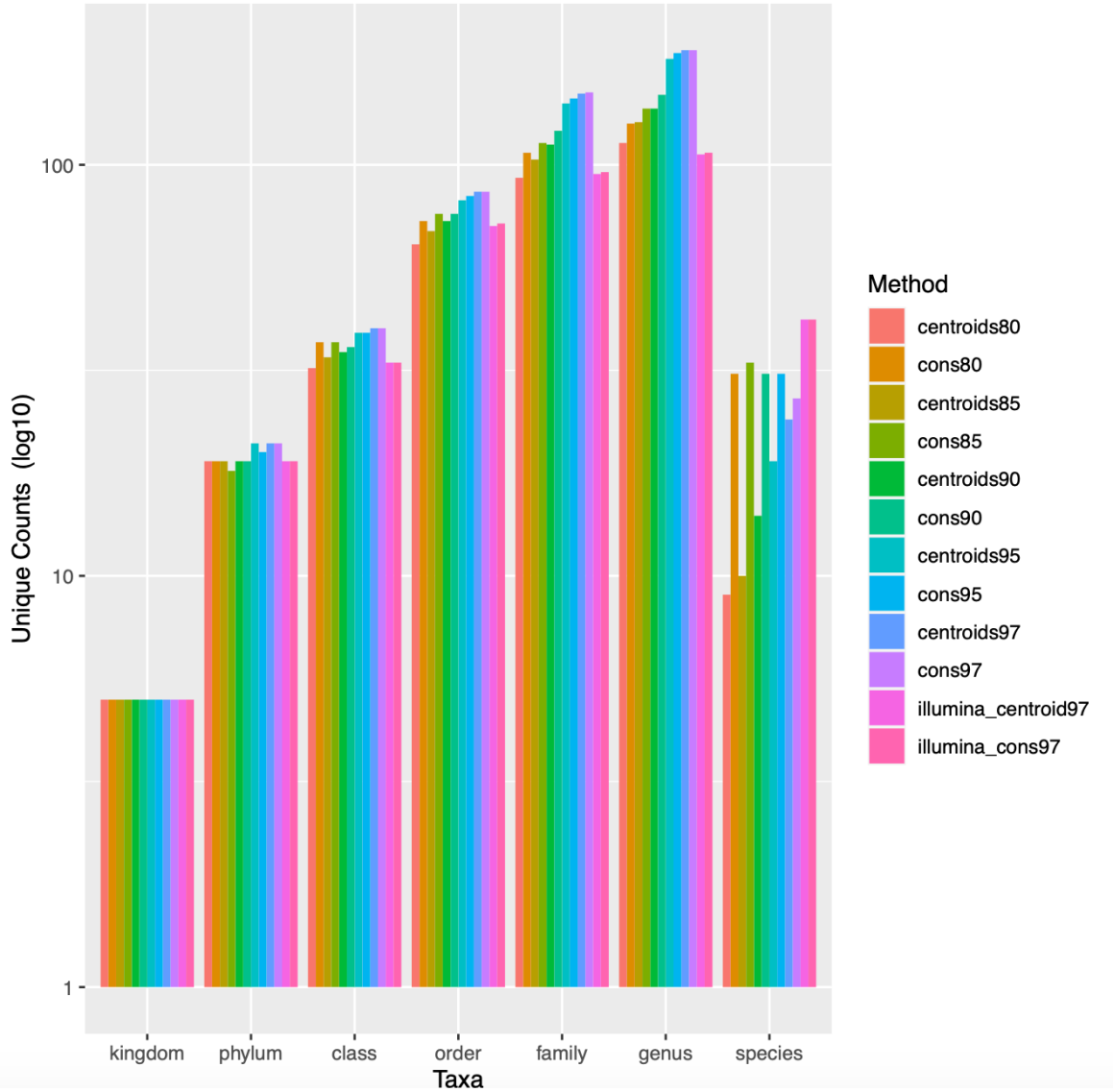


Figure 3. Comparison of the number of identifications after LCA analysis between clustering methods. If identification appeared more than once in the results, it was counted once.

