TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Towards standardized operating procedures for eDNA-based monitoring of marine coastal ecosystems

vom Fachbereich Biologie der Technischen Universität Kaiserslautern zur Verleihung des akademischen Grades Dr. rer. nat. genehmigte Dissertation

von

M. Sc. Verena Nicola Rubel, geb. Dully

geboren in Landau in der Pfalz

Datum der wissenschaftlichen Aussprache: 24.06.2022

Dekanin, Vorsitzende der Promotionskommission: Prof. Dr. Nicole Frankenberg-Dinkel

Berichterstatter: Prof. Dr. Thorsten Stoeck, Prof. Dr. Timo Mühlhaus

D 386

ERKLÄRUNG / DECLARATION

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig angefertigt und alle benutzten Hilfsmittel und Hilfestellungen in der Arbeit angegeben habe. Ich erkläre, dass die eingereichte Dissertation oder Teile hiervon nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht wurden. Ich erkläre, dass weder die vorliegende Dissertation oder eine andere Dissertation bei einem anderen Fachbereich oder einer anderen Universität als Dissertation eingereicht wurden.

I hereby declare that I have produced the present dissertation independently. The use of other sources or auxiliary material has been properly and fully acknowledged. Furthermore, I declare that the present dissertation has not been presented complete or partial to any other institution or university with the intent to obtain an academic degree.

_____

Verena Nicola Dully

Kaiserslautern, 04.04.2022

TABLE OF CONTENTS

TABLE OF ABBREVIATIONS **1**
PREFACE **3**
INTRODUCTION **5**
Aquaculture-induced eutrophication 5
Environmental monitoring 6
eDNA metabarcoding 9
Supervised machine learning applications in the field 12
The Random Forest algorithm in particular 14
A further approach: Quantile Regression Splines 17
Towards standard operating procedures 17
CHAPTER I **31**
Sample preservation for transport **31**
Summary 31
Publication: 41
Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems
CHAPTER II **65**
Reproducibility of eDNA metabarcoding-based sample processing **65**
Summary 65
Publication: 75
Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture
CHAPTER III **109**
Robustness and applicability of current approaches for EQ inference **109**
Summary 109
Background 110
Methods 114
Results 117
Discussion 124
References 138
CHAPTER IV **147**
Required sequencing depth for SML-based EQ inference **147**
Summary 147
Publication: 157
Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning

SYNTHESIS AND OUTLOOK **187**
Retrospective 187
Recommendations for eDNA metabarcoding-based monitoring approaches 191
Future monitoring scenarios 193
Conclusion 195
BIBLIOGRAPHY **203**
SUMMARY **229**
ZUSAMMENFASSUNG **233**
APPENDIX **237**
Statement of contributions 237
List of figures, tables, and equations 239
Supplementary files 241
Supplementary files of Chapter I 243
Supplementary files of Chapter II 247
Supplementary files of Chapter III 249
Supplementary files of Chapter IV 265
CURRICULUM VITAE **275**
List of publications 277
Journal articles 277
Presentations 278
Awards 278
ACKNOWLEDGEMENTS **281**

TABLE OF ABBREVIATIONS

AMBI	AZTI's Marine Biotic Index
ASV	Amplicon sequence variant
AZE	Allowable zone of effect (specific salmon farm site)
CART	Classification and regression tree
CE	Cage edge (specific salmon farm site)
CI	Confidence interval
CoV	Coefficient of variation
CV	Cross-validation
DADA2	Divisive amplicon denoising algorithm
DNA	Desoxyribonucleic acid
eDNA	Environmental desoxyribonucleic acid
EG	Ecological group, 'Eco-Group'
EIA	Environmental impact assessments
EQ	Ecological quality
Eq.	Equation
eRNA	Environmental ribonucleic acid
HQ	High-quality
HTS	High-throughput sequencing
IndVal	Indicator Value approach
IQI	Infaunal Quality Index
LOO-CV	Leave-one-out cross-validation
m	Meters
NGS	Next-generation sequencing
NMDS	Non-metric multidimensional scaling
OOB	Out-of-bag
OTU	Operational taxonomic units
PCR	Polymerase chain reaction
PSU	Practical salinity units
QRS	Quantile Regression Spline
REF	Reference site (specific salmon farm site)
Rep	Replicate
RF	Random Forest algorithm
RMSE	Root mean squared error
RNA	Ribonucleic acid
SBS	Sequencing by synthesis
SML	Supervised machine learning
SOP	Standard operating procedure
SSU	Small subunit
UK	United Kingdom
WFD	Water Framework Directive

PREFACE

While marine coastal ecosystems are precious for humanity, pollution and other anthropogenic influences can negatively affect them. To ensure biological conservation and ecosystem services, constant monitoring is required. To react rapidly and properly, monitoring should be conducted in a timely manner. Novel up-to-date molecular methods demonstrated their potential towards modern biomonitoring via extension or even replacement of traditional biomonitoring approaches. New approaches must be tested extensively for each habitat individually to be included in official regulations in the long term. There have been calls for method standardization from the scientific community and operators lately, since only standard operating procedures can be used to gain legislative recognition. By providing an insight into some of these processes yet to be standardized, this thesis aims to aid in the implementation of molecular methods into routine biomonitoring.

In total, this thesis is divided into six main parts, the Introduction, the Chapters I-IV, and a summarizing discussion. For the overall background and explanation of technical terms, please refer to the Introduction. Chapters I, II and IV provide an overview of research results that have already been published and place them into a larger context. Additionally, Chapter III comprises a study that is currently being prepared to be submitted for publication. The obtained conclusions are put into perspective by a summary discussion at the end of the thesis. Published supplementary material can be found in the appendix of this thesis. Literature cited is listed at the end of each respective section but is also incorporated at the end of the work in a summarized bibliography.

INTRODUCTION

Marine coastal environments offer a huge variety of ecosystem services to humanity. Ecosystem services are considered as benefits that people can obtain directly or indirectly from the respective ecosystem (Hassan et al., 2005), ranging from recreational and touristic benefits to habitat functions (Barbier et al., 2011; Daily, 2013; Drius et al., 2019). Habitat functions of coastal marine ecosystems include water purification, climate control, erosion control, and providing of nursery habitats for fishes (Barbier et al., 2011; Heal et al., 2005; Mcleod et al., 2011; Miller et al., 2011). By implication, big parts of the shoreline are inhabited, resulting in an ever-increasing anthropogenic impact on the marine coastal ecosystem (Small and Nicholls, 2003). Therefore, marine coastal environments belong to the mostly disturbed ecosystems worldwide and are primarily influenced by pollution (Gray, 1997; Karr, 1991; Lotze et al., 2006). Marine pollution is caused predominantly by nutrient input (eutrophication), as nutrients are discharged from industrial, residential, and agricultural areas (Weis, 2015). Additionally, aquaculture industry promotes coastal ecosystem pollution, where especially coastal net-pen farming of finfish like salmonids contributes to the eutrophication of the surrounding environment (Weis, 2015). Apart from finfish aquaculture carried out globally, production within Europe has been steadily increasing since the nineties (FAO, 2020). Northern European Countries focus on the production of the finfish Atlantic salmon (*Salmo salar*), with especially Norway holding over 46% of total European aquaculture production (EEA, 2018). Worldwide, the demand for fish is growing by approximately 3.1% annually, resulting in 156 million tons of fish ending up on our plates in 2018, facing enormous environmental impacts such as overfished oceans and aquaculture-induced pollution (FAO, 2020). It is assumed that by 2030, 62% of all fish consumed will originate from aquaculture to continue to serve the increasing global demand (Kobayashi et al., 2015).

Aquaculture-induced eutrophication

Early studies regarding pollution of the environment by aquaculture have shown that uneaten food and fish feces sink to the seafloor below the fish cages affecting the naturally occurring biotic community (Brown et al., 1987; Forrest et al., 2007).

However, organic substances do not just accumulate in the immediate vicinity of the aquaculture installations but are dispersed along with the prevailing current up to 1000m from the cages (Sarà et al., 2004; 2006). This dispersal results in an enrichment gradient, showing decreasing influence on the benthic assemblages with increasing distance from the cages (Forrest et al., 2007). Especially at near-cage sites, certain bacteria begin to decompose carbon from the excessive nutrient input requiring oxygen (Bannister et al., 2014). Such bacteria-induced degradation can deplete the surrounding sediment from oxygen, sometimes even inducing anoxic conditions (Bannister et al., 2014; Carroll et al., 2003). Oxygen scarcity prevents various organisms from living, thus changing the macrofauna community composition (Brown et al., 1987; Forrest et al., 2007; Grall and Glémarec, 1997). The marine benthic macrofauna refers to bottom-dwelling organisms like polychaete worms, starfishes, and marine mollusks, which are between 2mm and 20mm in size (Hintermeier-Erhard and Zech, 1997). Additionally, high nutrient input rates can cause local algal blooms, resulting not only in further decreasing oxygen levels but also in the production of harmful substances like ammonia, methane, or hydrogen sulfide, which again influence the natural community of organisms (Weis, 2015).

Environmental monitoring

To maintain a balance between the economic usage of ecosystems and ecological conservation, constant surveillance of the environmental health status is required. There are various legal requirements worldwide that require tracking the health state of marine environments under anthropogenic influences, like the European Water Framework Directive (WFD, 2000). Prior to farm expansion or the construction of a new farm, environmental impact assessment (EIA) is required to describe the potential impact of the new salmon farm (FAO, 1996; Hughes, 1975). For established farms, so-called routine compliance monitoring is required to assure fulfillment of legal regulations. To assess the impact of aquaculture installations, the operators are instructed to take sediment samples along the enrichment gradient, starting directly at the fish cages and proceeding along with the current, up to unpolluted reference areas. Traditionally, macrofauna organisms are used for such biomonitoring approaches. Within the macrofauna are organisms that can adapt to environmental impacts to a greater or lesser degree, therefore their community composition contains information on the quality of the environment (Markert et al., 1999).

Due to different frequencies of occurrence, they reflect different environmental conditions and are therefore considered as bioindicators (Forrest et al., 2007; Pearson and Rosenberg, 1978). Stenoecious organisms, also known as less opportunistic organisms, have little tolerance for fluctuations in environmental factors such as pollution, therefore they are found preferentially in uncontaminated to slightly contaminated sediments. Exemplary, the frequency of the stenoecious sea snail *Tritonoharpa leali* decreases drastically once the organic load of the marine sediment increases (Pearson and Rosenberg, 1978; Wenqian et al., 2013). In contrast, the polychaete worm *Capitella capitata* is an opportunistic or euryecious organism, implying it can tolerate a wide variation in environmental factors. *C. capitata* occurs preferentially where other organisms can no longer appear numerously due to increased pollution (Reish, 1955). Because such opportunistic species experience little competition in heavily polluted sediments, they can be often found in large numbers (Holmer et al., 2005).

