Beyond taxonomy: Validating functional inference approaches in the context of fish-farm impact assessments

Olivier Laroche[1], Xavier Pochon[2], Susanna Wood[3], and Nigel Keeley[1]

[1]Institute of Marine Research
[2]Cawthron Institute
[3]Cawthron

October 12, 2020

Abstract

Characterization of microbial assemblages via environmental DNA metabarcoding is increasingly being used in routine monitoring programs due to its sensitivity and cost-effectiveness. Several programs have been developed recently which infer functional profiles from 16S rRNA gene data using hidden-state prediction (HSP) algorithms. These might offer an economic and scalable alter-native to shotgun metagenomics. To date, HSP-based methods have seen limited use for benthic marine surveys and their performance in these environments remains unevaluated. In this study, 16S rRNA metabarcoding was applied to sediment samples collected at 0 and [?] 1200 m from Norwegian salmon farms, and three metabolic inference approaches (PAPRICA, PICRUSt2 and TAX4FUN2) evaluated against metagenomics and environmental data. While metabarcoding and metagenomics recovered a comparable functional diversity, the taxonomic composition differed be-tween approaches, with genera richness up to 20× higher for metabarcoding. Comparisons between the sensitivity (highest true positive rates) and specificity (lowest true negative rates) of HSP-based programs in detecting functions found in metagenomics data ranged, respectively, from 0.52 and 0.60 to 0.76 and 0.79. However, little correlation was observed between the relative abundance of their specific functions. Functional beta-diversity of HSP-based data was strongly associated with that of metagenomics (r [?] 0.86 for PAPRICA and TAX4FUN2) and responded similarly to the impact of fish farm activities. Our results demonstrate that although HSP-based metabarcoding approaches provide a slightly different functional profile than metagenomics, partly due to recovering a distinct community, they represent a cost-effective and valuable tool for characterizing and assessing the effects of fish farming on benthic ecosystems.

1 | Introduction

Aquatic biomonitoring has drastically changed in the last decade with the advent of High Throughput-Sequencing (HTS) and the substantial cost reduction of sequencing (Thomsen & Willerslev, 2015; Valentini et al., 2016; Deiner et al., 2017; Lobo et al., 2017; Ruppert, Kline & Rahman, 2019; Wang et al., 2019). Environmental DNA (eDNA) amplicon-based metabarcoding is increasingly being used for rapid and cost-effective community characterization, and is now often promoted and is being gradually integrated into routine monitoring programs (Leese et al., 2016; Danovaro et al., 2016; Pawlowski et al., 2018; Pilliod et al., 2019).

Metagenomics, herein defined as the study of the entire genetic material recovered from environmental samples, has also received considerable attention in the last few years (Bourlat et al., 2013; Grossart et al., 2020). While amplicon-based eDNA metabarcoding provides information on which organisms are present, metagenomics enables insights into the functions they possess. This is particularly relevant when microorganisms

are being used as indicator organisms. For organisms such as bacteria, little is known about the ecology of the vast majority of species. As such, it is their metabolic capability rather than their identity that is usually of greatest interest, and likely to provide more information about current environmental conditions. Taxonomic and functional profiles can respond differently to biogeography, abiotic environmental variables (e.g. organic content, metal concentration) and community processes and interactions. As such, they can exhibit different level of stochasticity and temporality, and provide complementary information that may increase our understanding of the mechanisms behind community turnover (Barberan et al., 2012; Louca, Parfrey & Doebeli, 2016; Hornick & Buschmann, 2018; Cordier et al., 2020). Having both taxonomic and functional information also enables the computation of functional redundancy within the community, which may also help assess resilience (Escalas et al., 2019).

While metagenomics may eventually replace metabarcoding for whole community assessment, it's mainstream use is still limited by its relatively low sample throughput and cost-efficiency, and heavy computational and data management requirements (Bowman & Ducklow, 2015; Nagpal, Haque & Mande, 2016; Breitwieser, Lu & Salzberg, 2018). To circumvent these issues, several programs have been developed to infer metabolic profiles from eDNA 16S ribosomal RNA (rRNA) gene data (e.g. Functional Annotation of Prokaryotic Taxa [Faprotax] (Louca, Parfrey & Doebeli, 2016), Piphillin (Iwai et al., 2016), Vikodak (Nagpal, Haque & Mande, 2016), Prediction by phylogenetIC plAcement [Paprica] (Bowman & Ducklow, 2015), Phylogenetic Investigation of Communities by Reconstruction of Unobserved States [Picrust2; Douglas et al., 2019) and Tax4Fun2 (Wemheuer et al., 2018). The last three methods use a hidden-state prediction (HSP) approach (Zaneveld & Thurber, 2014) where genomic content is inferred according to the position of the genome in a reference phylogenetic tree. These methods have been shown to provide functional profiles that correlate, with a varying degree of success, to metagenomics and metabolomics profiles (e.g. Bowman & Ducklow, 2015; Douglas et al., 2019; Sun, Jones, & Fodor, 2020; Wemheuer et al., 2018). While not as accurate as metagenomics functional analyses, these HSP methods can provide more complete metabolic profiles as they do not require high sequencing depth to assign functions (e.g. pathways of rare taxa can still be assigned; Langille et al., 2013), and may be useful in situations where metagenomics would be prohibitively expensive, such as in broad microbial routine monitoring surveys.

Metabolic inference methods have been evaluated and used in a large variety of studies, for example in clinical trials (Millares et al., 2015), oyster aquaculture (Arfken et al., 2017), aquatic urban systems (Wang et al., 2018), acid mine drainages (Aguinaga et al., 2018), subterranean estuaries (Hong et al., 2019), gut microbiota (Pacheco-Sandoval et al., 2019), marine biofilms (Salerno et al., 2018), and coral reef associated microbiomes (Pearman et al., 2019). However, there has been limited use in the context of benthic marine monitoring surveys (Hornick & Buschmann, 2018; Laroche et al., 2018; Cordier, 2020). Before this application can be routinely applied in such situation, it is essential to evaluate its performance. In particular, the accuracy of functional inference methods is strongly influenced by the completeness of the reference databases and by the genetic plasticity of some taxonomic groups (Bowman & Ducklow, 2015). For example, certain functions, especially those involving few genes, tend to occur at shallow phylogenetic depth (Martiny et al., 2015) and can be difficult to correctly predict.

The aim of this study was to evaluate the performance of three metabolic inference methods, Paprica, Picrust2 and Tax4Fun2, against metagenomics and environmental data, in the context of salmon farm benthic surveys. In particular, we aimed to: 1) compare predictions and abundance correlations between 16S rRNA amplicon-based metabarcoding functional inference approaches (hereafter referred to as HSP methods) and shotgun metagenomics functional profiling, 2) contrast the taxonomic and functional microbial diversity recovered from metabarcoding and metagenomics, and 3) assess and compare the sensitivity of functional communities derived from HSP methods and metagenomics with respect to microbial turnover and in correlation with environmental data.

2 | Materials and methods

2.1 | Sample collection

Benthic sediment samples (depths of 32-85 m) were collected at four large scale Atlantic salmon (*Salmo salar*) farms, two located in a semi-exposed coastal region of mid-Norway (farms: FRØ and SMØ) and two inside fjords in northern Norway (farms: NOR, STO) (Fig. 1). At FRØ, six samples were collected: three biological replicates located next to the pen (0 m), and three reference (control) replicates located 1,200 m away from the fish pens. The objective of this sampling design was to compare bacterial communities between impacted and non-impacted sediments. To test HSP methods across different geographic settings within Norway, samples were also collected next to the pen at the NOR and STO farms (one each) in northern Norway, and two samples at SMØ (Southern-Norway), one next to the pen and one located at a reference site located 7,920 m away from the farm. Approximately 5 g of the top (1 cm) sediment layer, collected with a van Veen grab sampler (surface area 0.1 m^2), was subsampled per grab with a sterile spatula and stored at -20°C until further processing.

2.2 | Physico-chemical and macrofaunal analyses

A complimentary suite of environmental parameters was obtained from parallel studies. The prevalence of three terrestrial fatty acids (oleic acid, 18:1n-9; linoleic acid, 18:2n-6; and α-linolenic acid, 18:3n-3) in the sediments, which indicate fish-feed-derived organic matter (White et al., 2017), were assessed by Folch lipid extraction and direct methylation as described in Woodcock et al. (2019). The organic and inorganic carbon content of the sediment was determined by drying the sediment at 40°C for 48 h followed by combustion at 450°C for 2 h (LOI450). Measures of benthic respiratory fluxes, including ammonium (NH_4), total carbon dioxide (TCO_2), and oxygen (O_2) were obtained from, and following the methods described in Keeley et al. (2019). Additionally, the macrofaunal communities were obtained from and characterized using the methods described in Keeley et al. (2019) and in Keeley et al. (*In prep* .).