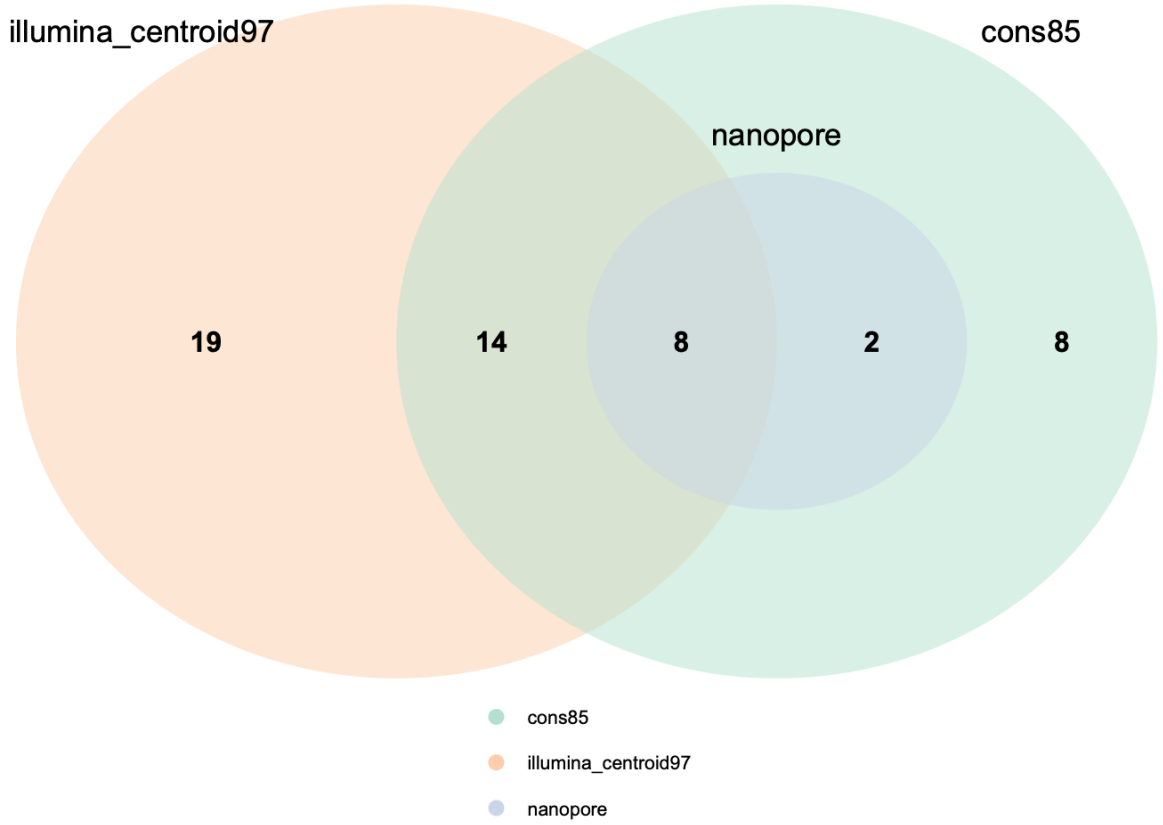