For traditional environmental biomonitoring, the macrofaunal community is assessed by taxonomical identification and enumeration of organisms. The AMBI (AZTI's Marine Biotic Index) is one of the most widely used indices for quantifying aquaculture-induced environmental impacts based on macrofaunal community composition (Borja et al., 2000). AZTI researchers have published a software suitable for environmental evaluations that includes organisms from the major soft-bottom communities from Asia, Oceania, the Arctic, North America, South America, and Europe (AZTI, 2022). The core of the software is a database consisting of recently 10,638 entries in which each macrofauna organism is assigned a so-called Eco-Group (EG). This EG is an indication of whether that organism occurs primarily in unpolluted sediments (e.g. *T. leali,* EG I) or heavily polluted sediments (e.g. *C. capitata,* EG V). Organisms that occur in transient conditions can be assigned to EG II, III, or IV, ascending with an increasing degree of resistance to environmental pollution. Organisms that occur nonspecifically are not considered as bioindicators and are therefore not assigned an EG. By entering the taxonomically identified species and their respective abundance in the software, an equation based on weighted proportions (*Eq.1*) is used to calculate the AMBI value, inferring the ecological quality (EQ), and thus indicating the level of pollution (Muxika et al., 2005).

$$AMBI = \frac{\{(0 * \%EGI) + (1.5 * \%EGII) + (3 * \%EGIII) + (4.5 * \%EGIV) + (6 * \%EGV)\}}{100} \quad \textit{(Eq.1)}$$

Where: %EGI-%EGV is the proportion of organisms categorized into the corresponding Eco-Group

The resulting AMBI can take values between 0 and 6. If the calculated AMBI is lower than 1.2, the examined environment is classified as ‘unpolluted’, representing a high EQ. Contrary, if the calculated AMBI is above 5.5, the EQ is assumed to be low, representing a ‘heavily polluted’ environment. As an alternative measure of ecosystem health, the United Kingdom (UK) has included an extension of the AMBI, the IQI (Infaunal Quality Index), in its legislation (Phillips et al., 2014; UKMMAS, 2014). For IQI inference, the traditional AMBI is determined and supplemented with additional parameters such as total species count and evenness, a measure of species distribution. Moreover, a reference sample under pristine conditions is factored as a gold standard. The newly generated IQI can take values between 0 and 1, from which the EQ can then be inferred. According to UK legislation, the EQ of an environment must be ‘high’ or ‘good’, which corresponds to an IQI value above 0.64. Under these circumstances, the disturbance of the natural habitat is expected to be ‘none’ to ‘slight’, compared to disturbances of ‘major’ to ‘severe’ nature which are assumed if IQI values are below 0.64 (Phillips et al., 2014). More alternative indices exist worldwide that are used to a greater or lesser extent, such as the multivariate AMBI (mAMBI, Muxika et al., 2007), the Infaunal Trophic Index ITI (Word, 1978), the Norwegian Sensitivity Index NSI, or the Indicator Species Index ISI (Rygg and Norling, 2013).

All the above-mentioned indices have in common that they determine the EQ based on macrofauna composition. Therefore, each individual macrofauna organism needs to be taxonomically identified and its abundance needs to be ascertained. It is essential to identify the organisms down to species level since some species from the same family or genus differ in their tolerance to environmental influences (Aylagas et al., 2014; Carignan and Villard, 2002). Exemplary, different sea snail species of the genus *Nassarius* can be found in the AMBI database, indicating contrasting EGs for ecological reasons. From the two almost identical looking snail species, *N. lima* is assigned to EG I, while *N. mendicus* is assigned to EG IV. Considering all 48 species of *Nassarius* listed in the database, EG II and EG III have also been assigned. Misidentifications can therefore easily lead to errors in the evaluation of the EQ (Martinez-Crego et al., 2010), sometimes also induced by the occurrence of different life stages of the organisms (Darling and Mahon, 2011). To be able to correctly identify the organisms, it takes staff with great expertise (Schander and Willassen, 2005) and lots of time as it can take up to months before the EQ can be assessed (Aylagas et al., 2014).

As a result, adjustments to operating processes by salmon farm operators cannot be timely and are therefore not effective (Danovaro et al., 2016; Lejzerowicz et al., 2015).

eDNA metabarcoding

To circumvent these difficulties, a new strategy using metabarcoding of environmental DNA (eDNA) has been applied in recent years, which is no longer dependent on morphological macrofauna identification (Aylagas et al., 2014; Pawlowski et al., 2014). Using genetic information, eDNA metabarcoding is a molecular technique allowing estimations on the environmental conditions (*Figure 1*). It is easily upscalable, as it allows parallel identification of multiple taxa among different samples (Bohmann et al., 2014; Taberlet et al., 2012; Valentini et al., 2016).

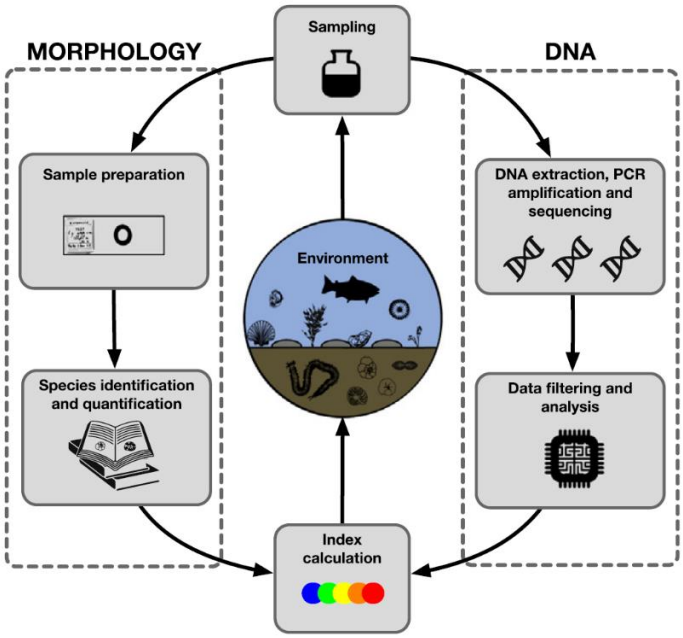


Figure 1) Schematic representation of the eDNA bioassessment method. *After sampling, the EQ of the environment can be inferred using either the traditional, morphology-based approach (left) or the novel eDNA-based metabarcoding approach (right). For the morphology-based approach, species have to be prepared, identified, and quantified, requiring personnel with taxonomic expertise. The molecular approach is based on eDNA extraction, amplification, and sequencing. From eDNA, taxa can be inferred bio-informatically and an index representing the EQ of the environment can be calculated. From Pawlowski et al. (2018)*

After extraction of multispecies eDNA from a variety of sample materials like water, stool, or sediment, a specific, highly conserved gene region is selected and amplified via polymerase chain reaction (PCR). This process is conducted using specialized artificial nucleic acids which can be referred to as group-specific primers. The selected gene region is referred to as a metabarcode or marker (*Figure 2*). Such genetic metabarcodes are required to contain highly conserved gene regions across all target organisms, which enable the annealing of the group-specific primers.

Therefore, they can be used for targeting the desired taxonomic group (Taberlet et al., 2018). Commonly used metabarcodes are the cytochrome-c-oxidase gene COI in metazoans (Elbrecht and Leese, 2015; Hebert et al., 2003), the internal transcribed spacer region ITS in fungi (Schoch et al., 2012), the small subunit SSU 18S rRNA region in eukaryotes (Hino et al., 2016), and the small subunit SSU 16S rRNA region in prokaryotes (Mizrahi-Man et al., 2013).

Metabarcodes are additionally required to contain a hypervariable gene region which ensures the taxonomic identification of individual organisms after sequencing (Taberlet et al., 2018). Hypervariable regions are for example the V4 region of the 18S SSU rRNA gene in eukaryotes (Cordier et al., 2019a; Forster et al., 2015) or the V3-V4 region of the 16S SSU rRNA gene in prokaryotes (Cordier et al., 2019a; Dowle et al., 2015; Frühe et al., 2020; Klemetsen et al., 2019; Stoeck et al., 2018a). Via amplification of the eDNA using the group-specific primers, the metabarcode region is extracted and multiplied, resulting in so-called amplicons. These amplicons are subsequently passed to sequencing.

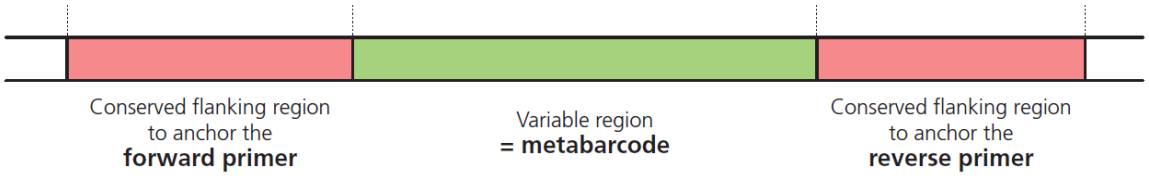


Figure 2) Example of a genetic metabarcoding region. *Prerequisites for the gene region which can be used for metabarcoding include containing a variable region (in green) flanked by group-specific, conserved regions (in red) which are used for primer annealing during PCR. From Taberlet et al. (2018), modified.*

The most commonly applied sequencing strategy of the next-generation sequencing (NGS) also known as high-throughput sequencing (HTS) technology is Illumina sequencing (Slatko et al., 2018). This approach is based on massive parallel sequencing, meaning it can provide simultaneous sequencing of hundreds to thousands of samples (Goodwin et al., 2016; Kozich et al., 2013; Taberlet et al., 2018). Specific nucleotide anchors are attached to a sequencing surface, the flow cell, to which the amplicons to be sequenced are added (Buermans and den Dunnen, 2014). The previously generated amplicons are additionally equipped with so-called Illumina tags, which function as flow cell binding sites as they interfere with the anchors attached to the flow cell. By using the so-called ‘sequencing by synthesis’ (SBS) strategy, the eDNA is amplified via bridge amplification resulting in clusters of identical molecules.

During the elongation process, light signals are emitted which are used for the identification of the progression of the nucleotides (Taberlet et al., 2018). The process is easily upscalable, as additional identifier nucleotide tags can also be appended, a process which is referred to as multiplexing (Meyer et al., 2007; Nielsen et al., 2006). These tags can be identified after sequencing by using bioinformatic tools to recover individual samples which is referred to as demultiplexing (Illumina, 2017; Renaud et al., 2015).