2.3 | DNA extraction

Sediment samples were homogenized with a sterile stainless-steel micro spatula (soaked in 50% bleach solution - commercial bleach diluted with double-distilled water [ddH_2O]) for a minimum of 5 min and rinsed with ddH_2O between each sample), subsampled (0.25 g), and processed with the DNeasy PowerSoil kit (QIAGEN, California, USA) following the manufacturer's protocol. DNA purity was measured with a Nanophotometer (Implen, Munich, Germany), integrity assessed on 1.5% agarose gels, and quantity measured with a Qubit® Fluorometer (Life Technologies).

To assess potential cross-contamination, extraction blanks (sediment replaced by ddH_2O) were included. All sample handling and extraction steps took place in a dedicated DNA laboratory that was decontaminated with 50% bleach solution prior to DNA extraction.

2.4 | Targeted 16S rRNA library preparation

DNA extracts were set at equimolar concentration (5 ng/μL) with ddH_2O for a total volume of 25 μL per sample, stabilized with DNAstable (Biomatrica, California, USA) following the manufacturer's protocol, and dry-shipped to the Cawthron Institute (Nelson, NZ) for 16S rRNA library preparation. Upon arrival, DNA extracts were rehydrated with 25 μL of ddH_2O, and a segment of the V3-V4 region of the 16S ribosomal RNA gene (approximately 450 base pairs [bp]) was PCR amplified using the forward S-D-Bact-0341-b-S-17: 5'-CCT ACG GGN GGC WGC AG-3' and reverse S-D-Bact-0785-a-A-21: 5'- GAC TAC HVG GGT ATC TAA TCC-3' primers from Klindworth et al. (2013), modified to include Illumina™ overhang adaptors.

PCR reactions consisted of 22 μL of AmpliTaq Gold ? 360 PCR Master Mix (Life Technologies), 8 μL of ddH_2O, 1 μL of each primer (10 μM, IDT, Iowa, USA), 5 μL of GC enhancer (Life Technologies), and 2 μL of

template DNA (5 ng/μL). The reaction cycling conditions were 94°C for 3 min, followed by 30 cycles of 94°C for 30 s, 52°C for 30 s, 72°C for 1 min, with a final extension step at 72°C for 5 min. Each PCR included a negative control (no template sample) to ensure the absence of cross-contamination.

Amplicon purification and normalization (2 ng/uL) were performed with the SequalPrep^{TM} Normalization plates (Invitrogen, California, USA) following the manufacturer's instructions, and submitted to New Zealand Genomics Ltd (Auckland, New Zealand) for indexing with the Nextera^{TM} DNA library Prep Kit (Illumina, California, USA), pooling and paired-end (2 × 250 bp) sequencing on a Illumina^{TM} MiSeq. One blank sample (ddH_2O) was included prior to indexing and sequencing to control for potential contamination. Sequences are available from the NCBI Sequence Read Archive (SRA) under project number PRJNA661323.

2.5 | Metagenomics library preparation

DNA extracts were set at equimolar concentration (8 ng/μL) with 10 Mm Tris and sent on dry ice to the Norwegian Sequencing Center (NSC, Oslo) for library preparation with the Nextera^{TM} DNA Flex Tagmentation kit (Illumina, California, USA) following the manufacturer's protocol. After indexing, samples were pooled and sequenced on $\frac{1}{2}$ Novaseq SP flow cell (Illumina, California, USA) with a 2 x 150 bp paired-end protocol. One blank sample (ddH_2O) was included prior to indexing and sequencing to assess potential contamination. Sequences are available from the NCBI Sequence Read Archive (SRA) under project number PRJNA661323.

2.6 | Bioinformatic analysis of 16S rRNA data

Primers from the demultiplexed fastq files were removed with cutadapt (version 2.6; Martin, 2011) and reads quality filtered, denoised, merged and chimera filtered with the DADA2 R program (version 1.14; Callahan et al., 2016). Prior to quality filtering, reads were truncated at 226 and 220 bp on the 5' end to remove the lower quality section and reduce the number of reads lost during quality trimming. Quality filtering and denoising were performed using the default parameters, and merged using a perfect minimum overlap of 10 bp. Chimera removal was performed using the consensus method where sequences found to be chimeric in the majority of samples (default value = 90 %) are discarded. Sequences found in negative controls, including DNA extraction, PCR, indexing and sequencing blanks were examined and subsequently removed from across all samples. Sequences unclassified at kingdom level or not identified as bacteria were discarded. Additionally, rare amplicon sequence variants (ASVs; less than 10 reads across the entire study) were removed from the dataset. The resulting sequences were used for taxonomic and functional profiling using three pipelines: Paprica (version 0.5.2; Bowman & Ducklow, 2015), Picrust2 (version 2.2.0_b; Douglas et al., 2019) and Tax4Fun2 (version 1.1.4; Wemheuer et al., 2018). These methods were chosen for their popularity in the scientific literature, their compatibility with large datasets, and their reliance on KO (KEGG ortholog numbers) and EC (Enzyme commission numbers), which can be easily assessed against metagenomics results obtained from Humann2 (Franzosa et al., 2018), our chosen functional profiling methodology. Prior to the taxonomic assignment and metabolic inference, sequencing depth per sample was visualized with the "rarecurve" function of the "vegan" R package (version 2.5.6; Oksanen et al., 2019) to ensure that all samples had sufficient sequencing depth to recover most of the diversity (Fig. S1). The default parameters implemented within each method were used to keep the analysis as simple as possible.

2.7 | Bioinformatic analysis of metagenomics data

The metagenomics pipeline used is based on the fully automated workflow of the Humann2 software (version 2.8.2; Franzosa et al., 2018). Reads were first pre-processed with KneadData (version 0.7.4; https://bitbucket.org/biobakery/kneaddatae) to perform both quality filtering with Trimmomatic (version 0.39; Bolger, Lohse, & Usadel, 2014), and screening of undesired reads (herein phix, human and salmon DNA) with Bowtie2 (version 2.3.5.1; Langmead & Salzberg, 2012). Profiling of taxa was performed with MetaPhlAn2 (version 2.7.8; Truong et al., 2015) and results used to construct a sample-specific database from functionally

annotated pangenomes (referred to as the ChocoPhlAn database). Humann2 then performed a nucleotide-level mapping with Bowtie2 of all reads against the custom database, followed by a translated search with Diamond (version 0.8.36.98; Buchfink, Xie, & Huson, 2015) for those that did not align. Reads that remained unaligned were subjected to an additional translated search against the UniRef50 protein database (Suzek et al., 2015). The gene family abundance table was then converted to a KO and EC tables using the "Humann2_regroup_table" function and the uniref50_ko and uniref50_level4ec groups, respectively.

2.8 | Data analysis and statistics

Taxonomic and functional richness differences between metabarcoding and metagenomics data were visualized with box and bar plots using the "ggplot2" R package (version 3.3.1; Wickham, 2016).

Using presence/absence data, the prediction of pathways from HSP methods (herein Paprica, Picrust2 and Tax4Fun2) was tested against the metagenomics profiles using the "caret" R package (version 6.0.86; Kuhn, 2020) and the "confusionMatrix" function, and visualized with the "alluvial" R package (version 0.1.2; Bojanowski & Edwards, 2016).

Spearman correlations of functions across all samples and per samples between the HSP methods and the metagenomics data were assessed using the "stats" R package (version 3.6.1; R Core Team, n.d.) and the "cor" function. For this analysis, EC and KO abundances were transformed to the centered-log ratio with the "clr" function of the "composition" R package (version 1.40.4; van den Boogaart, Tolosana-Delgado, & Bren, 2020) and only functions shared between inference methods and metagenomics were maintained. Because several ECs and KOs co-occur within pathways and genomes, correlation of functions between HSP and metagenomic samples can be naturally high, even between completely unrelated samples (Douglas et al., 2019; Sun, Jones & Fodor, 2020). To take this dependency into account, we added a randomized dataset referred to as "null expectation" for each HSP method. Based on the original ASV abundance table, the permaswap function of the "vegan" R package (parameters: times = 1, burnin = 20000, thin = 500, mtype = "count", shuffle = "both") was first used to create a dataframe with permuted samples and ASVs. This table was then inputted in to each HSP method to obtain "null expectation" datasets of functional inferences. Differences between results of the actual and "null expectation" data were tested with Welch two sample t-tests.