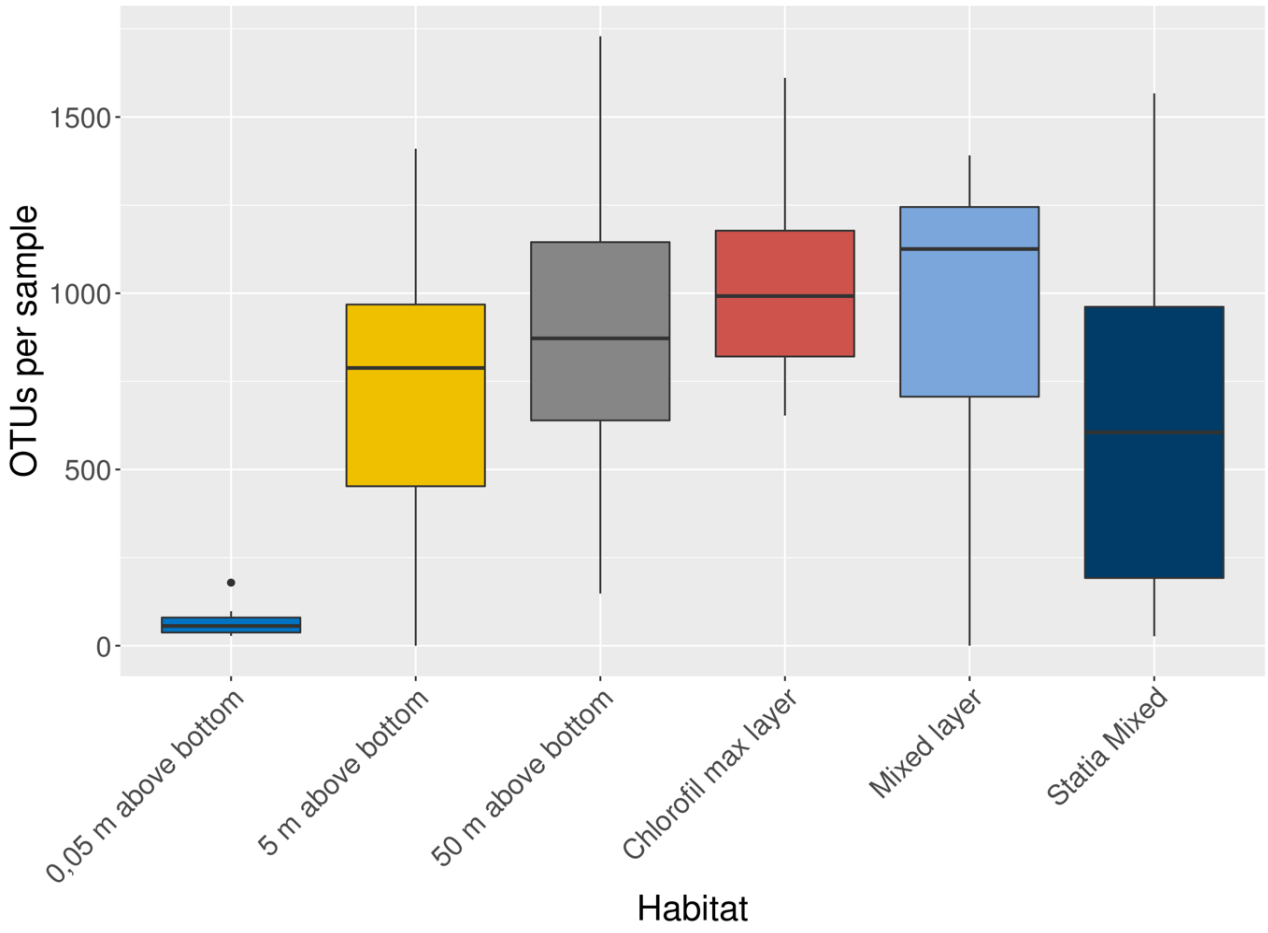