After demultiplexing, subsequent bioinformatic applications are used to trim extant primers from the desired sequences and to filter sequences by desired quality before grouping them into operational categories. In recent years, algorithms regarding categorical sequencing errors, such as the DADA2 algorithm, evolved as useful tools towards high-resolution sequence categorization (Amir et al., 2017; Callahan et al., 2016; Edgar, 2016). The tools can be used for initial length trimming and quality filtering of sequences, although most notably, they can determine expected sequencing error rates for each independent sequencing run, and therefore identify erroneous sequences. Exemplary, DADA2 corrects and clusters high-quality (HQ) sequences to amplicon sequence variants (ASVs) which act as operational units and can be used for downstream bioinformatic applications (Callahan et al., 2016). ASVs offer a higher resolution and sensitivity compared to traditional operational taxonomic units (OTUs) constructed using earlier tools (Callahan et al., 2017; Prodan et al., 2020). The desired output from DADA2 bioinformatic processing is an ASV-to-sample matrix in which the sequence counts per sample have been itemized for each ASV. This type of inventory can then be used for downstream bioinformatic pipelines and statistics.

Using such metabarcoding-based inventories to infer the composition of the community of organisms has already been suggested several times, as the community composition can faithfully reflect morphology-based indices (Lejzerowicz et al., 2015; Pawlowski et al., 2014). Applying molecular data for environmental biomonitoring was described as '*bigger, better and faster*' (Dafforn et al., 2014) than the traditional macrofauna identification and counting protocols and has therefore been recommended on several occasions for marine coastal biomonitoring (Aylagas et al., 2016; Lejzerowicz et al., 2015; Pochon et al., 2015; Stoeck et al., 2018a; 2018b; Valentini et al., 2016). Furthermore, it has been repeatedly shown that bacteria are particularly well suited for the evaluation of the EQ of the ecosystem (Aylagas et al., 2021; Dowle et al., 2015; Keeley et al., 2018; Stoeck et al., 2018a).

Due to their short generation times, they can react quickly to environmental influences, even faster than metazoans, and thus mirror the ecological status in a timely manner (Lear et al., 2011; Nogales et al., 2011). A schematic representation of changing bacterial communities along an enrichment gradient underneath a salmon farm net-pen system is presented in *Figure 3*.

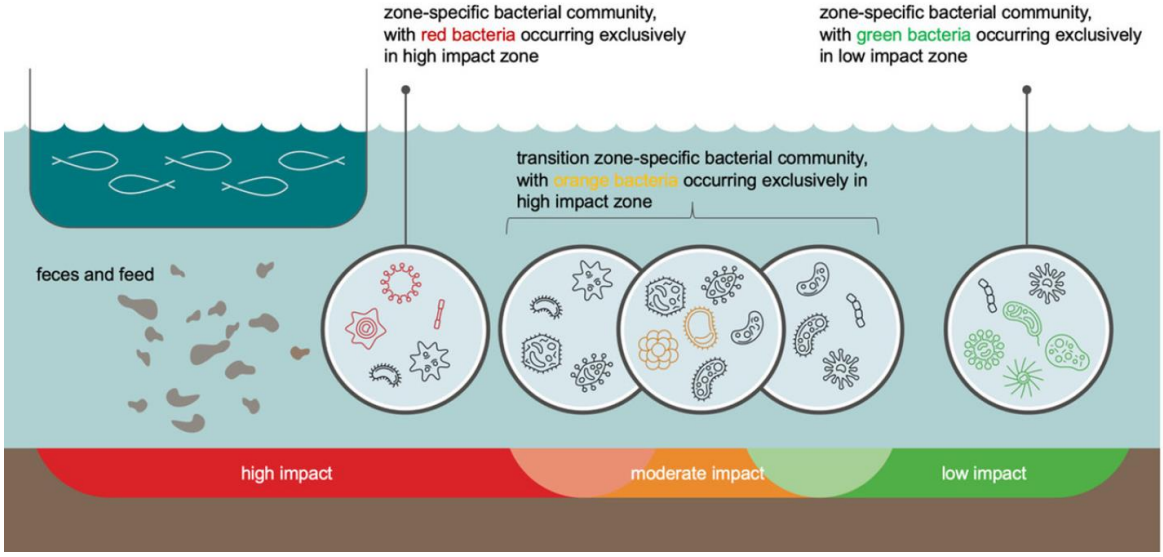


Figure 3) Schematic representation of bacterial inventories mirroring aquaculture impact in open net-pen systems. *Fish feces and leftover feed accumulate beneath the fish cages inducing a high impact on the marine sediment. From left to right, a decreasing organic enrichment gradient along the prevailing current is generated. Pristine-like conditions can be found in the low-impacted site on the right, while a moderate impact of organic enrichment is verified for transitional zones. The high-impact zone on the left side shows the greatest extent of impact and alteration of bacterial communities. From Frühe et al 2021, modified.*

Supervised machine learning applications in the field

A recent gamechanger in various fields of science including medicine, information technology, or environmental protection is supervised machine learning (SML; Ahmed et al., 2021; Elmogy et al., 2021; Muhammad et al., 2021; Osarogiagbon et al., 2021; Wilhelm et al., 2022). SML is a type of statistical learning that uses labeled data to develop models that can then be applied to new data (Hastie et al., 2009b). For biomonitoring, it has already been shown that SML can predict the ecological influences based on eDNA-based metabarcoding (Cordier et al., 2017; Cordier et al., 2019b; Frühe et al., 2020; Smith et al., 2015). The big advantage of this method is that environmental evaluation can be conducted taxonomy-free, as it is not dependent on the completeness of curated entries in databases. Databases are mostly incomplete, so many organisms can be found of which the classification into opportunists/specialists or similar is not known (Chariton et al., 2010; Lanzén et al., 2016; Lejzerowicz et al., 2015).

In 2015, Smith et al. tested such a taxonomy-free approach where bacterial eDNA was used to distinguish anthropogenic polluted areas from unpolluted areas. SML was able to detect pollution from nitrate (e.g., deriving from fertilizers), uranium (e.g., deriving from nuclear waste), and oil (e.g., deriving from oil spills). Tree-based SML methods such as Random Forest (RF) were found to give the best results predicting up to 98% of the samples correctly, which indicates an almost perfect agreement between predicted values and reference values. In 2017, Cordier et al. tied to the previous study by expanding the scope of human impact by including the aquaculture industry as a factor of environmental impact. The authors demonstrated that SML based on HTS amplicon eDNA sequencing can predict the EQ of the habitat under surveillance very well. The authors were able to build robust, taxonomy-free models for the prediction of the status based on foraminifers, which are also known to react to marine pollution (Alve et al., 2016; Frontalini and Coccioni, 2011). For model building, they used EQ measures gained from macroinvertebrate data as a reference. They predicted the ecological quality using foraminiferal operational taxonomic units in a RF regression model analysis. The authors pointed out that SML can create robust models which are independent of taxonomic databases, suggesting eDNA-based SML to infer biotic indices as an alternative for traditional benthic biomonitoring. Further, they confirmed that SML was able to handle the highly dimensional eDNA dataset containing rare sequences because they used the RF algorithm to prevent overfitting (Breiman, 2001; Hastie et al., 2009a). Around one year later, the authors published another study comparing different genetic markers of foraminifers, eukaryotes, and bacteria. They demonstrated that each RF model based on high-resolution sequence data can predict the ecological status of a sample better than the taxonomy-based data, which paved the way for taxonomy-free RF approaches into eDNA-based EQ assessments (Cordier et al., 2018).

Recently, Frühe et al. (2020) confirmed that SML outperforms another taxonomy-free method for biomonitoring attempts, the Indicator Value (IndVal) approach developed by Dufrene and Legendre (1997). They compared molecular AMBI values, which were inferred using IndVal or SML, against macrofauna-based AMBI values as a benchmark. They found SML to be less sensitive to noisy eDNA metabarcoding data and uneven data coverage among categories compared to the IndVal approach. They could also show that bacterial eDNA metabarcodes outperform ciliate eDNA metabarcodes regarding their prediction accuracies for EQ. Thus, Frühe et al. (2020) recommend that metabarcodes derived from bacterial eDNA should be evaluated with SML to assess the EQ of the environment.

The Random Forest algorithm in particular

It was formerly assumed that a particularly high predictive performance can be expected using RF for large and complex ecological datasets (Fox et al., 2017; Freeman et al., 2015; Gislason et al., 2006; Prasad et al., 2006). Among others, the previously mentioned studies could show that tree-based methodologies like RF are suitable to analyze metabarcoding-based datasets for this reason (Cordier et al., 2017; Gerhard and Gunsch, 2019; Smucker et al., 2020). The RF algorithm is a part of SML, as it is a supervised ensemble learning method (James et al., 2013). RF is based on classification and regression decision trees (CARTs) which are suitable for building classification models for categorical predictions, as well as regression models to predict continuous values (Breiman, 2001; Hastie et al., 2009a). The fundamental concept of the RF approach consists of using predictor variables, which can be also referred to as explanatory variables or features, to predict response variables also known as labels.

The first step of each SML approach consists of learning explanatory variables while a reference outcome, considered as a response variable or a label, is also given for each observation. This dataset is called the training dataset. Learning results in the construction of a model, which can then be applied for the prediction of a new dataset containing explanatory variables with unknown labels (Breiman, 2001). For evaluation of the predictive power of the constructed model, this new dataset is an artificial testing dataset containing samples with known but removed reference labels. During each RF learning, CARTs are constructed and passed by predictor variables. At each split of the CART, a specific number *m* of random predictors is used. The default *m* value is the number of predictors divided by three for regression analysis, or the square root of the number of predictions for classification analysis (Liaw and Wiener, 2002). After the passing of predictor variables through the CART, terminal nodes suggest a vote for the prediction outcome. The concept of RF incorporates a step from the single ‘tree’ to the combined ‘forest’: based on bagging, also known as bootstrap aggregation or majority voting, each tree holds one decision, and all the trees together influence the outcome (Breiman, 2001). For classification approaches, the class voted for by the majority of the trees, is chosen as the final prediction. For regression analysis, the mean value of all trees is aggregated and averaged, enabling variance reduction (*Figure 4*).