Correspondence of ASV and functional communities between HSP methods with metagenomics was assessed with a procrustes test using the "protest" function (symmetric analysis with 9,999 permutations and scores = "sites") of the "vegan" R package (version 2.5.6; Oksanen et al., 2019). The correspondence of the ASV and functional communities derived from HSP and metagenomics with macrofaunal communities (transformed with the Wisconsin method implemented in "vegan") as well as physico-chemical data (scaled with the rda function of the "vegan") was also assessed with procrustes tests using the same parameters. Physico-chemical data were only completely available for the FRØ locality, therefore only samples from this site were kept for the latter analysis. In addition to the procrustes tests, the sensitivity of each dataset towards fish farming was assessed using permutational analyses of variance (PERMANOVA; Anderson, 2005) between samples collected at the 0 m and [?] 1200 m from the pen. For the procrustes and PERMANOVA analyses, two methodologies were evaluated: 1) Euclidean distance matrices computed from centered-log ratio transformed abundances, as suggested in Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue (2017) for compositional data, and 2) Jaccard distance matrices for presence/absence data. Multivariate homogeneity of groups dispersions analyses between distance categories (0 m and [?] 1200 m) were performed with the "betadisper" function of the "vegan" R package.

Compared to reference sites, benthic environments in proximity to fish farm activities are typically characterized by higher concentrations of organic matter and nutrients, especially phosphorus and nitrogen (Buschmann et al., 2006; Wang et al., 2012), which can lead to eutrophic conditions and anaerobic microbial degradation (e.g. sulphate reduction and methanogenesis; Valdemarsen, Kristensen & Holmer, 2009). To explore this, the response of pathways associated to the nitrogen and sulfur cycle between near and far-field

sites, and between metagenomics and HSP-based pathways profiles were compared. Pathways were clustered to their parent class based on the MetaCyc database (https://metacyc.org/) and only classes involved in nitrogen (fixation, ammonification, nitrification, denitrification and degradation) and sulfur cycle (oxidation, reduction and degradation) were maintained. Groups of pathways that associated to more than one parent class were filtered out. The response of the remaining classes between pen and reference sites was analyzed using centered-log ratio transformed data and the 'lm' function of the 'stats' R package and visualized with barplots using the 'ggplot2' R package. Only Paprica and Picrust2 provide pathways data derived from the MetaCyc database and are similar to those of Humann2, therefore only these two HSP methods were assessed. In addition, functional groups determined by the Faprotax methodology (version 1.2.1; Louca, Parfrey & Doebeli, 2016), where ecologically relevant groups are assigned to ASVs based on available literature from cultured strains, was performed and differential abundance assessed with Ancom2 (version 2.1; Kaul et al., 2017) between pen and reference sites in order to compare results with metagenomics and HSP data.

3 | Results

3.1 | 16S rRNA sequencing and pre-processing

Excluding the controls, a total of 1,374,973 reads (mean of 122,640 per sample) were sequenced. Quality filtering, denoising, merging and removing chimeric sequences resulted in 36% of reads being discarded (Table S1). Discarding sequences found in the blanks (Table S2) reduced read count by 2.1% and resulted in the loss of 12 ASVs. Removing non-bacterial sequences and those either assigned to chloroplast or unidentified at kingdom level resulted in the loss of 8.5% of the reads. Discarding ASVs with less than 10 reads further decreased the number of ASVs and reads by 44.8% and 3.8%, respectively. Final reads count per sample averaged 65,966 (standard deviation [sd] = 21,526), with ASV richness reaching a plateau for all samples (Fig. S1).

3.2 | Metagenomics sequencing and pre-processing

A total of 314,003,232 reads with a mean of 31,400,323 per sample were obtained from the 1/2 Novaseq SP flow cell (Table S3). Trimming low quality reads reduced their number by 19.3% and removing reads associated to phix and human DNA by 0.6%, and to salmon DNA by another 0.6% (Table S3).

On average, 4,425.6 gene families per sample matched the ChocoPhLAn database after nucleotide alignment with Bowtie2 (Table S4). Translating reads with diamond and using the uniref50 database, an average of 36% of reads could be aligned, with a mean of 1,084,334 gene families identified per sample (Table S4).

3.3 | Metabarcoding versus metagenomics-based functional profiling

In both the 16S rRNA metabarcoding and metagenomics dataset, the two main bacterial Phyla were Proteobacteria and Bacteroidetes (Fig. 2A). The Proteobacteria families Desulfobacteraceae, Psychromonadaceae and Vibrionaceae, and the Bacteroidetes family Flavobacteriaceae were the most abundant taxa. The total number of bacterial families detected was substantially higher for metabarcoding (224) compared to metagenomics (15; Fig. 2B). Subsequently, the mean number of families detected per sample was also considerably higher for metabarcoding (114) versus metagenomics (4.3) (Fig. 2B). Conversely, the overall and mean number of functions per sample, either ECs or KOs, were relatively similar between HSP methods and metagenomics (Fig. 2C).

Differences in the detection of functions (ECs and KOs) by HSP methods and metagenomics were assessed with confusion matrices (Fig. 3). The results indicate that Paprica had the highest sensitivity (highest true positives rate; 0.76), followed by Picrust2 (0.66) and Tax4Fun2 (0.52). Conversely, Paprica showed lowest

specificity (lowest true negative rate) of 0.6, while Picrust2 and Tax4Fun2 showed similar results with 0.78 and 0.79, respectively.

Overall, correlations of the abundance of HSP-derived functions with metagenomics was low, with a mean value of 0.07, 0.2 and 0.16 for Paprica, Picrust2 and Tax4Fun2, respectively, and hardly differed from the null expectation datasets (Fig. 4). At the sample level, Spearman correlations of the HSP datasets with metagenomics was highest for Picrust2 (0.81), followed by Tax4Fun2 (0.67) and Paprica (0.6), but no significant difference with null expectation datasets was observed (Fig. 4 and Table S5).

The correspondence of the ASV (16S rRNA metabarcoding) community with the functional community derived from metagenomics was relatively weak using either centered-log ratio transformation (r = 0.49, p.value = 0.21) or presence/absence (r = 0.4, p.value = 0.44) data (Fig. 5). This contrasts with the strong and significant correspondence of the HSP-derived functional communities with metagenomics, with Paprica and Tax4Fun2 showing highest correlation (r [?] 0.87; Fig. 5). Correspondence of these two HSP methods with metagenomics noticeably improved when using presence/absence data and Jaccard dissimilarity indices (r = 0.92 and 0.04, respectively; Fig. 5).

3.4 | Sensitivity of 16S rRNA metabarcoding and metagenomics

The response of the taxonomic (ASVs) and functional communities toward fish farm activities was tested for each dataset with a PERMANOVA, by comparing the variance between samples collected at the pen (0 m) versus those collected at reference sites ([?] 1200 m). Except for Picrust2, the data transformed to the centered-log ratio showed significant differences between near and far-field samples. In general, highest sensitivity was achieved when using presence/absence data and Jaccard dissimilarity indices, with Paprica and Tax4Fun2 being the most responsive (R^2 = 0.39 and 0.4, respectively; Table 1). While betadisper analysis of homogeneity of groups dispersions showed higher variance in samples collected at the pen, no significant difference was detected between distance groups (Fig. S2 and S3, Table S6).

Correlation of the taxonomic (ASVs) and functional communities with macrofaunal communities and physico-chemical data were explored with Protest analyses. While the taxonomic profile did not significantly correlate with the macrofauna, all functional communities were strongly and significantly correlated (r [?] 0.68), with the strongest associations with Paprica, Tax4Fun2 and metagenomics data (Table 2). Conversely, only the EC-based datasets (Paprica and Humann2 EC) were significantly associated with the physico-chemical data (r [?] 0.96, p.value [?] 0.048), although correlations were relatively strong with all molecular datasets (r [?] 0.5; Table 2).

Pathways of particular interest involved in the nitrogen and sulfur cycle were compared between pen and reference sites and between the metagenomics (Humann2) and 16S rRNA HSP-based data (Fig. 6). Only four out of the nine pathways investigated were found to be affected by fish farm activities within the metagenomics and Picrust2 datasets. These included nitrate-reduction (increased prevalence; +), allantoin-degradation (reduced prevalence; -), sulfur-oxidation (-) and glycosaminoglycan-degradation (+). In comparison, Paprica identified four additional pathways affected by fish farming including nitrogen-fixation (+), sulfur-reduction (-), sulfite and sulfide-reduction (-), and dimethylsulfide-degradation (+). All pathways found to be either positively or negatively affected by fish farm activities in the metagenomes were found to respond similarly in both Paprica and Picrust2 datasets. Since we expected sulfur related pathways of both metagenomics and 16S rRNA HSP-based data to be more strongly affected by fish farm activities, we also assessed functional groups determined by the Faprotax methodology (based on the taxonomic identification of the ASVs). Using Faprotax, functional groups more prominent near the pens included ureolysis, dark hydrogen oxidation, sulfite respiration and nitrogen fixation, while those more abundant at the reference sites included aerobic ammonia oxidation and oxygenic photoautotrophy (Fig. S4).