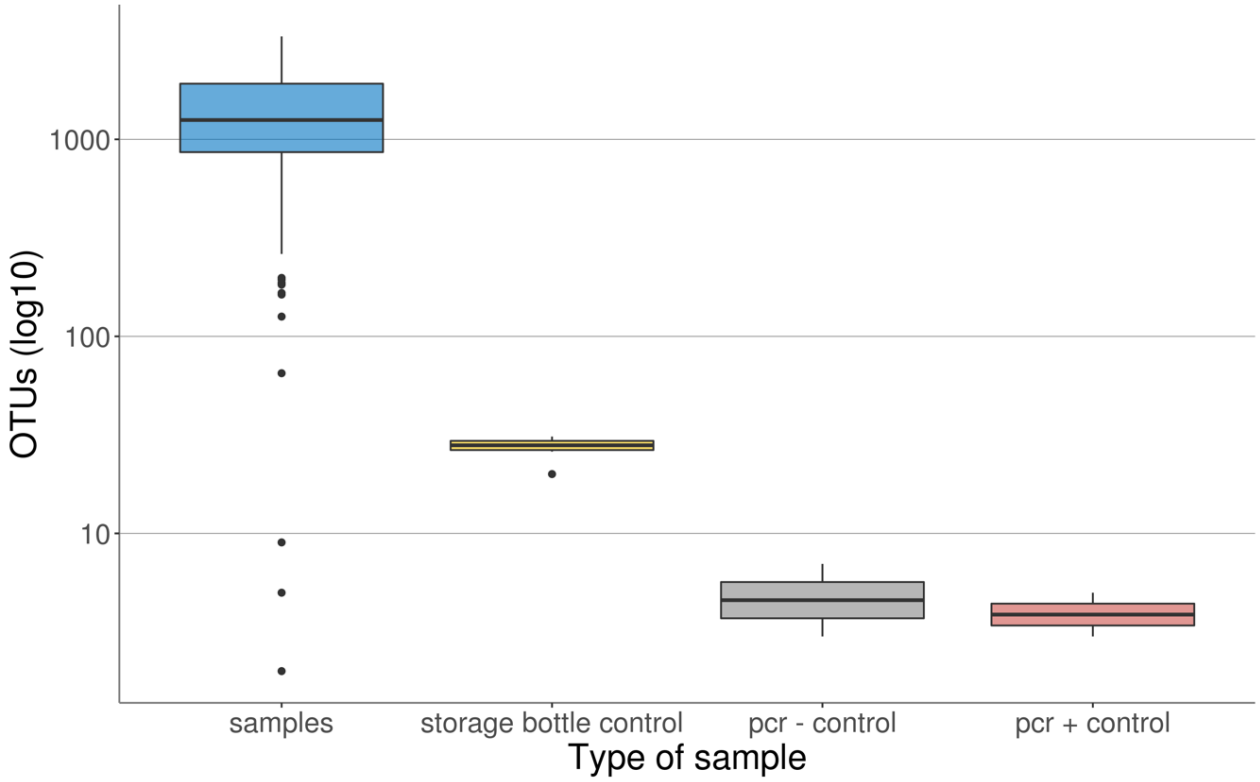
Figure 4. Venn diagram of overlap between unique identifications at the species level of clustering consensus at 85%, the unclustered nanpore data, and the centoroids of Illlumina data clustered at 97%. Only the unique identifications were taken into account for the Venn diagram, meaning that if a species was found twice by a method, it was considered as one identification in the construction of the diagram.