For the evaluation of a classification model, the prediction accuracy can be inferred by comparing predicted labels to reference labels. For regression tasks, another measure of model fit, the root mean squared error (RMSE), is used (Hastie et al., 2009b). Additionally, the explanatory variables of the training dataset are analyzed for their capability to foretell the response variable at each split in the CART decision trees, resulting in a variable importance measure (Hastie et al., 2009b). Therefore, formerly left out samples, referred to as out-of-bag (OOB) samples, are passed down the trees with features randomly permuted. Comparing prediction accuracies and RMSE errors, the variable importance for each explanatory variable can then be inferred (Hastie et al., 2009a).

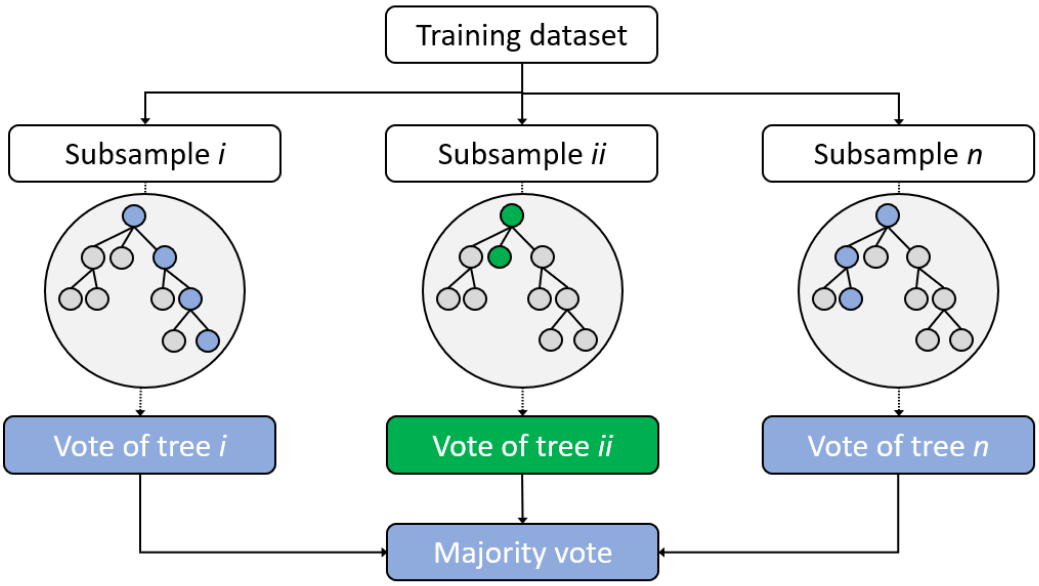


Figure 4) Schematic representation of the RF algorithm. *A dataset is randomly split into n subsamples constructing various decision trees. After the variables have been passed along the trees, each tree has its own vote (tree i: 'blue', tree ii: 'green', tree n: 'blue'). RF predictions are built based on bagging, as the votes of each decision tree are aggregated into a majority vote (majority vote: 'blue').*

Another common terminology concerning RF is cross-validation (CV). CV is important for complex datasets because it reduces variance and can therefore help to improve the model (Hastie et al., 2009b). The concept of CV describes the bootstrap sampling of the training dataset into subsamples, meaning for each training iteration *k*, different observations are being used for RF model construction (Hastie et al., 2009a). Hence, this CV approach is named *k*-fold CV depending on the number of iterations (*Figure 5*).

A popular type of CV in the field of molecular biology is the leave-one-out approach (LOO-CV; Cordier et al., 2018; Gerhard and Gunsch, 2019). When using LOO-CV, the training dataset is subsampled the number of times as it contains samples. For each model construction, one sample is left out, while the model is built based on the information of all remaining samples. In the next training step, the formerly left out sample is included again, but therefore, another sample is excluded. The new model construction relies on the set of remaining samples. Therefore, the number of CV iterations *k* corresponds to the total number of samples in a LOO approach. LOO-CV is often used when the predictive power for new samples with unknown labels should be anticipated (Hastie et al., 2009a).

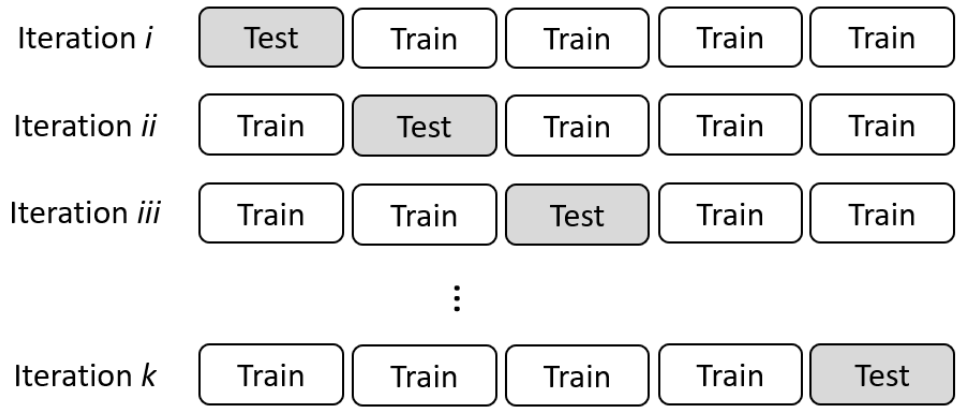


Figure 5) Schematic representation of the k-fold CV approach. *CV is based on subsampling steps of the dataset which are called iterations. As k represents the number of iterations, it is called k-fold CV. For the first iteration, the dataset is split into a testing and a training dataset. The model is constructed using the training data and evaluated using the testing data. For the second iteration, another subsample from the dataset is chosen as test data. The remaining data is used for model training. If five iterations are performed, a so-called five-fold CV is being conducted. If the number of iterations corresponds to the total number of samples, a special type of CV, referred to as LOO-CV, is conducted.*

Contrary to RF, noisy datasets consisting of many explanatory variables in complex relationships often trigger overfitting in ordinary regression models (Babyak, 2004; Faraway, 2004). Overfitting results in a prediction relying on the exact set of data recognition rather than actual learning of variables. It can be circumvented by using a majority vote approach like RF (Hastie et al., 2009b). Ecological datasets, such as datasets inferred by eDNA metabarcoding, are often highly dimensional and contain noise which can be referred to as a lot of variables in nonlinear relationships, therefore RF applications on those kinds of datasets are advised (Cutler et al., 2007; Evans et al., 2011).

A further approach: Quantile Regression Splines

A further promising approach for using eDNA metabarcoding for environmental quality inference was introduced by Keeley et al. in 2018. As in previous metabarcoding-based studies, the eDNA was extracted, barcoded, sequenced, and bioinformatically processed. For the molecular inference of ecological quality, the authors developed a novel index. This index is multitrophic, meaning that it considers eDNA from different organisms such as foraminifera, eukaryotes, and bacteria simultaneously. Following a publication by Anderson (2008), a Quantile Regression Spline (QRS) analysis was performed to anticipate organic enrichment gradient stages in New Zealand marine sediments instead of using SML. QRS are based on changes in taxonomic distributions as a function of the enrichment stage. In this approach, taxonomic units are assigned a specific Eco-Group (EG) regarding their occurrence along a pollution gradient (compare to EG in Borja et al., 2000). With the ratios of the different inferred EG per taxonomic unit, a new metabarcoding biotic index can be calculated representing the EQ. This outcome can then be evaluated by comparing it to the macrofauna inferred reference ecological status. The most promising results were obtained with the combination of bacterial and eukaryotic eDNA, which together showed a concordance with the reference EQs of $R^2 > 0.9$. The authors suggest that in the future, the QRS method could augment or even replace current biomonitoring efforts that monitor the impact of fish farms on the surrounding environment.

Towards standard operating procedures

Summarizing all previous studies on eDNA metabarcoding as a biomonitoring tool, a heterogeneity of applications is quickly apparent which makes it difficult to include this high-potential method in legislation (Goldberg et al., 2016). Calls for standardization in terms of standard operating procedures (SOP) are being made by researchers and industry to form the basis for the implementation of eDNA-based methods into regulatory legislature (Goldberg et al., 2016; Helbing and Hobbs, 2019; Kelly et al., 2014). Constant collaboration between researchers and regulators is needed to support the confidence of eDNA-based methods (Helbing and Hobbs, 2019). A special conference was held in 2018 to help drive the standardization of eDNA, the ‘Pathway to Increase Standards and Competency of eDNA Surveys’ (PISCeS) conference.

One of the main outcomes of the symposium was the demonstration of the collaboration needs of academics, regulators, and industry, leading to a publication of a special issue regarding eDNA metabarcoding standardization approaches (Loeza-Quintana et al., 2020; Morey et al., 2020; Skinner et al., 2020). However, the best practices for survey design, sample collection, sample preservation, eDNA isolation, eDNA analysis, statistical analysis, and interpretation need to be discussed and harmonized (Helbing and Hobbs, 2019). Although, best practices may differ between the system to be surveyed and target species (Loeza-Quintana et al., 2020). For the implementation of eDNA-based methods into routine compliance monitoring, technical restrictions of the eDNA monitoring method must be identified to ensure reproducibility. Only if robust ecological evaluations of the ecosystems under surveillance can be achieved, the holistic method has the potential to be included into regulatory frameworks. Therefore, SOPs are an important step towards the implementation of eDNA metabarcoding-based methods into environmental biomonitoring regulations, determining how it could be used to complement or even replace traditional macrofauna-based monitoring. For this reason, this thesis focuses on topics of possible technical restricting factors that could have an impact on the overall ecological evaluation via DNA-based monitoring, namely the sample preservation method, the molecular lab treatment reproducibility, and the ecological evaluation method. It has to be determined to what extent various limitations exist and if assessments of EQ based on eDNA metabarcoding are sufficiently robust and reproducible.

Outline of this thesis

The first chapter of this thesis focuses on the reproducibility of the eDNA metabarcoding-based ecological evaluation when different preservation strategies are used for sample transportation to the molecular laboratory. Preservation is necessary to prevent eDNA alteration, which can occur through nuclease-induced degradation, mechanical influences, or spontaneous chemical reactions. In current studies regarding eDNA metabarcoding-based biomonitoring, two main methods for sample preservation are used. Either the samples are combined with a preservation solution, which, according to the manufacturer, protects nucleic acids from degradation, or the samples are frozen in a timely manner after sampling.