4 | Discussion

In this study, our main objectives were to assess the level of correspondence between the taxonomic and functional profiles derived from amplicon-based 16S rRNA metabarcoding data and shotgun metagenomics, evaluate the strengths and weaknesses of both approaches, and determine whether functional profiles from HSP methods can be used as a substitute to metagenomics for monitoring functional changes associated with fish farming in marine environments.

The overall taxonomic richness recovered by 16S rRNA metabarcoding was up to 20-fold higher (depending on the taxonomic level) than metagenomics. This is due, in part, to the differences in the ability of identifying sequences of different lengths (˜ 450 bp [metabarcoding] versus ˜ 150 bp [metagenomics]) and by differences in the taxonomic assignment methods used. However, it is also well recognized that metagenomics requires much more sequencing effort than metabarcoding to reach equivalent 16S rRNA coverage as it captures all DNA material (Cottier et al., 2018; Singer, Shekarriz, McCarthy, Fahner, & Hajibabaei, 2020). This is especially problematic in highly diverse communities such as the marine sediment samples assessed in this study. Small differences in taxonomic composition were anticipated because 16S rRNA primer sets are never truly universal (Pollock et al., 2018). While both datasets were relatively similar in terms of dominant Phyla and Families, several taxa such as BD2-2 (Bacteroidetes), Sulfurovaceae, Thermoanaerobaculaceae and Rubritaleaceae were more predominant in the metabarcoding data. These results clearly illustrate the advantage of using targeted amplicon 16S profiling over metagenomics when it comes to providing a comprehensive overview of communities in complex environments, although biases due to preferential PCR amplification and primer specificity are inevitably introduced.

Despite the differences in taxonomic richness, functional richness between the HSP methods and metagenomics was relatively similar. This counter intuitive result occurs because most ECs/KOs are typically redundant across microbial communities (Louca et al., 2018; Starke et al., 2020), many of which performing functions that are essential to cellular activities. Additionally, several non-ubiquitous functions, and especially those occurring at shallow phylogenetic depth, are difficult to accurately predict (Martiny et al., 2015) and may therefore be omitted by HSP methods (Bowman & Ducklow, 2015). These false negatives were prominent for Paprica, which had the lowest specificity. The opposite scenario where functional richness is artificially increased due to false positive predictions is also a possibility. For example, phylogenetic plasticity and genomic variability can result in loss of functions within taxa that can't be detected by HSP methods. However, because of the substantial differences in terms of recovered taxonomic diversity between metagenomics and 16S metabarcoding methods and limits imposed by sequencing depth, it is also possible that some functions predicted by the HSP methods were not detected by metagenomics. As such, the true sensitivity of the HSP methods, which was lowest for Tax4Fun2 and highest for paprica, was likely underestimated.

In general, we observed little correlation and/or no significant difference with the null expectation datasets based on the abundance of functions shared between HSP methods and metagenomics. Other studies have reported weak correlations of HSP derived pathways abundance with metagenomics when compared to null expectation datasets, with decreasing performance for more complex and/or less characterized environments (Douglas et al., 2019; Sun, Jones & Fodor, 2020). These low correlations could be due to preferential amplification of certain DNA sequences, primers biases, and varying gene copy numbers of 16S rRNA per taxa, although HSP methods typically try to correct this bias (Bowman & Ducklow, 2015; Wemheuer et al., 2018; Douglas et al., 2019). The increased detection sensitivity of 16S rRNA metabarcoding can also create a bias in the number of contributing taxa to certain functions, which can negatively affect correlations with functions derived from metagenomics. Considering that amplicon-based 16S rRNA metabarcoding and metagenomics uncovered a substantially different bacterial diversity, the weak correlation in functional abundance between the two methods is expected.

An alternative and possibly more appropriate approach to comparing functional profiles of HSP and metagenomics approaches is by contrasting their correlation with metadata (Sun et al., 2020) or by evaluating the correspondence between the ordination of their functional communities. Using procrustes analyses, there

was a very strong and significant correlation between HSP methods and metagenomics, especially when using presence/absence data. The ASV communities showed no significant relationship with the functional profiles, suggesting that the taxonomic and functional communities were influenced differently by the biological and/or environmental conditions. We also tested the correspondence of the bacterial taxonomic and functional profiles with macrofaunal communities and physico-chemical data. While both profiles correlated relatively well to physico-chemical data (r [?] 0.5), with the EC-based data performing best, only functional profiles correlated strongly and significantly with the macrofaunal communities. A higher association of functional versus taxonomic beta-diversity with macrofaunal data was also reported by Laroche et al. (2018), which suggest that interactions between these communities are especially driven by microbial metabolic capabilities rather than specific phylogenetic association.

The sensitivity of the different datasets in detecting the effect of fish farm activities was evaluated by comparing changes in community composition between near-field (0 m from pen) and far-field samples (>= 1,200 m from pen). In general, we found higher sensitivity for the HSP methods, and especially for Paprica and Tax4Fun2, compared to metagenomics and ASV communities. These results indicate that despite the lower accuracy and increased detection sensitivity of HSP methods, they may be more accurate in assessing how microbial communities respond to environmental changes than metagenomics. This is likely enhanced when complex microbial communities are present, such as in marine sediments. The results improved when transforming functional abundance data, including those of metagenomics, to presence/absence data, as it reduced within group variability. In addition, we observed higher stochasticity of microbial taxonomic shifts in response to a contamination gradient compared to functional community changes. This observation has also been reported by Hornick & Buschmann (2018), Laroche et al. (2018) and Ren et al. (2016) and is likely due to several taxa sharing the same metabolic capabilities and their succession in the ecosystem may have less to do with environmental changes than with biological properties (e.g. growth cycle and bacterial interactions) and geo-topographic factors (e.g. depth and geographic distance). As such, functional profiles may be slightly more robust and sensitive in detecting environmental alterations caused by fish farm activities, although further research is needed to properly test this assumption.

Benthic environments under cage fin-fish aquaculture are usually enriched in organic waste and nutrients such as phosphorus and nitrogen compounds (from faeces and uneaten fish feed for example), which can lead to eutrophic conditions, microbial anaerobic activities and the production of ammonia and hydrogen sulfide gasses (Brooks & Mahnken, 2003; Buschmann et al., 2006; Valdemarsen, Kristensen & Holmer, 2009; Wang et al., 2012). In the present study, we were particularly interested in comparing the response of classes of pathways associated to the nitrogen and sulfur cycles between the pen and reference sites, and between the metagenomics and HSP-based data. Overall, results from both approaches were very similar, with an increase near the pens of pathways associated to nitrate reduction and glycosaminoglycan degradation, and a decrease of pathways affiliated to allantoin degradation and sulfur oxidation. Additionally, the Paprica analysis showed a decrease in pathway abundance associated with sulfur reduction, sulfite and sulfide reduction, and an increase of dimethylsulfide degradation near the pens. While we expected pathways of nitrate-reduction to be in higher abundance near the fish farms, due to enriched nutrients and possibly anoxic conditions, it was somewhat surprising that pathways associated with sulfur compounds were less abundant in both the metagenomics and HSP datasets. However, sulfite respiration was found to be associated with the pens when using taxonomic information of the 16S rRNA data and literature-based functional association (Faprotax methodology). It is likely that certain pathways associated to the sulfur cycle, such as sulfite oxidation and reduction, were indeed more prominent near the pens but were not fully picked-up by metagenomics and HSP-based functional profiling, possibly due to the incompleteness of reference databases. For example, pathways associated to sulfite respiration were absent from both the Humann2 and Picrust2 datasets. Glycosaminoglycan degradation is responsible for the degradation of long linear polysaccharides made of repeating disaccharide units, also referred to as mucopolysaccharides (Ernst et al., 1995). It is probable that high quantities of mucopolysaccharides originate from mucus produced and excreted by the caged salmon (see Reverter et al., 2018; Jacobsen et al., 2019) and this is being catabolized by a specialized group of bacteria. Allantoin represents a product of uric acid, an important metabolic

intermediate compound produced by both animals and bacteria. Under limited nutrient conditions, allantoin can be degraded into ammonia by some bacteria, to serve as a secondary source of nitrogen (Switzer et al., 2020). Correspondingly, our results suggest that pathways associated to allantoin degradation are less abundant near the pens, where strong nutrient enrichment occurs. Overall, these results show congruence between metagenomics and HSP methods for the classes of pathways of particular interest, and highlight both the potential and caveats of the current functional profiling methods in providing further understanding of the metabolic and environmental changes occurring in benthic ecosystems. Further research involving more samples and taking into account regional and temporal variability is needed to confidently identify potential functional indicators of fish farm ecological impacts that could be eventually integrated into benthic health indexes.