811

812 **Figure 5**. Boxplot of mean OTUs per sample by habitat. All Statia data was binned into "Statia

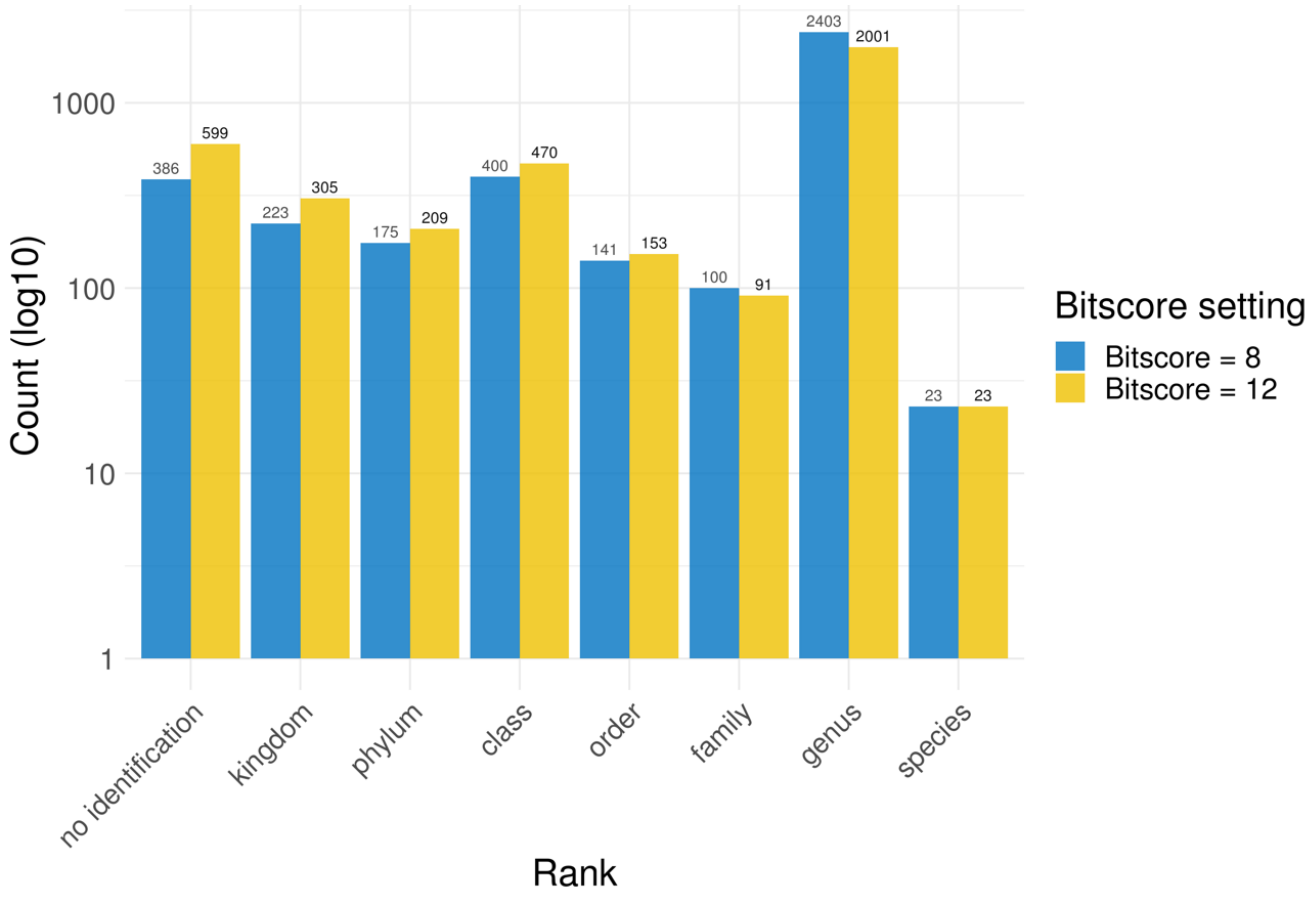
813 mixed".



814

815 **Figure 6.** Boxplot of number of OTUs per sample type, after filtering on minimal abundance.