In **Chapter I, 'Sample preservation for transport'**, replicate biological samples have been subjected to both preservation methods independently and have been compared regarding their subsequent evaluation of EQ. The results of the study were published in:

Dully, V., Rech, G., Wilding, T.A., Lanzén, A., MacKichan, K., Berrill, I, & Stoeck, T. (2021). Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems. *Marine Pollution Bulletin 173*, e113129. doi:10.1016/j.marpolbul.2021.113129

In the second part of this thesis, another component crucial to the successful implementation of the eDNA method into SOPs, namely reproducibility of amplicon generation via PCR and subsequent Illumina sequencing, is addressed. Only if PCR and Illumina sequencing are reproducible among independent analyses, eDNA metabarcoding can be implemented into routine biomonitoring frameworks. In **Chapter II, 'Reproducibility of eDNA metabarcoding-based sample processing'**, a study was conducted comparing ecological interpretations among two independent laboratories which processed replicate samples using the same molecular protocol. Therefore, amplicons have been constructed independently via PCR and were then sequenced on two independent Illumina runs. To assess the overall robustness of the method, the obtained results among laboratories were compared statistically and via SML. The findings of the study were published in:

Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., & Stoeck, T. (2021). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environ-mental biomonitoring tool in coastal salmon aquaculture – An inter laboratory study. *Ecological Indicators 121*, e107049. doi:10.1016/j.ecolind.2020.107049

In order to achieve most reliable biomonitoring results, the conversion of eDNA into the actual environmental assessment, a process which is referred to as EQ inference, is required to be optimized. In **Chapter III, 'Robustness and applicability of current approaches for EQ inference'**, the newly introduced QRS method following Keeley et al. (2018), was tested against an SML approach. The comparison of these current EQ inference methods was conducted to determine which is the best method to develop a universally applicable monitoring framework.

Therefore, two huge salmon farm sediment datasets which have derived from Norway and Scotland were analyzed independently. Besides the predictive performance of the model, inferred indicators were compared between the methods. The advantages and disadvantages of each EQ inference application are discussed, and there is a recommendation for implementation in SOPs.

Chapter IV, 'Required sequencing depth for SML-based EQ inference', describes an optimization study that was conducted to enhance the eDNA-based SML method. The aim was to anticipate to what extent sequencing processes and therefore computational resources can be reduced without losing informational value for eDNA-based EQ inference. Therefore, holistic eDNA datasets, displaying various influences of urban structures, aquaculture, and shipping traffic, have been analyzed regarding their ecological interpretations based on SML. The predictive performance has been subsequently compared to the results based on artificially reduced sequence numbers. Thus, this study helps to understand to what extent SML-based ecological evaluations of the marine coastal ecosystems are dependent on sequencing depth. The results of the study were published in:

Dully, V., Wilding, T. A., Mühlhaus, T., & Stoeck, T. (2021).
Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning.
Computational and Structural Biotechnology Journal 19, 2256-2268.
doi:10.1016/j.csbj.2021.04.005

References

Ahmed, K. R., Akter, S., Marandi, A., & Schüth, C. (2021). A simple and robust wetland classification approach by using optical indices, unsupervised and supervised machine learning algorithms. *Remote Sensing Applications: Society and Environment, 23*, e100569. doi:10.1016/j.rsase.2021.100569

Alve, E., Korsun, S., Schönfeld, J., Dijkstra, N., Golikova, E., Hess, S., Husum, K., & Panieri, G. (2016). Foram-AMBI: A sensitivity index based on benthic foraminiferal faunas from North-East Atlantic and Arctic fjords, continental shelves and slopes. *Marine Micropaleontology, 122*, 1-12. doi:10.1016/j.marmicro.2015.11.001

Amir, A., McDonald, D., Navas-Molina Jose, A., Kopylova, E., Morton James, T., Zech Xu, Z., et al. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems, 2*(2), e00191-00116 doi:10.1128/mSystems.00191-16

Anderson, M. (2008). Animal-sediment relationships re-visited: Characterising species' distributions along an environmental gradient using canonical analysis and quantile regression splines. *Journal of Experimental Marine Biology and Ecology, 366*, 16-27. doi:10.1016/j.jembe.2008.07.006

Aylagas, E., Atalah, J., Sánchez-Jerez, P., Pearman, J. K., Casado, N., Asensi, J., Toledo-Guedes, K., & Carvalho, S. (2021). A step towards the validation of bacteria biotic indices using DNA metabarcoding for benthic monitoring. *Molecular Ecology Resources, 21*(6), 1889-1903. doi:10.1111/1755-0998.13395

Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Frontiers in Marine Science, 3*(96). doi:10.3389/fmars.2016.00096

Aylagas, E., Borja, A., & Rodriguez-Ezpeleta, N. (2014). Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). *Plos One, 9*(3), e90529. doi:10.1371/journal.pone.0090529

AZTI (2022). AMBI by AZTI, AZTI's Marine Biotic Index. *Retrieved on 15.03.2022 from https://ambi.azti.es/*

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411-421. doi:10.1097/01.psy.0000127692.23278.a9

Bannister, R. J., Valdemarsen, T., Hansen, P. K., Holmer, M., & Ervik, A. (2014). Changes in benthic sediment conditions under an Atlantic salmon farm at a deep, well-flushed coastal site. *Aquaculture Environment Interactions, 5*(1), 29-47. doi:10.3354/aei00092

Barbier, E. B., Hacker, S. D., Kennedy, C., Koch, E. W., Stier, A. C., & Silliman, B. R. (2011). The value of estuarine and coastal ecosystem services. *Ecological Monographs, 81*(2), 169-193. doi:10.1890/10-1510.1

Borja, A., Franco, J., & Pérez, V. (2000). A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments. *Marine Pollution Bulletin, 40*(12), 1100-1114. doi:10.1016/S0025-326X(00)00061-8

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Brown, J. R., Gowen, R. J., & McLusky, D. S. (1987). The effect of salmon farming on the benthos of a Scottish sea loch. *Journal of Experimental Marine Biology and Ecology, 109*(1), 39-51. doi:10.1016/0022-0981(87)90184-5

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1842*(10), 1932-1941. doi:10.1016/j.bbadis.2014.06.015

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal, 11*(12), 2639-2643. doi:10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Carignan, V., & Villard, M.-A. (2002). Selecting Indicator Species to Monitor Ecological Integrity: A Review. *Environmental Monitoring and Assessment, 78*(1), 45-61. doi:10.1023/A:1016136723584

Carroll, M. L., Cochrane, S., Fieler, R., Velvin, R., & White, P. (2003). Organic enrichment of sediments from salmon farming in norway: environmental factors, management practices, and monitoring techniques. *Aquaculture, 226,* 165-180. doi:10.1016/S0044-8486(03)00475-7

Chariton, A. A., Court, L. N., Hartley, D. M., Colloff, M. J., & Hardy, C. M. (2010). Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment, 8*(5), 233-238. doi:10.1890/090115

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01518

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources, 18*(6), 1381-1391. doi:10.1111/1755-0998.12926

Cordier, T., Frontalini, F., Cermakova, K., Apothéloz-Perret-Gentil, L., Treglia, M., Scantamburlo, E., Bonamin, V., & Pawlowski, J. (2019a). Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Marine Environmental Research, 146,* 24-34. doi:10.1016/j.marenvres.2018.12.009

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019b). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology, 27*(5), 387-397. doi:10.1016/j.tim.2018.10.012

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for classification in Ecology. *Ecology, 88*(11), 2783-2792. doi:doi.org/10.1890/07-0539.1

Dafforn, K., Baird, D., Chariton, A., Sun, M., Brown, M., Simpson, S., Kelaher, B., & Johnston, E. (2014). Chapter One - Faster, Higher and Stronger? The Pros and Cons of Molecular Faunal Data for Assessing Ecosystem Condition. In G. Woodward, A. J. Dumbrell, D. J. Baird, & M. Hajibabaei (Eds.), *Advances in Ecological Research: Big Data in Ecology* (pp. 1-40). Essex: Academic Press, Elsevier Ltd.

Daily, G. C. (2013). Nature's Services: Societal Dependence on Natural Ecosystems (1997). In L. Robin, S. Sörlin, & P. Warde (Eds.), *The Future of Nature: Documents of Global Change* (pp. 454-464). London: Yale University Press.

Danovaro, R., Carugati, L., Berzano, M., Cahill, A. E., Carvalho, S., Chenuil, A., et al. (2016). Implementing and Innovating Marine Monitoring Approaches for Assessing Marine Environmental Status. *Frontiers in Marine Science, 3*, 213. doi:10.3389/fmars.2016.00213

Darling, J., & Mahon, A. (2011). From molecules to management: Adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research, 111*, 978-988. doi:10.1016/j.envres.2011.02.001

Dowle, E., Pochon, X., Keeley, N., & Wood, S. A. (2015). Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *Fems Microbiology Ecology, 91*(8), fiv089. doi:10.1093/femsec/fiv089

Drius, M., Bongiorni, L., Depellegrin, D., Menegon, S., Pugnetti, A., & Stifter, S. (2019). Tackling challenges for Mediterranean sustainable coastal tourism: An ecosystem service perspective. *Science of the Total Environment, 652*, 1302-1317. doi:10.1016/j.scitotenv.2018.10.121

Dufrene, M., & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs, 67*(3), 345-366. doi:10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2

Edgar, R. (2016). Preprint: UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. doi:10.1101/081257

EEA (2018). European Environment Agency. Indicator Assessment - Aquaculture production in Europe. *Retrieved on 15.02.2022 from www.eea.europa.eu/data-and-maps/indicators/aquaculture-production-4/assessment*

Elbrecht, V., & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass-Sequence Relationships with an Innovative Metabarcoding Protocol. *Plos One, 10*(7), e0130324. doi:10.1371/journal.pone.0130324

Elmogy, A. M., Tariq, U., Mohammed, A., & Ibrahim, A. (2021). Fake Reviews Detection using Supervised Machine Learning. *International Journal of Advanced Computer Science and Applications, 12*(1), 301-606. doi:10.14569/IJACSA.2021.0120169

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2011). Modeling Species Distribution and Change Using Random Forest. In C. A. Drew, Y. F. Wiersma, & F. Huettmann (Eds.), *Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications* (pp. 139-159). New York: Springer.

FAO (1996). Food and Agriculture Organization of the United Nations. The contributions of science to integrated coastal management, Joint Group of Experts on the Scientific Aspects of Marine Environmental Protection (GESAMP). *Retrieved on 21.01.2022 from http://www.gesamp.org/site/assets/files/1239/the-contributions-of-science-to-integrated-coastal-management-en.pdf.*

FAO (2020). Food and Agriculture Organization of the United Nations. The State of World Fisheries and Aquaculture 2020. *Retrieved on 03.03.2022 from https://www.fao.org/publications/card/en/c/CA9229EN.*

Faraway, J. J. (2004). *Linear models with R*. New York: Chapman and Hall/CRC.