Collectively our results suggest that the lower specificity of HSP methods may be offset by the ability of amplicon-based metabarcoding to provide a much more exhaustive assessment of the taxonomic community, and hence of functions that have low genomic variability. This allows HSP methods to provide functional profiles that are relatively similar to those of metagenomics, and which respond similarly to environmental changes. Although the accuracy and sensitivity of HSP methods are still strongly affected by the incompleteness of reference databases, our results demonstrate that they provide a useful functional profiling alternative to metagenomics, and a valuable tool in detecting and evaluating the effects of salmon farming on benthic ecosystems.

5 | Acknowledgements

Support and funding for this project was provided by the Norwegian Research Council under the project 267829 SUSTAINable AQUAculture in the North: identifying thresholds, indicators and tools for future growth. We thank U. von Ammon at Cawthron Institute for preparing the 16S rRNA gene libraries, C. Gebbie at Auckland Genomics for sequencing of the metabarcoding data, G. D. Gilfillan and T. Ribarska from the Norwegian High-Throughput Sequencing Centre (NSC) for laboratory support, library preparation and sequencing of the metagenomics data, and the Norwegian e-infrastructure for Research and Education (Uninett-Sigma2) for storage and computing resources.

6 | Author contributions

NK and OL designed the study. NK, OL and XP conducted field work and sampling at sea. OL, NK and XP generated the data and OL performed the analyses and wrote the manuscript with intellectual contributions from all co-authors. NK provided grant and equipment support.

7 | Data Accessibility

Unprocessed sequences are accessible from the NCBI Sequence Read Archive (SRA) under project number PRJNA661323. Metadata for the samples are available in the Supplementary Tables.

Figures and tables

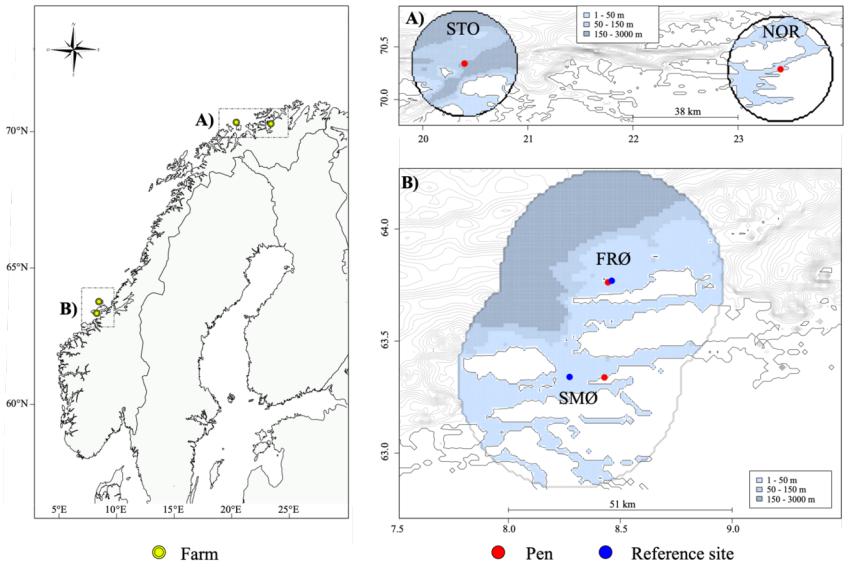

Fig 1 . Overview of farm location within Northern and Southern Norway (left panel), and station location and bathymetry at the STO and NOR farms (**A** ; upper right panel) and at the FRØ and SMØ farms (**B** ; bottom right panel).

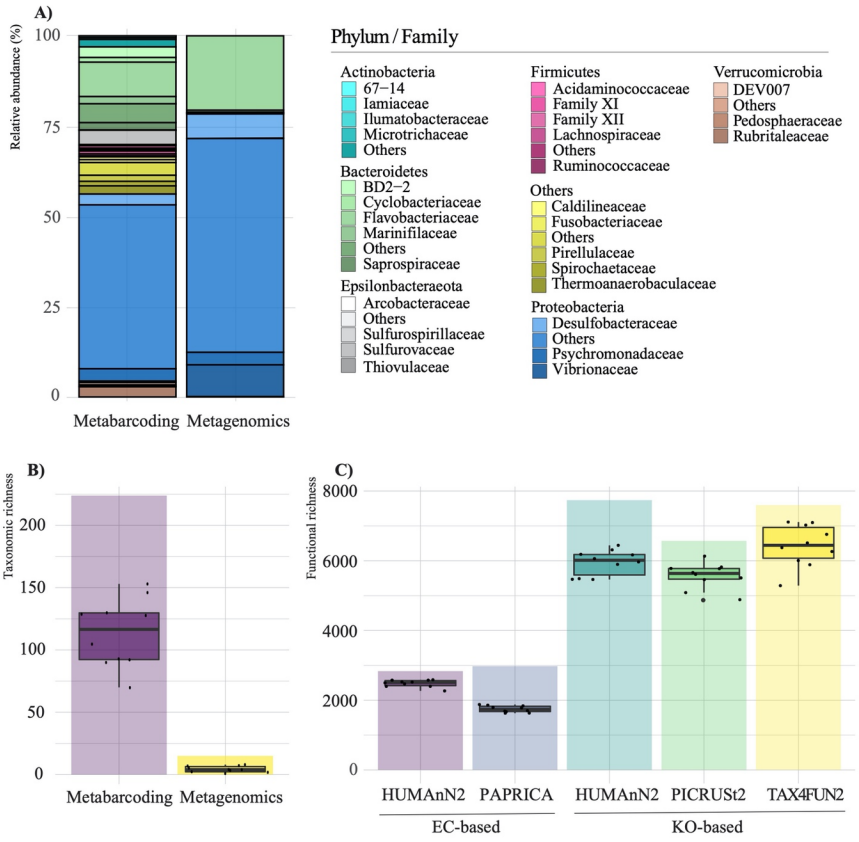

Fig 2 . Taxonomic composition at family and phylum level (**A**), and taxonomic (**B**) and functional (**C**) richness per dataset. In **A** , unclassified families and those out of the most five abundant families within their phylum were grouped under "Others". Similarly, Phyla out of the top six most abundant were grouped under "Others". In **B** and **C** , barplots represent the total richness per dataset while boxplots indicate richness per sample and dataset. Taxonomic richness comparison is made at family level. In **C** , Humann2 columns represent the metagenomics data. EC= Enzyme commission numbers, and KO = KEGG ortholog numbers.

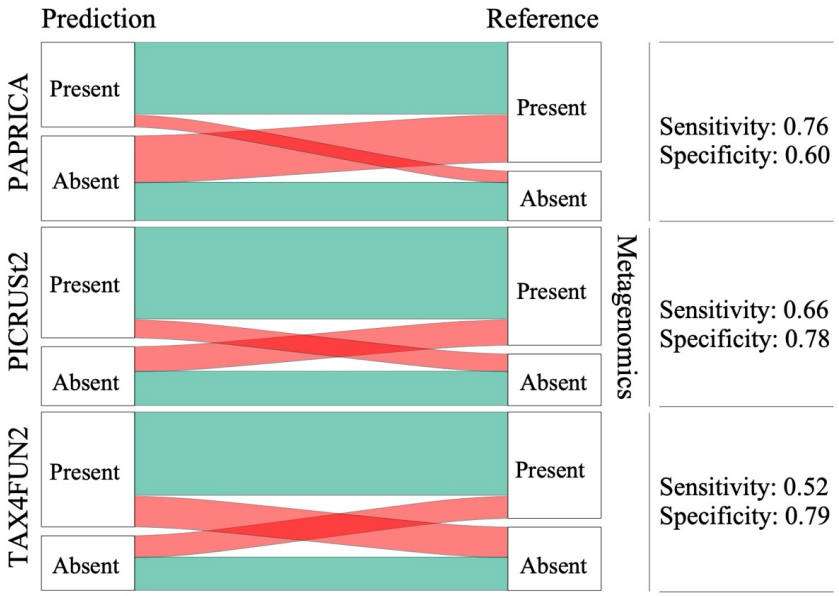

Fig 3 . Functional inferences of hidden state prediction methods (Paprica, Picrust2 and Tax4Fun2) (Prediction), versus metagenomics data (Reference). Metrics on the right side of the alluvial figures derive from the confusion Matrix function of the 'caret' R package.