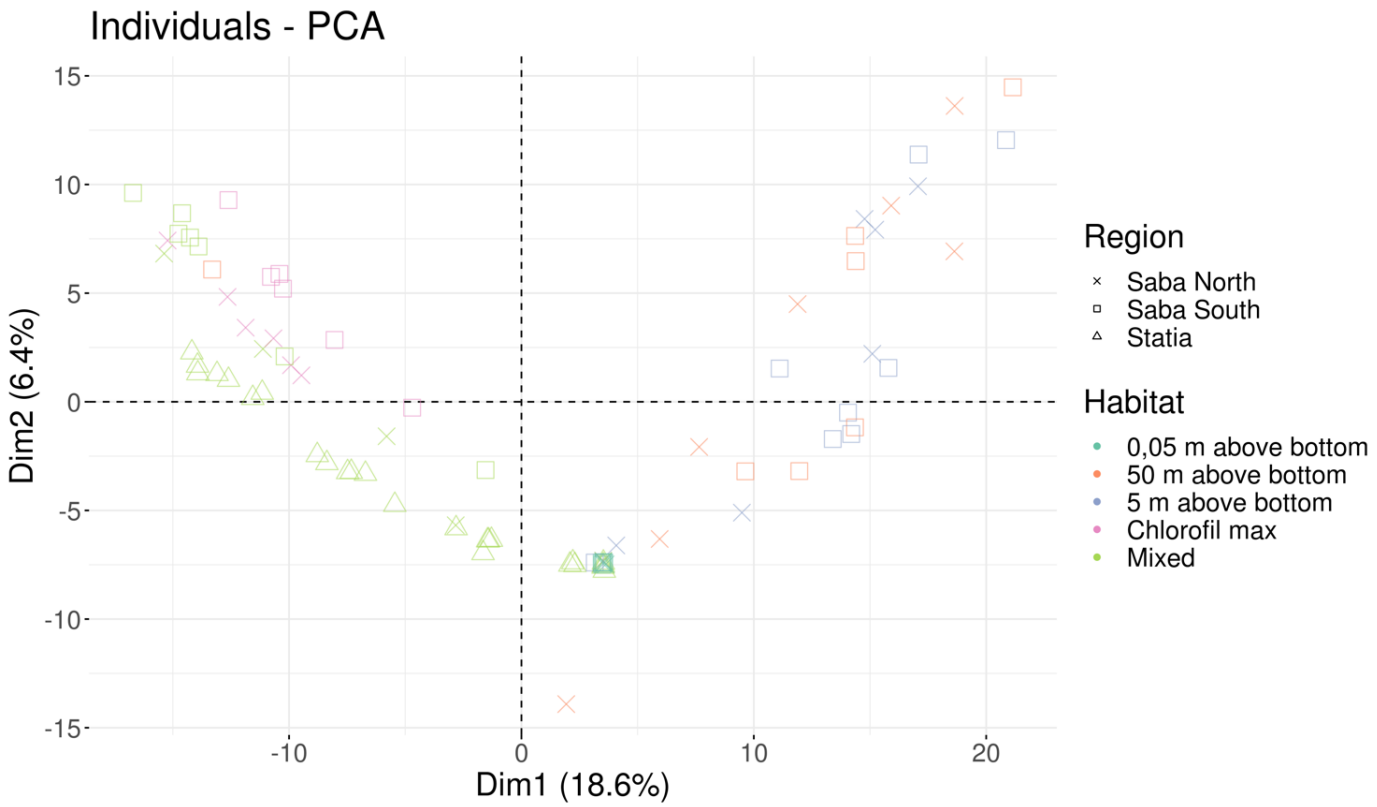
816 the y axis in log scale.



817

818 **Figure 7.** Bar plot of the number of LCA identifications per rank by bitscore setting. OTUs that could