Forrest, B. M., Keeley, N., Gillespie, P., Hopkins, G., Knight, B., & Govier, D. (2007). Review of the Ecological Effects of Marine Finfish Aquaculture: Final Report. The Ministry of Fisheries, New Zealand. *Retrieved on 18.01.2022 from https://www.yumpu.com/en/document/read/20615870/review-of-the-ecological-effects-of-marine-finfish-aquaculture-final*

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., Lopez, P., Stoeck, T., & Bapteste, E. (2015). Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *Bmc Biology, 13*(1), 16. doi:10.1186/s12915-015-0125-5

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment, 189*(7), 316. doi:10.1007/s10661-017-6025-0

Freeman, E. A., Moisen, G. G., Coulston, J. W., & Wilson, B. T. (2015). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Canadian Journal of Forest Research, 46*(3), 323-339. doi:10.1139/cjfr-2014-0562

Frontalini, F., & Coccioni, R. (2011). Benthic foraminifera as bioindicators of pollution: a review of Italian research over the last three decades. *Revue de micropaléontologie, 54*(2), 115-127. doi:10.1016/j.revmic.2011.03.001

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

Gerhard, W., & Gunsch, C. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environment International, 124*, 312-319. doi:10.1016/j.envint.2018.12.038

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300. doi:10.1016/j.patrec.2005.08.011

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307. doi:10.1111/2041-210X.12595

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351. doi:10.1038/nrg.2016.49

Grall, J., & Glémarec, M. (1997). Using biotic indices to estimate macrobenthic community perturbations in the Bay of Brest. *Estuarine, Coastal and Shelf Science, 44*, 43-53. doi:10.1016/S0272-7714(97)80006-6

Gray, J. S. (1997). Marine biodiversity: patterns, threats and conservation needs. *Biodiversity & Conservation, 6*(1), 153-175. doi:10.1023/A:1018335901847

Hassan, R., Scholes, R., & Ash, N. (2005). Millennium Ecosystem Assessment Series. Ecosystems and Human Well-Being: Current State and Trends: Findings of the Condition and Trends Working Group. *Retrieved on 15.02.2022 from https://www.millenniumassessment.org/documents/document.766.aspx.pdf.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). Random forests. In *The elements of statistical learning* (2nd ed., pp. 587-604). New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009b). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Heal, G. M., Barbier, E. B., Boyle, K. J., Covich, A. P., Gloss, S. P., Carlton H. Hershner, J., et al. (2005). *Valuing Ecosystem Services: Toward Better Environmental Decision-Making*. Washington DC: The National Academies Press.

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society - Biological Sciences, 270*(1512), 313-321. doi:10.1098/rspb.2002.2218

Helbing, C. C., & Hobbs, J. (2019). Environmental DNA standardization needs for fish and wildlife population assessments and monitoring. Canadian Standards Association. *Retrieved on 10.02.2022 from https://www.csagroup.org/wp-content/uploads/CSA-Group-Research-Environmental-DNA.pdf.*

Hino, A., Maruyama, H., & Kikuchi, T. (2016). A novel method to assess the biodiversity of parasites using 18S rDNA Illumina sequencing; parasitome analysis method. *Parasitology International, 65*(5), 572-575. doi:10.1016/j.parint.2016.01.009

Hintermeier-Erhard, G., & Zech, W. (1997). *Wörterbuch der Bodenkunde - Systematik, Genese, Eigenschaften, Ökologie und Verbreitung von Böden*. Stuttgart: Springer Spektrum.

Holmer, M., Wildish, D., & Hargrave, B. T. (2005). Organic Enrichment from Marine Finfish Aquaculture and Effects on Sediment Biogeochemical Processes. In B. T. Hargrave (Ed.), *Environmental Effects of marine Finfish Aquaculture* (pp. 182-206). New York: Springer.

Hughes, S. (1975). *National environmental policy act of 1969*. Washington, D.C.: Congressional Research Service, Library of Congress.

Illumina (2017). Metagenomic Sequencing Library Preparation. *Retrieved on 01.01.2022 from https://www.illumina.com/content/dam/illumina-marketing/documents/ products/ illumina_sequencing_introduction.pdf.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Karr, J. R. (1991). Biological Integrity: A Long-Neglected Aspect of Water Resource Management. *Ecological Applications, 1*(1), 66-84. doi:10.2307/1941848

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *Plos One, 9*(1), e86175. doi:10.1371/journal.pone.0086175

Klemetsen, T., Willassen, N., & Karlsen, C. (2019). Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis. *MicrobiologyOpen, 8*. doi:10.1002/mbo3.898

Kobayashi, M., Msangi, S., Batka, M., Vannuccini, S., Dey, M. M., & Anderson, J. L. (2015). Fish to 2030: The Role and Opportunity for Aquaculture. *Aquaculture economics & management, 19*(3), 282-300. doi:10.1080/13657305.2015.994240

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology, 79*(17), 5112-5120. doi:10.1128/aem.01043-13

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology, 25*(17), 4392-4406. doi:10.1111/mec.13761

Lear, G., Dopheide, A., Ancion, P., & Lewis, G. D. (2011). A comparison of bacterial, ciliate and macroinvertebrate indicators of stream ecological health. *Aquatic Ecology, 45*(4), 517-527. doi:10.1007/s10452-011-9372-x

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D., & Pawlowski, J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports, 5*, e13932 doi:10.1038/srep13932.

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R news, 2(3)*, 18-22.

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

Lotze, H. K., Lenihan, H. S., Bourque, B. J., Bradbury, R. H., Cooke, R. G., Kay, M. C., et al. (2006). Depletion, degradation, and recovery potential of estuaries and coastal seas. *Science, 312*(5781), 1806-1809. doi:10.1126/science.1128035

Markert, B., Wappelhorst, O., Weckert, V., Herpin, U., Siewers, U., Friese, K., & Breulmann, G. (1999). The use of bioindicators for monitoring the heavy-metal status of the environment. *Journal of Radioanalytical and Nuclear Chemistry, 240*(2), 425-429. doi:10.1007/BF02349387

Martinez-Crego, B., Alcoverro, T., & Romero, J. (2010). Biotic indices for assessing the status of coastal waters: a review of strengths and weaknesses. *Journal of Environmental Monitoring, 12*(5), 1013-1028. doi:10.1039/b920937a

Mcleod, E., Chmura, G. L., Bouillon, S., Salm, R., Björk, M., Duarte, C. M., Lovelock, C. E., Schlesinger, W. H., & Silliman, B. R. (2011). A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO2. *Frontiers in Ecology and the Environment, 9*(10), 552-560. doi:10.1890/110004

Meyer, M., Stenzel, U., Myles, S., Prüfer, K., & Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research, 35*(15), e97. doi:10.1093/nar/gkm566

Miller, M. E., Belote, R. T., Bowker, M. A., & Garman, S. L. (2011). Alternative states of a semiarid grassland ecosystem: implications for ecosystem services. *Ecosphere, 2*(5), 55. doi:10.1890/ES11-00027.1

Mizrahi-Man, O., Davenport, E. R., & Gilad, Y. (2013). Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *Plos One, 8*(1), e53608. doi:10.1371/journal.pone.0053608

Morey, K. C., Bartley, T. J., & Hanner, R. H. (2020). Validating environmental DNA metabarcoding for marine fishes in diverse ecosystems using a public aquarium. *Environmental DNA, 2*(3), 330-342. doi:10.1002/edn3.76

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN computer science, 2*(1), 1-13. doi:10.1007/s42979-020-00394-7

Muxika, I., Borja, A., & Bald, J. (2007). Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. *Marine Pollution Bulletin, 55*(1-6), 16-29. doi:10.1016/j.marpolbul.2006.05.025

Muxika, I., Borja, A., & Bonne, W. (2005). The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. *Ecological Indicators, 5*(1), 19-31. doi:10.1016/j.ecolind.2004.08.004

Nielsen, K. L., Høgh, A. L., & Emmersen, J. (2006). DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Research, 34*(19), e133. doi:10.1093/nar/gkl714

Nogales, B., Lanfranconi, M. P., Piña-Villalonga, J. M., & Bosch, R. (2011). Anthropogenic perturbations in marine microbial communities. *FEMS Microbiology Reviews, 35*(2), 275-298. doi:10.1111/j.1574-6976.2010.00248.x

Osarogiagbon, A. U., Khan, F., Venkatesan, R., & Gillard, P. (2021). Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. *Process Safety and Environmental Protection, 147*, 367-384. doi:10.1016/j.psep.2020.09.038

Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources, 14*(6), 1129-1140. doi:10.1111/1755-0998.12261

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (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

Pearson, T., & Rosenberg, R. (1978). Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology, 16*, 229-311. doi:10.2983/035.034.0121u1.10

Phillips, G. R., Anwar, A., Brooks, L., Martina, L. J., Miles, A. C., & Prior, A. (2014). Infaunal quality index: Water Framework Directive classification scheme for marine benthic invertebrates. *Retrieved on 01.01.2022 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/314673/Water_Framework_Directive_classification_scheme_for_marine_benthic_invertebrates_-_report.pdf*

Pochon, X., Wood, S. A., Keeley, N. B., Lejzerowicz, F., Esling, P., Drew, J., & Pawlowski, J. (2015). Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Marine Pollution Bulletin, 100*(1), 370-382. doi:10.1016/j.marpolbul.2015.08.022

Prasad, A., Iverson, L., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems, 9*, 181-199. doi:10.1007/s10021-005-0054-1

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *Plos One, 15*(1), e0227434. doi:10.1371/journal.pone.0227434

Reish, D. J. (1955). The relation of polychaetous Annelids to harbor pollution. *Public health reports, 70*(12), 1168-1174. doi:10.2307/4589315

Renaud, G., Stenzel, U., Maricic, T., Wiebe, V., & Kelso, J. (2015). deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics, 31*(5), 770-772. doi:10.1093/bioinformatics/btu719

Rygg, B., & Norling, K. (2013). Norwegian Sensitivity Index (NSI) for marine macroinvertebrates, and an update of Indicator Species Index (ISI). *Retrieved on 23.02.2022 from http://hdl.handle.net/11250/216238.*

Sarà, G., Scilipoti, D., Mazzola, A., & Modica, A. (2004). Effects of fish farming waste to sedimentary and particulate organic matter in a southern Mediterranean area (Gulf of Castellammare, Sicily): A multiple stable isotope study. *Aquaculture, 234*, 199-213. doi:10.1016/j.aquaculture.2003.11.020