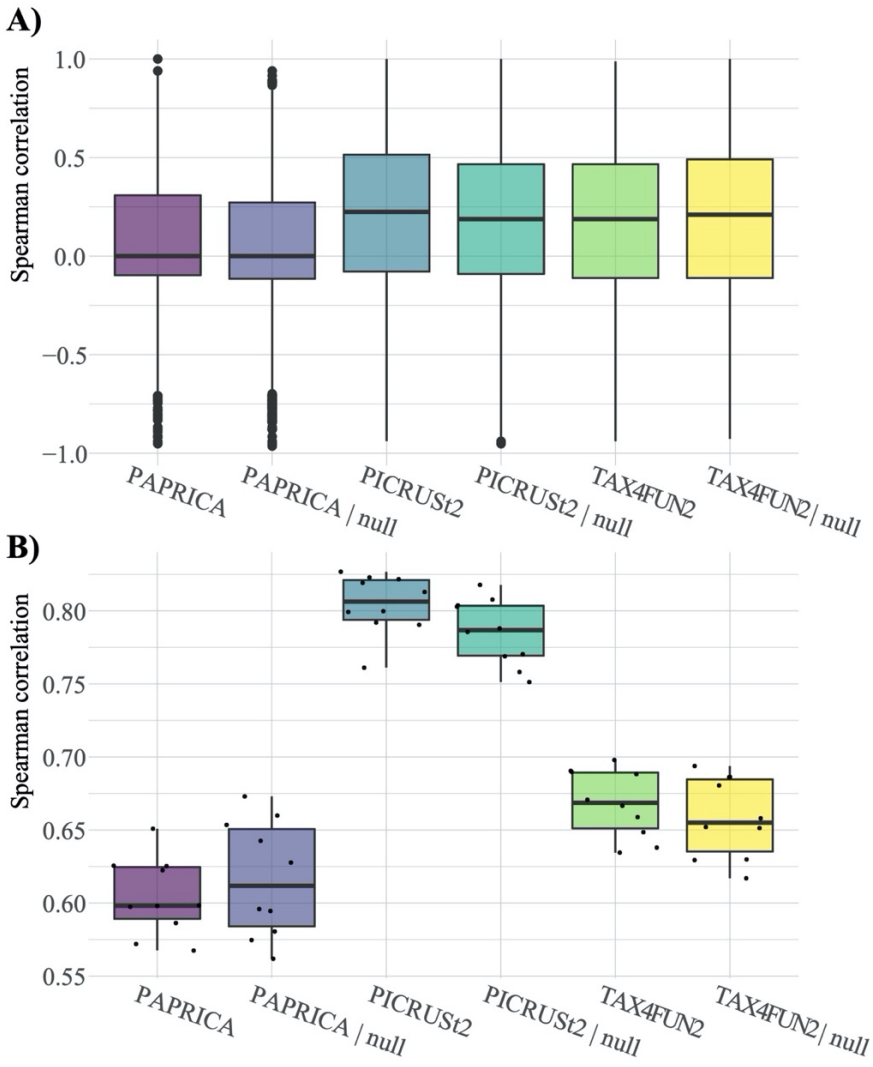

Fig 4 . Spearman correlations of the functions per Enzyme commission numbers (EC) and KEGG ortholog numbers (KO) (**A**) and per samples (**B**) between hidden state prediction methods (Paprica, Picrust2 and Tax4Fun2) and metagenomics data, each a priori transformed to centered-log ratio transformed.

Hosted file

`image5.emf` available at https://authorea.com/users/366620/articles/486313-beyond-taxonomy-validating-functional-inference-approaches-in-the-context-of-fish-farm-impact-assessments

Fig 5. Procrustes analysis of the 16S rRNA gene metabarcoding-based datasets with the functional community of the metagenomics data. The amplicon sequence variant (ASVs) and Paprica datasets were fitted with the Enzyme commission numbers (EC) profiles while Picrust2 and Tax4Fun2 were fitted with the KEGG ortholog numbers (KO) profiles of the metagenomics data.

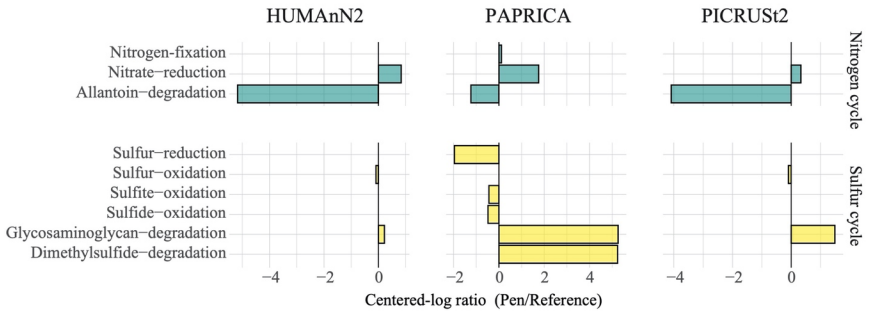

Fig 6. Centered-log ratio (CLR) abundance of pathway classes of particular interest between metagenomics (Humann2) and hidden-state prediction (HSP) methods (Paprica and Picrust2).

Classes of pathways with positive CLR were more prevalent at the pens while those with negative CLR were more prevalent at the reference sites.

Table 1 . Permutational analysis of variance of the taxonomic (amplicon sequence variant; ASVs) and functional communities between distance categories (pen versus reference sites) per methodology using 999 permutations. Significant responses (p.value [?] 0.05) are displayed in bold.

		Distance from pen	Distance from pen
Transf.	Method	R^2	p.value
Centered-log ratio	ASVs	0.19	**0.009**
	Paprica	0.26	**0.007**
	Picrust2	0.16	0.151
	Tax4Fun2	0.26	**0.009**
	Humann2 (EC)	0.18	**0.040**
	Humann2 (KO)	0.17	**0.027**
Presence Absence	ASVs	0.17	**0.014**
	Paprica	0.39	**0.012**
	Picrust2	0.26	**0.025**
	Tax4Fun2	0.40	**0.008**
	Humann2 (EC)	0.24	**0.045**
	Humann2 (KO)	0.24	**0.029**

Table 2 . Protest analysis of the 16S rRNA gene metabarcoding-based and metagenomics data with macrofauna and physico-chemical data, using 9,999 permutations. Significant responses (p.value [?] 0.05) are displayed in bold. ASV = amplicon sequence variant.

		Macrofauna	Macrofauna	Physico-chemical*	Physico-chemical*	
Transf.	Method	r	p.value	r	p.value	p.value
Centered-log ratio	ASVs	0.49	0.214	0.72	0.111	0.111
	Paprica	0.88	**<0.001**	0.98	**0.048**	**0.048**
	Picrust2	0.68	**0.019**	0.71	0.135	0.135
	Tax4Fun2	0.87	**<0.001**	0.74	0.126	0.126

		Macrofauna	Macrofauna	Physico-chemical*	Physico-chemical*	
	Humann2 (EC)	0.88	**0.001**	0.96	**0.004**	**0.004**
	Humann2 (KO)	0.88	**0.001**	0.67	0.211	0.211
Presence Absence	ASVs	0.41	0.453	0.72	0.111	0.111
	Paprica	0.90	**<0.001**	0.98	**0.048**	**0.048**
	Picrust2	0.69	**0.011**	0.50	0.126	0.126
	Tax4Fun2	0.95	**<0.001**	0.74	0.126	0.126
	Humann2 (EC)	0.90	**<0.001**	0.96	**0.004**	**0.004**
	Humann2 (KO)	0.89	**<0.001**	0.67	0.211	0.211

* Analysis with the physico-chemical data only included samples (n=6) from the FRØ locality.

References

Aguinaga OE., McMahon A., White KN., Dean AP., Pittman JK. 2018. Microbial community shifts in response to acid mine drainage pollution within a natural wetland ecosystem. *Frontiers in Microbiology*9:1–14. DOI: 10.3389/fmicb.2018.01445.

Anderson MJ. 2005. PERMANOVA: A FORTRAN computer program for permutational multivariate analysis of variance. *Department of Statistics* .

Arfken A., Song B., Bowman JS., Piehler M. 2017. Denitrification potential of the eastern oyster microbiome using a 16S rRNA gene based metabolic inference approach. *PLOS ONE* 12:e0185071. DOI: 10.1371/journal.pone.0185071.

Barberan A., Fernandez-Guerra A., Bohannan BJM., Casamayor EO. 2012. Exploration of community traits as ecological markers in microbial metagenomes. *Molecular Ecology* 21:1909–1917. DOI: 10.1111/j.1365-294X.2011.05383.x.