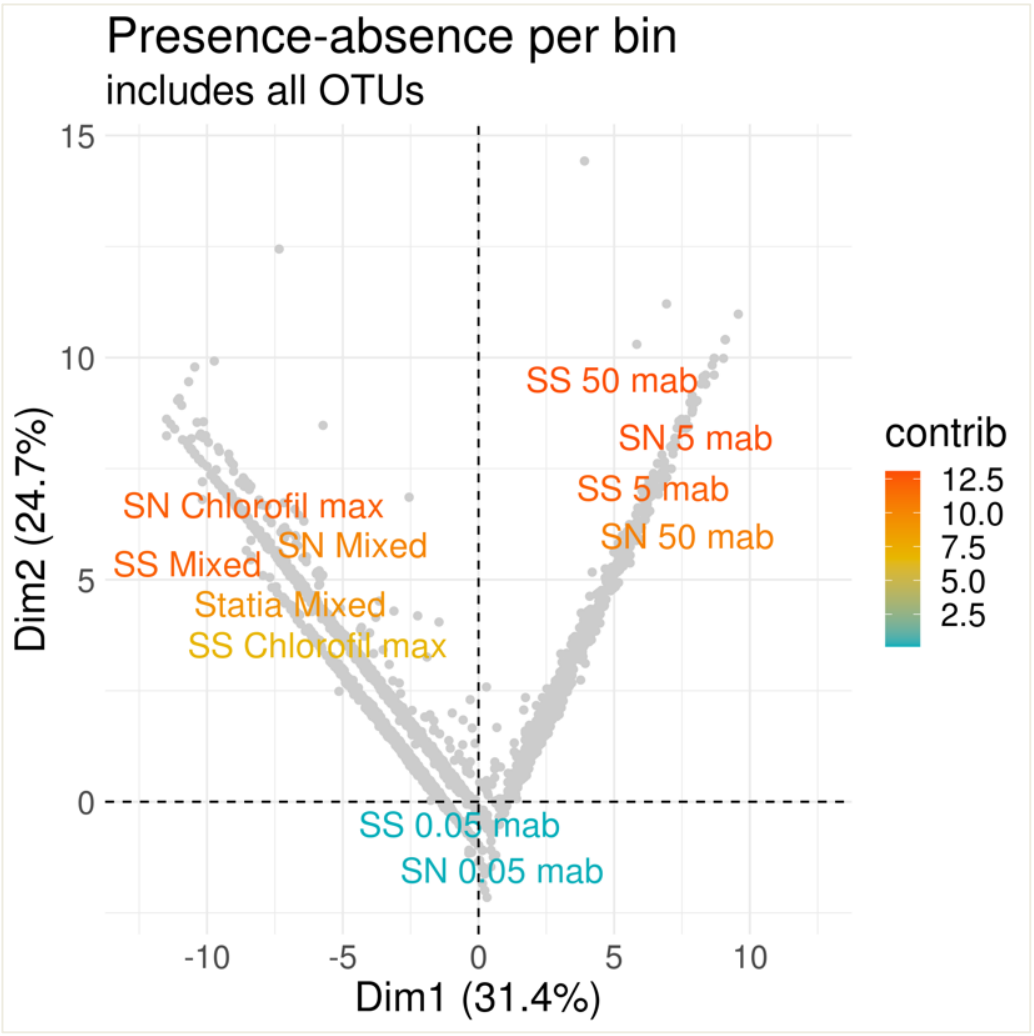
819 not be assigned an LCA were denoted “no identification”.



820

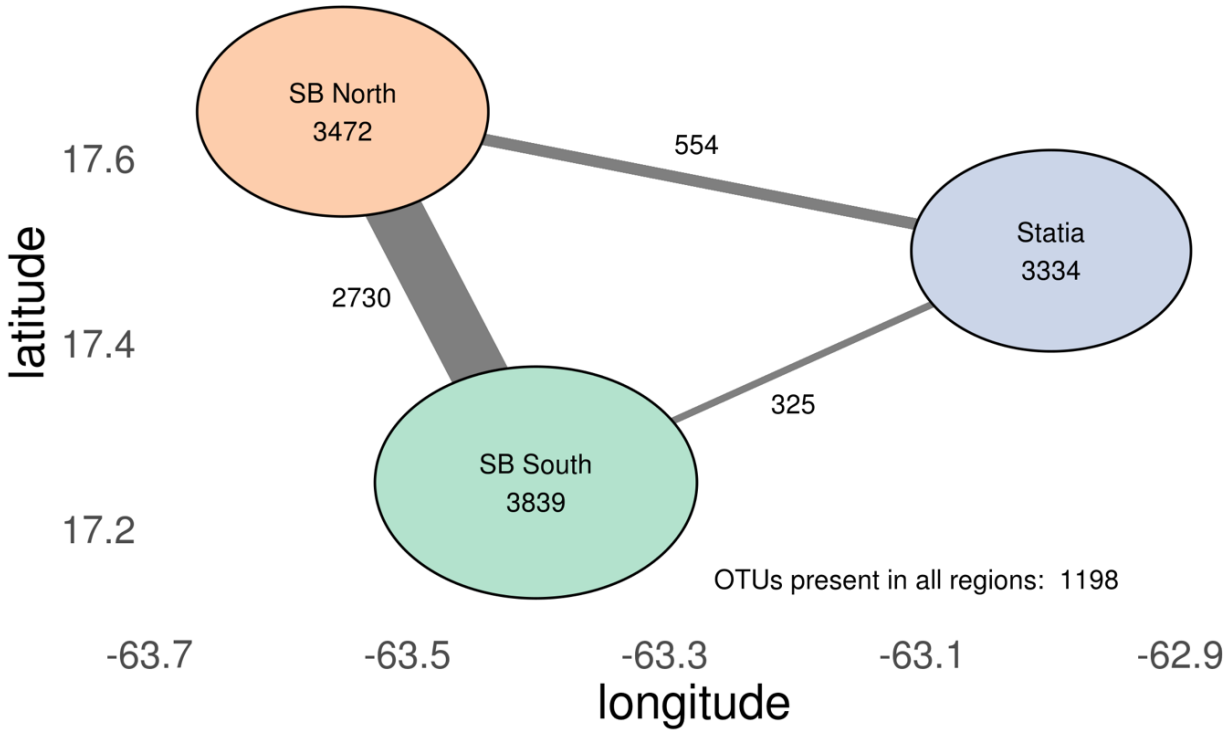
821 **Figure 8.** PCA of all samples. The samples are colored by their corresponding habitat. The shape of