Sarà, G., Scilipoti, D., Milazzo, M., & Modica, A. (2006). Use of stable isotopes to investigate dispersal of waste from fish farms as a function of hydrodynamics. *Marine Ecology Progress Series, 313*, 261-270. doi:10.3354/MEPS313261

Schander, C., & Willassen, E. (2005). What can biological barcoding do for marine biology? *Marine Biology Research, 1*(1), 79-83. doi:10.1080/17451000510018962

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., & Chen, W. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America, 109*(16), 6241-6246. doi:10.1073/pnas.1117018109

Skinner, M., Murdoch, M., Loeza-Quintana, T., Crookes, S., & Hanner, R. (2020). A mesocosm comparison of laboratory-based and on-site eDNA solutions for detection and quantification of striped bass (Morone saxatilis) in marine ecosystems. *Environmental DNA, 2*(3), 298-308. doi:10.1002/edn3.61

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current protocols in molecular biology, 122*(1), e59. doi:10.1002/cpmb.59

Small, C., & Nicholls, R. J. (2003). A Global Analysis of Human Settlement in Coastal Zones. *Journal of Coastal Research, 19*(3), 584-599. doi:10.2307/4299200

Smith, M., B., Rocha, A., M., Smillie, C., S., Olesen, S., W., Paradis, C., Wu, L., et al. (2015). Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. *mBio, 6*(3), e00326-00315. doi:10.1128/mBio.00326-15

Smucker, N. J., Pilgrim, E. M., Nietch, C. T., Darling, J. A., & Johnson, B. R. (2020). DNA metabarcoding effectively quantifies diatom responses to nutrients in streams. *Ecological Applications, 30*(8), e02205. doi:10.1002/eap.2205

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., & Pawlowski, J. (2018b). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecological Indicators, 85*, 153-164. doi:10.1016/j.ecolind.2017.10.041

Taberlet, P., Bonin, A., Zinger, L., & Coissac, É. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press.

UKMMAS (2014). United Kingdom Marine Monitoring & Assessment Strategy UKMMAS. *Retrieved on 14.02.2022 from https://moat.cefas.co.uk/biodiversity-food-webs-and-marine-protected-areas/benthic-habitats/infaunal-quality-index/.*

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., et al. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

Weis, J. S. (2015). *Marine Pollution: What Everyone Needs to Know*. Oxford: Oxford University Press.

Wenqian, C., Meng, W., Zhu, Y., Zhou, J., & Liu, L. (2013). Assessing benthic ecological status in stressed Liaodong Bay (China) with AMBI and M-AMBI. *Chinese Journal of Oceanology and Limnology, 31*, 482-492. doi:10.1007/s00343-013-2177-0

WFD (2000). Water Framework Directive 2000/60/EC of the European Parliament and of The Council of 23 October 2000 establishing a framework for Community action in the field of water policy. *Official Journal of the European Commission, 327*, 1-73.

Wilhelm, R. C., van Es, H. M., & Buckley, D. H. (2022). Predicting measures of soil health using the microbiome and supervised machine learning. *Soil Biology and Biochemistry, 164*, 108472. doi:10.1016/j.soilbio.2021.108472

Word, J. Q. (1978). The infaunal trophic index. *Southern California Coastal Water Research Project Annual Report. El Segundo*, 19-40.

CHAPTER I
Sample preservation for transport

Summary

Background

Alternative biomonitoring strategies, such as eDNA metabarcoding-based on bacterial communities, can only be applied if the DNA to be analyzed remains unaffected between sample collection and eDNA extraction in the laboratory (Bowers et al., 2021; Hestetun et al., 2021a). Therefore, preservation of samples during transport is necessary to inhibit changes of the eDNA, which can be caused by nuclease-induced degradation, mechanical impacts, or spontaneous chemical reactions (Lindahl, 1993; Nielsen et al., 2007; Thomsen et al., 2012). Insufficient eDNA preservation can result in a distorted picture of the bacterial community, which can lead to a misinterpretation of the ecological status in the subsequent evaluation (Morgan et al., 2010; Rubin et al., 2013). Ideally, eDNA from environmental samples is extracted directly after sampling, requiring on-board processing and extraction on research vessels (Hestetun et al., 2021b; McCarthy et al., 2015). In most cases, this is not feasible due to a lack of personnel and time on the vessels, meaning that samples need to be transported to the site of their extraction. To preserve DNA, two main methods have been used so far, either preservation of the samples with a preservation solution specially developed for this purpose or preservation of samples by freezing (-18°C) without further addition of preservation solutions.

Preservation solutions such as LifeGuard® (MoBio, 2011) aim to suppress the activity of nucleic acid degrading enzymes, thus precluding RNA and DNA modification during transport (MoBio, 2011). As preservation solutions are costly, one must expect increased project expenses. For example, LifeGuard® preservation solution currently retails for 2222 € per liter[1]. Furthermore, the addition of a certain amount of preservation solution to each sample individually increases expedition time and personnel costs. MoBio states that the advantage of using LifeGuard® is that there is no need for freezing samples in order to preserve them sufficiently. Freezing units are sometimes cumbersome to carry and operate in the field.

[1] LifeGuard®, meanwhile distributed by Qiagen, price accessed via www.qiagen.com on 20.03.2022

Therefore, preservation solutions have been used several times for bacteria-based eDNA metabarcoding despite the high costs (Cordier et al., 2019a; Laroche et al., 2017; Lejzerowicz et al., 2015). Comparatively, freezing of untreated samples at -18°C is a more budget-friendly and time-saving alternative to adding expensive preservation solutions (Aylagas et al., 2014; Lanzén et al., 2020; Polinski et al., 2019; Steyaert et al., 2020). Sample preservation by freezing is applicable if cooled transport to the analyzing facility can be done in a timely manner. For refrigerated transportation, it should be noted that dry ice or other freezing options might add to transportation costs.

Regardless of the economic aspects, it is important for the establishment of SOPs to achieve reproducible conventions (Bowers et al., 2021; Goldberg et al., 2016). Therefore, it is essential to compare ecological interpretations between preservation strategies with a reliable reference to find the method which allows drawing correct conclusions about the ecological status. This chapter deals with the potential influence of the preservation method on the composition of the bacterial community and thus on the assessment of the ecological state of the environment.

Methods

Twenty sediment samples from three different coastal marine sites impacted by different anthropogenic activities have been sampled during compliance monitoring. The samples were immediately split into 40 aliquots which were then preserved in two different ways: 20 aliquots were frozen without further treatment at -18°C, while the other 20 aliquots were treated with LifeGuard® preservation solution (MoBio, 2011). In the following, frozen aliquots are referred to as non-treated aliquots, while LifeGuard®-preserved aliquots are referred to as treated aliquots. After simultaneous extraction of eDNA, the V3-V4 hypervariable gene region of 16S DNA was amplified and Illumina-sequenced.

To find potential influences of the conservation method on the quality of the extracted eDNA, quality losses along the bioinformatic sequence filtering process were compared between treated and non-treated aliquots using a two-way analysis of variance (ANOVA; Chambers and Hastie, 1992). Bacterial inventories were bioinformatically inferred from treated and non-treated aliquots separately. The properties of those inventories were compared to each other in various statistical approaches.

Analysis of alpha diversity indicating the diversity within each aliquot was conducted for the two sets of data subjected to the different preservation strategies. The alpha diversity measures ASV richness and Shannon diversity index were compared using a model II regression analysis (Sokal and Rohlf, 2012). Additionally, a correlation analysis was conducted for both measures (Kendall, 1948; Pearson and Henrici, 1896). To account for differences in community composition among treated and non-treated aliquots, beta diversity was determined and compared between the preservation strategies using a multivariate analysis of variance (ADONIS; Anderson, 2001). For one of the three sites of sample origin, an additional macrofauna-to-sample matrix was available, therefore it was possible to compare eDNA-based beta diversity to the macrofauna-derived beta diversity. Additionally, shared bacterial families and shared ASVs among treated and non-treated aliquots were determined. To account for differences in ecological significance due to frequency of occurrence, the ASV community was analyzed either including or excluding rare ASVs. Intersections among treatment methods were illustrated using Venn diagrams.

Results and Discussion

The number of sequences that could pass various sequence quality filters in the bioinformatic processing did not differ significantly between aliquots that were preserved using different preservation methods. This congruency suggests that the tested strategies of preservation for coastal marine sediments do not affect DNA extraction efficiency. Two alpha diversity measures, ASV richness and Shannon index, were compared between treated and non-treated aliquots. For both measures, correlation analysis demonstrated a strong concordance of the treated and non-treated aliquots (ASV richness: correlation coefficient 0.92 with $p < 0.001$; Shannon index: correlation coefficient 0.90 with $p < 0.001$). Exemplary, the results of the ASV richness analysis per sample are presented in *Figure 6.*

Additionally, the ASV richness measures from treated and non-treated aliquots were subjected to a regression analysis, where a high degree of agreement was observed ($R^2 = 0.85$). For the regression and correlation analysis of the Shannon index among the different preservation methods, a similar degree of correspondence could be demonstrated ($R^2 = 0.95$). Therefore, both alpha diversity measures revealed a strong correspondence among treated and non-treated aliquots indicating no noticeable influence of the preservation method on diversity recovery.