Bojanowski M., Edwards R. 2016. alluvial: R Package for Creating Alluvial Diagrams.

Bolger AM., Lohse M., Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.

van den Boogaart KG., Tolosana-Delgado R., Bren M. 2020. compositions: Compositional Data Analysis.

Bourlat SJ., Borja A., Gilbert J., Taylor MI., Davies N., Weisberg SB., Griffith JF., Lettieri T., Field D., Benzie J., Glöckner FO., Rodríguez-Ezpeleta N., Faith DP., Bean TP., Obst M. 2013. Genomics in marine monitoring: New opportunities for assessing marine health status.*Marine Pollution Bulletin* 74:19–31. DOI: 10.1016/j.marpolbul.2013.05.042.

Bowman JS., Ducklow HW. 2015. Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the coastal West Antarctic Peninsula. *PLoS ONE* 10:1–18. DOI: 10.1371/journal.pone.0135868.

Breitwieser FP., Lu J., Salzberg SL. 2018. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics* 20:1125–1139. DOI: 10.1093/bib/bbx120.

Brooks KM., Mahnken CVW. 2003. *Interactions of Atlantic salmon in the Pacific northwest environment. II. Organic wastes* . DOI: 10.1016/S0165-7836(03)00064-X.

Buchfink B., Xie C., Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12:59–60. DOI: 10.1038/nmeth.3176.

Buschmann AH., Riquelme VA., Hernández-González MC., Varela D., Jiménez JE., Henríquez LA., Vergara PA., Guíñez R., Filún L. 2006. A review of the impacts of salmonid farming on marine coastal ecosystems in the southeast Pacific. *ICES Journal of Marine Science* 63:1338–1345. DOI: 10.1016/j.icesjms.2006.04.021.

Callahan BJ., McMurdie PJ., Rosen MJ., Han AW., Johnson AJA., Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13:581–583. DOI: 10.1038/nmeth.3869.

Cordier T. 2020. Bacterial communities' taxonomic and functional turnovers both accurately predict marine benthic ecological quality status. *Environmental DNA* 2:175–183. DOI: 10.1002/edn3.55.

Cordier T., Alonso-Saez L., Apotheloz-Perret-Gentil L., Aylagas E., Bohan DA., Bouchez A., Chariton A., Creer S., Fruhe L., Keck F., Keeley N., Laroche O., Leese F., Pochon X., Stoeck T., Pawlowski J., Lanzen A. 2020. Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology* :mec.15472. DOI: 10.1111/mec.15472.

Danovaro R., Carugati L., Berzano M., Cahill AE., Carvalho S., Chenuil A., Corinaldesi C., Cristina S., David R., Dell'Anno A., Dzhembekova N., Garces E., Gasol JM., Goela P., Feral J-P., Ferrera I., Forster RM., Kurekin AA., Rastelli E., Marinova V., Miller PI., Moncheva S., Newton A., Pearman JK., Pitois SG., Rene A., Rodriguez-Ezpeleta N., Saggiomo V., Simis SGH., Stefanova K., Wilson C., Lo Martire M., Greco S., Cochrane SKJ., Mangoni O., Borja A. 2016. Implementing and innovating marine monitoring approaches for assessing marine environmental status.*Frontiers in Marine Science* 3:213. DOI: 10.3389/fmars.2016.00213.

Deiner K., Bik HM., Machler E., Seymour M., Lacoursiere-Roussel A., Altermatt F., Creer S., Bista I., Lodge DM., Vere N., Pfrender ME., Bernatchez L. 2017. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*26:5872–5895. DOI: 10.1111/mec.14350.

Douglas GM., Maffei VJ., Zaneveld J., Yurgel SN., Brown JR., Taylor CM., Huttenhower C., Langille MGI. 2019. PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv* :672295. DOI: 10.1101/672295.

Ernst S., Langer R., Cooney CL., Sasisekharan R. 1995. Enzymatic Degradation of GlycosaminogIycans. *Critical Reviews in Biochemistry and Molecular Biology* 30:387–444. DOI: 10.3109/10409239509083490.

Escalas A., Hale L., Voordeckers JW., Yang Y., Firestone MK., Alvarez-Cohen L., Zhou J. 2019. Microbial Functional Diversity: From Concepts to Applications. *Ecology and Evolution* 9:1–17. DOI: 10.1002/ece3.5670.

Franzosa EA., McIver LJ., Rahnavard G., Thompson LR., Schirmer M., Weingart G., Lipson KS., Knight R., Caporaso JG., Segata N., Huttenhower C. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* 15:962–968. DOI: 10.1038/s41592-018-0176-y.

Gloor GB., Macklaim JM., Pawlowsky-Glahn V., Egozcue JJ. 2017. Microbiome datasets are compositional: And this is not optional.*Frontiers in Microbiology* 8:1–6. DOI: 10.3389/fmicb.2017.02224.

Grossart H., Massana R., McMahon KD., Walsh DA. 2020. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles.*Limnology and Oceanography* 65. DOI: 10.1002/lno.11382.

Hong Y., Wu J., Wilson S., Song B. 2019. Vertical stratification of sediment microbial communities along geochemical gradients of a subterranean estuary located at the gloucester beach of Virginia, United States. *Frontiers in Microbiology* 10:1–11. DOI: 10.3389/fmicb.2018.03343.

Hornick KM., Buschmann AH. 2018. Insights into the diversity and metabolic function of bacterial communities in sediments from Chilean salmon aquaculture sites. *Annals of Microbiology* 68:63–77. DOI: 10.1007/s13213-017-1317-8.

Iwai S., Weinmaier T., Schmidt BL., Albertson DG., Poloso NJ., Dabbagh K., DeSantis TZ. 2016. Piphillin: Improved prediction of metagenomic content by direct inference from human microbiomes. *PLoS ONE*11:1–18. DOI: 10.1371/journal.pone.0166104.

Jacobsen A., Shi X., Shao C., Eysturskarð J., Mikalsen S-O., Zaia J. 2019. Characterization of Glycosaminoglycans in Gaping and Intact Connective Tissues of Farmed Atlantic Salmon (Salmo salar) Fillets by Mass Spectrometry. *ACS Omega* 4:15337–15347. DOI: 10.1021/acsomega.9b01136.

Kaul A., Mandal S., Davidov O., Peddada SD. 2017. Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*8:1–10. DOI: 10.3389/fmicb.2017.02114.

Keeley N., Valdemarsen T., Woodcock S., Holmer M., Husa V., Bannister R. 2019. Resilience of dynamic coastal benthic ecosystems in response to large-scale finfish farming. *Aquaculture Environment Interactions*11:161–179. DOI: 10.3354/aei00301.

Klindworth A., Pruesse E., Schweer T., Peplies J., Quast C., Horn M., Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* 41:1–11. DOI: 10.1093/nar/gks808.

Kuhn M. 2020. caret: Classification and Regression Training.

Langille MGI., Zaneveld J., Caporaso JG., McDonald D., Knights D., Reyes JA., Clemente JC., Burkepile DE., Vega Thurber RL., Knight R., Beiko RG., Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* 31:814–821. DOI: 10.1038/nbt.2676.

Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359. DOI: 10.1038/nmeth.1923.

Laroche O., Pochon X., Tremblay LA., Ellis JI., Lear G., Wood SA. 2018. Incorporating molecular-based functional and co-occurrence network properties into benthic marine impact assessments. *FEMS Microbiology Ecology* 94:1–12. DOI: 10.1093/femsec/atx167.

Leese F., Altermatt F., Bouchez A., Ekrem T., Hering D., Mergen P., Pawlowski J., Piggott J., Abarenkov K., Beja P., Bervoets L., Boets P., Bones A., Borja Á., Bruce K., Carlsson J., Coissac E., Costa F., Costache M., Creer S., Csabai Z., Deiner K., DelValls Á., Duarte S., Fazi S., Graf W., Hershkovitz Y., Japoshvili B., Jones J., Kahlert M., Kalamujic Stroil B., Kelly-Quinn M., Keskin E., Mächler E., Mahon A., Marečková M., Mejdandzic M., Montagna M., Moritz C., Mulk V., Navodaru I., Pálsson S., Panksep K., Penev L., Petrusek A., Pfannkuchen M., Rinkevich B., Schmidt-Kloiber A., Segurado P., Strand M., Šulčius S., Traugott M., Turon X., Valentini A., van der Hoorn B., Vasquez Hadjilyra M., Viguri J., Vogler A., Zegura B. 2016. DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes* 2:e11321. DOI: 10.3897/rio.2.e11321.