822 the point by determined by region.



823

824 **Figure 9.** PCA analysis of OTUs and their presence in samples. Mab = meter above bottom.



825

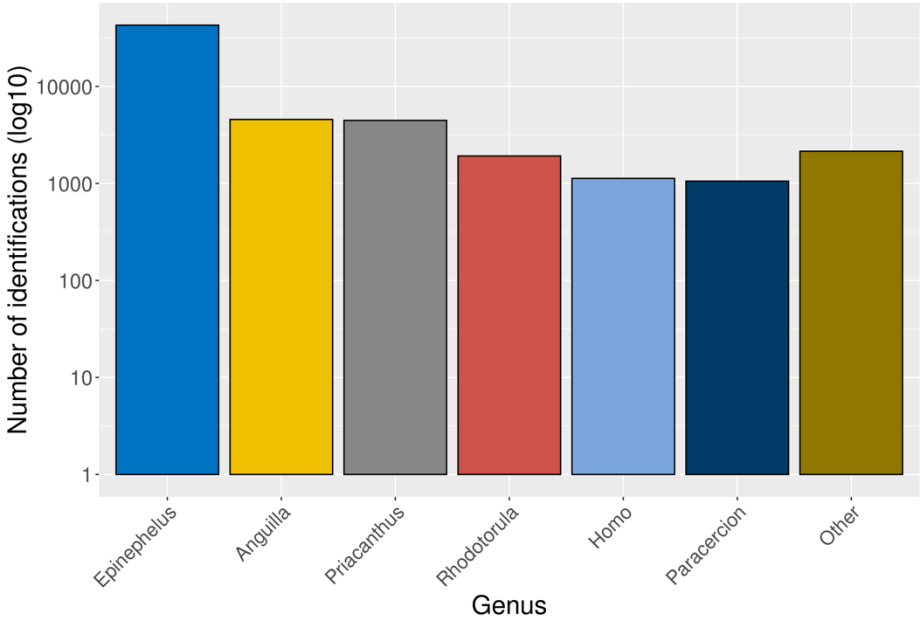
826 **Figure 10.** Shared OTUs between regions, the number of OTUs are relative to the number of
827 UTUs found in Statia. The circles show the number of OTUs that are unique to that region. The
828 lines between the regions denote the degree of similarity in OTU composition.

Tables

Table 1. Example of an LCA analysis of an OTU with two blast hits. This LCA analysis below will return the family "Euphausiidae" for the corresponding OTU.

Kingdom	**"Animalia"**	**"Animalia"**
Phylum:	"Arthropoda"	"Arthropoda"
Class:	"Malacostraca"	"Malacostraca"
Order:	"Euphausiacea"	"Euphausiacea"
Family:	"Euphausiidae"	"Euphausiidae"
Genus:	"Nematoscelis"	"Stylocheiron"

Supplementary Figures



Supplementary Figure 1. Genera represented in strorage bottles. The genus Epinephelus is highly overrepresented.