Lobo J., Shokralla S., Costa MH., Hajibabaei M., Costa FO. 2017. DNA metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. *Scientific Reports* 7:15618. DOI: 10.1038/s41598-017-15823-6.

Louca S., Parfrey LW., Doebeli M. 2016. Decoupling function and taxonomy in the global ocean microbiome. *Science* 353:1272–1277. DOI: 10.1126/science.aaf4507.

Louca S., Polz MF., Mazel F., Albright MBN., Huber JA., O'Connor MI., Ackermann M., Hahn AS., Srivastava DS., Crowe SA., Doebeli M., Parfrey LW. 2018. Function and functional redundancy in microbial systems.*Nature Ecology & Evolution* 2:936–943. DOI: 10.1038/s41559-018-0519-1.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10. DOI: 10.14806/ej.17.1.200.

Martiny JBH., Jones SE., Lennon JT., Martiny AC. 2015. Microbiomes in light of traits: A phylogenetic perspective. *Science* 350. DOI: 10.1126/science.aac9323.

Millares L., Pérez-Brocal V., Ferrari R., Gallego M., Pomares X., García-Núñez M., Montón C., Capilla S., Monsó E., Moya A. 2015. Functional Metagenomics of the Bronchial Microbiome in COPD. *PLOS ONE* 10:e0144448. DOI: 10.1371/journal.pone.0144448.

Nagpal S., Haque MM., Mande SS. 2016. Vikodak - A modular framework for inferring functional potential of microbial communities from 16S metagenomic datasets. *PLoS ONE* 11:1–19. DOI: 10.1371/journal.pone.0148347.

Oksanen J., Blanchet FG., Friendly M., Kindt R., Legendre P., McGlinn D., Minchin PR., O'Hara RB., Simpson GL., Solymos P., Stevens MHH., Szoecs E., Wagner H. 2019. vegan: Community Ecology Package.

Pacheco-Sandoval A., Schramm Y., Heckel G., Brassea-Pérez E., Martínez-Porchas M., Lago-Lestón A. 2019. The Pacific harbor seal gut microbiota in Mexico: Its relationship with diet and functional inferences. *PLoS ONE* 14:1–21. DOI: 10.1371/journal.pone.0221770.

Pawlowski J., Kelly-Quinn M., Altermatt F., Apothéloz-Perret-Gentil L., Beja P., Boggero A., Borja A., Bouchez A., Cordier T., Domaizon I., Feio MJ., Filipe AF., Fornaroli R., Graf W., Herder J., van der Hoorn B., Iwan Jones J., Sagova-Mareckova M., Moritz C., Barquín J., Piggott JJ., Pinna M., Rimet F., Rinkevich B., Sousa-Santos C., Specchia V., Trobajo R., Vasselon V., Vitecek S., Zimmerman J., Weigand A., Leese F., Kahlert M. 2018. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems.*Science of the Total Environment* 637–638:1295–1310. DOI: 10.1016/j.scitotenv.2018.05.002.

Pearman JK., Aylagas E., Voolstra CR., Anlauf H., Villalobos R., Carvalho S. 2019. Disentangling the complex microbial community of coral reefs using standardized Autonomous Reef Monitoring Structures (ARMS).*Molecular Ecology* 28:3496–3507. DOI: 10.1111/mec.15167.

Pilliod DS., Laramie MB., MacCoy D., Maclean S. 2019. Integration of eDNA-Based Biological Monitoring within the U.S. Geological Survey's National Streamgage Network. *Journal of the American Water Resources Association* :1–14. DOI: 10.1111/1752-1688.12800.

Pollock J., Glendinning L., Wisedchanwet T., Watson M. 2018. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and Environmental Microbiology*84:1–12. DOI: 10.1128/AEM.02627-17.

R Core Team. 2017. R: A Language and Environment for Statistical Computing.

Ren Y., Niu J., Huang W., Peng D., Xiao Y., Zhang X., Liang Y., Liu X., Yin H. 2016. Comparison of microbial taxonomic and functional shift pattern along contamination gradient. *BMC Microbiology* 16:110. DOI: 10.1186/s12866-016-0731-6.

Reverter M., Tapissier-Bontemps N., Lecchini D., Banaigs B., Sasal P. 2018. Biological and ecological roles of external fish mucus: A review.*Fishes* 3:1–19. DOI: 10.3390/fishes3040041.

Ruppert KM., Kline RJ., Rahman MS. 2019. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation* 17:e00547. DOI: 10.1016/j.gecco.2019.e00547.

Salerno JL., Little B., Lee J., Hamdan LJ. 2018. Exposure to crude oil and chemical dispersant may impact marine microbial biofilm composition and steel corrosion. *Frontiers in Marine Science* 5:1–14. DOI: 10.3389/fmars.2018.00196.

Starke R., Capek P., Morais D., Callister SJ., Jehmlich N. 2020. The total microbiome functions in bacteria and fungi. *Journal of Proteomics* 213:103623. DOI: 10.1016/j.jprot.2019.103623.

Sun S., Jones RB., Fodor AA. 2020. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* 8:1–9. DOI: 10.1186/s40168-020-00815-y.

Suzek BE., Wang Y., Huang H., McGarvey PB., Wu CH. 2015. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31:926–932. DOI: 10.1093/bioinformatics/btu739.

Switzer A., Burchell L., McQuail J., Wigneshweraraj S. 2020. The adaptive response to long-term nitrogen starvation in Escherichia coli requires the breakdown of allantoin . *Journal of Bacteriology* :1–11. DOI: 10.1128/jb.00172-20.

Thomsen PF., Willerslev E. 2015. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity.*Biological Conservation* 183:4–18. DOI: 10.1016/j.biocon.2014.11.019.

Truong DT., Franzosa EA., Tickle TL., Scholz M., Weingart G., Pasolli E., Tett A., Huttenhower C., Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 12:902–903. DOI: 10.1038/nmeth.3589.

Valdemarsen T., Kristensen E., Holmer M. 2009. Metabolic threshold and sulfide-buffering in diffusion controlled marine sediments impacted by continuous organic enrichment. *Biogeochemistry* 95:335–353. DOI: 10.1007/s10533-009-9340-x.

Valentini A., Taberlet P., Miaud C., Civade R., Herder J., Thomsen PF., Bellemain EVA., Besnard A., Coissac E., Boyer F., Gaboriaud C., Jean P., Poulet N., Roset N., Copp GH., Geniez P., Pont D., Argillier C., Baudoin J-M., Peroux T., Crivelli AJ., Olivier A., Acqueberge M., Le Brun M., Møller PR., Willerslev E., Dejean T., Civade L., Valentini A., Taberlet P., Miaud C., Rapha E., Besnard L., Thomsen PF., Bellemain EVA., Aur E., Argillier C., Baudoin J-M., Peroux T., Alain J. 2016. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology* 25:929–942. DOI: 10.1111/mec.13428.

Wang X., Olsen L., Reitan K., Olsen Y. 2012. Discharge of nutrient wastes from salmon farms: environmental effects, and potential for integrated multi-trophic aquaculture. *Aquaculture Environment Interactions* 2:267–283. DOI: 10.3354/aei00044.

Wang P., Yan Z., Yang S., Wang S., Zheng X., Fan J., Zhang T. 2019. Environmental DNA: An Emerging Tool in Ecological Assessment.*Bulletin of Environmental Contamination and Toxicology*103:651–656. DOI: 10.1007/s00128-019-02720-z.

Wang L., Zhang J., Li H., Yang H., Peng C., Peng Z., Lu L. 2018. Shift in the microbial community composition of surface water and sediment along an urban river. *Science of the Total Environment*627:600–612. DOI: 10.1016/j.scitotenv.2018.01.203.

Wemheuer F., Taylor JA., Daniel R., Johnston E., Meinicke P., Thomas T., Wemheuer B. 2018. Tax4Fun2: a R-based tool for the rapid prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene marker gene sequences. *bioRxiv* :490037. DOI: 10.1101/490037.

White CA., Woodcock SH., Bannister RJ., Nichols PD. 2017. Terrestrial fatty acids as tracers of finfish aquaculture waste in the marine environment. *Reviews in Aquaculture* 11:133–148. DOI: 10.1111/raq.12230.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis* . Springer-Verlag New York.

Woodcock SH., Meier S., Keeley NB., Bannister RJ. 2019. Fate and longevity of terrestrial fatty acids from caged fin-fish aquaculture in dynamic coastal marine systems. *Ecological Indicators* 103:43–54. DOI: 10.1016/j.ecolind.2019.03.057.

Zaneveld JRR., Thurber RL V. 2014. Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. *Frontiers in Microbiology* 5:1–8. DOI:

10.3389/fmicb.2014.00431.