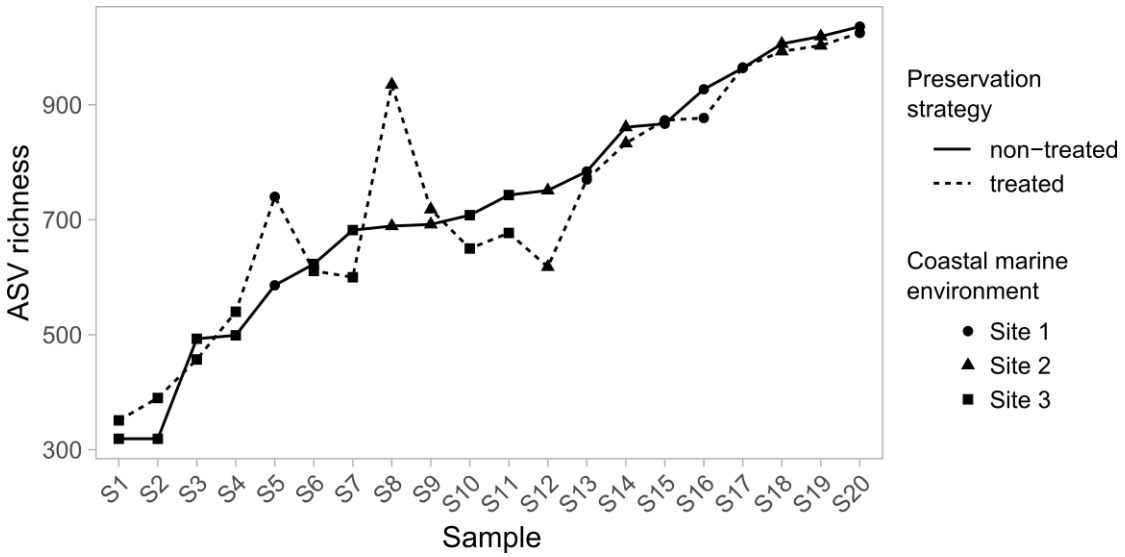


Figure 6) ASV richness comparison between non-treated and treated aliquots*. The samples were sorted according the ascending ASV richness of the non-treated aliquot. The sample ID is plotted at the x-axis. Each sample consists of two aliquotes, one non-treated aliquot (solid line) and one Lifeguard®-treated aliquot (dashed line). The shapes represent the ASV richness of the individual replicates. The shape of the dots indicates the coastal marine environment of origin (circle = Site 1; triangle = Site 2, square = Site 3).*

Subsequently, all aliquots were subjected to a beta diversity analysis, which provided information on how similar or dissimilar the bacterial communities were to each other. After statistical analysis, it was found that the aliquots could be distinguished primarily by their sample origin but hardly by the preservation method. This was confirmed using an ADONIS statistical analysis which demonstrated that the variations in community composition between aliquots subjected to different preservation methods were marginal compared to natural occurring variations. Those natural variances occur as marine sediment is a highly complex and patchy environment. On a small scale, the composition of the sediment can differ in particle topography, particle structure, sediment permeability, oxygen content, nutrient supply, and particulate matter deposition (Parsons et al., 1977; Thrush et al., 2021). However, it has already been shown several times that despite biological variations, sediment samples allow an evaluation of the ecosystem (Dowle et al., 2015; Hestetun et al., 2021a; Stoeck et al., 2018a) and are less pronounced for smaller, single-celled organisms like bacteria (Lanzén et al., 2017). Since the variation introduced by different preservation methods was within the natural variation, the divergence potentially introduced by the preservation method is not a holdback for the implementation of eDNA-based methods into standard protocols for monitoring.

Finally, inventories of commonly occurring sequence variants (ASVs) were compared among the preservation methods. At Sites 2 and 3, the inventories of the treated and the non-treated aliquots were identical (*Figure 7B, 7C*). For Site 1, the inventories were 97.8% identical *(Figure 7A).* This high correspondence among aliquots indicated that no bias was introduced by choosing preservation solution or freezing over another. This resulted in corresponding community compositions, which is why equivalent ecological conclusions were expected.

For Site 2, additional reference macrofauna inventories were available, so eDNA-based beta diversity could directly be compared with macrofauna-derived beta diversity. As expected, the macrofauna community structure revealed an ecological gradient describing diminishing community changes depending on increasing distance to the point of input of the pollution (Brown et al., 1987; Frühe et al., 2021; Keeley et al., 2012). Regarding the analyzed macrofauna of Site 2, samples deriving from the allowable zone of effect (AZE) were found to be more similar to reference samples than to highly impacted samples taken directly at the salmon farm cages. The eDNA-based beta diversity analysis using either treated or non-treated aliquots mirrored the macrofauna-based picture equally well. The analysis therefore resulted in a high congruency of the ecological interpretation independent of the preservation strategy used for the eDNA-based data.

Figure 7) Venn diagrams of the shared ASV among non-treated and treated aliquots at the different sites*. Solid circles indicate non-treated ASV inventories, dotted circles indicate treated ASV inventories. For A) Site 1, the intersection is 97.8%, which means 97.8% of all ASVs were shared between non-treated and treated aliquots. For B) Site 2 and C) Site 3, ASVs inventories shared 100%, they were therefore identical.*

Our results were partly consistent with the literature. For soil samples in general, a change in the community due to the use of preservation solutions has been shown several times (McCarthy et al., 2015; Pavlovska et al., 2021; Tatangelo et al., 2014). Regarding DNA extraction efficiency and ecological interpretation, differences between treated and non-treated samples were documented (Iturbe-Espinoza et al., 2021).

One reason for this is perhaps the different nature of gram-positive and gram-negative bacteria. Since gram-positive bacteria might grow better under the influence of preservation solutions, gram-negative bacteria may disintegrate faster (Iturbe-Espinoza et al., 2021). However, the results of other studies must be evaluated with caution, as they are not directly transferable to other ecosystems such as marine coastal sediments (Loeza-Quintana et al., 2020). In our study, no substantial shifts in the bacterial community comparing treated and non-treated samples were found. These results were consistent with another study published by Pavlovska et al. (2021), who focused on the bacterial community of forest soils. The authors found that no preservation method induced change in alpha diversity measurements, which was consistent with our results.

In conclusion, neither of the investigated conservation methods entailed a distortion of the bacterial community. Both methods, nucleic acid preservation solution utilization and freezing of samples without further treatment, concluded the same ecological evaluation. Therefore, both preservation methods are suitable for eDNA-based monitoring approaches of marine coastal environments. Accordingly, we recommend the implementation of both sample preservation methods into SOPs for marine coastal monitoring. Depending on budget and cooling capabilities, the operator can decide which strategy is more appropriate for the specific use case. However, it should be noted that multiple thawing-freezing cycles may influence eDNA integrity (Bowers et al., 2021; Cardona et al., 2012; Cuthbertson et al., 2015). If the samples are subjected to further analysis which includes several thawing-freezing cycles, usage of a preservation solution is recommended despite high costs.

References

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32-46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Aylagas, E., Borja, A., & Rodriguez-Ezpeleta, N. (2014). Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). *Plos One, 9*(3), e90529. doi:10.1371/journal.pone.0090529

Bowers, H. A., Pochon, X., von Ammon, U., Gemmell, N., Stanton, J.-A. L., Jeunen, G.-J., Sherman, C. D. H., & Zaiko, A. (2021). Towards the Optimization of eDNA/eRNA Sampling Technologies for Marine Biosecurity Surveillance. *Water, 13*(8), 1113. doi:10.3390/w13081113

Brown, J. R., Gowen, R. J., & McLusky, D. S. (1987). The effect of salmon farming on the benthos of a Scottish sea loch. *Journal of Experimental Marine Biology and Ecology, 109*(1), 39-51. doi:10.1016/0022-0981(87)90184-5

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., et al. (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol, 12*(1), 158. doi:10.1186/1471-2180-12-158

Chambers, J. M., & Hastie, T. (1992). *Statistical Models in S*. New York: Chapman and Hall/CRC.

Cordier, T., Frontalini, F., Cermakova, K., Apothéloz-Perret-Gentil, L., Treglia, M., Scantamburlo, E., Bonamin, V., & Pawlowski, J. (2019a). Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Marine Environmental Research, 146*, 24-34. doi:10.1016/j.marenvres.2018.12.009

Cuthbertson, L., Rogers, G. B., Walker, A. W., Oliver, A., Hoffman, L. R., Carroll, M. P., Parkhill, J., Bruce, K. D., & van der Gast, C. J. (2015). Implications of multiple freeze-thawing on respiratory samples for culture-independent analyses. *Journal of Cystic Fibrosis, 14*(4), 464-467. doi:10.1016/j.jcf.2014.10.004

Dowle, E., Pochon, X., Keeley, N., & Wood, S. A. (2015). Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *Fems Microbiology Ecology, 91*(8), fiv089. doi:10.1093/femsec/fiv089

Frühe, L., Dully, V., Forster, D., Keeley, N. B., Laroche, O., Pochon, X., Robinson, S., Wilding, T. A., & Stoeck, T. (2021). Global Trends of Benthic Bacterial Diversity and Community Composition Along Organic Enrichment Gradients of Salmon Farms. *Frontiers in Microbiology, 12*, e637811. doi:10.3389/fmicb.2021.637811

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307. doi:10.1111/2041-210X.12595

Hestetun, J. T., Lanzén, A., & Dahlgren, T. G. (2021a). Grab what you can—an evaluation of spatial replication to decrease heterogeneity in sediment eDNA metabarcoding. *PeerJ, 9*, e11619. doi:10.7717/peerj.11619

Hestetun, J. T., Lanzén, A., Skaar, K. S., & Dahlgren, T. G. (2021b). The impact of DNA extract homogenization and replication on marine sediment metabarcoding diversity and heterogeneity. *Environmental DNA, 3*, 997–1006. doi:10.1002/edn3.223

Iturbe-Espinoza, P., Brandt, B. W., Braster, M., Bonte, M., Brown, D. M., & van Spanning, R. J. M. (2021). Effects of DNA preservation solution and DNA extraction methods on microbial community profiling of soil. *Folia Microbiologica, 66*, 597–606. doi:10.1007/s12223-021-00866-0

Keeley, N. B., Forrest, B. M., Crawford, C., & Macleod, C. K. (2012). Exploiting salmon farm benthic enrichment gradients to evaluate the regional performance of biotic indices and environmental indicators. *Ecological Indicators, 23*, 453-466. doi:10.1016/j.ecolind.2012.04.028

Kendall, M. G. (1948). *Rank correlation methods*. Oxford: Griffin.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2017). DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *Plos One, 12*(6), e0179443. doi:10.1371/journal.pone.0179443

Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology, 30*, 2969-2987. doi:10.1111/mec.15489

Laroche, O., Wood, S. A., Tremblay, L. A., Lear, G., Ellis, J. I., & Pochon, X. (2017). Metabarcoding monitoring analysis: the pros and cons of using co-extracted environmental DNA and RNA data to assess offshore oil production impacts on benthic communities. *PeerJ, 5*, e3347. doi:10.7717/peerj.3347

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D., & Pawlowski, J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports, 5*, e13932 doi:10.1038/srep13932.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature, 362*(6422), 709-715. doi:10.1038/362709a0

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

McCarthy, A., Chiang, E., Schmidt, M. L., & Denef, V. J. (2015). RNA Preservation Agents and Nucleic Acid Extraction Method Bias Perceived Bacterial Community Composition. *Plos One, 10*(3), e0121659. doi:10.1371/journal.pone.0121659

MoBio (2011). LifeGuard Soil Preservation Solution - Instruction Manual. *Retrieved on 01.01.2022 from https://www.qiagen.com/us/resources/download.aspx?id=982fd584-9776-4dff-9324-587291cfe0fb&lang=en.*

Morgan, J. L., Darling, A. E., & Eisen, J. A. (2010). Metagenomic Sequencing of an In Vitro-Simulated Microbial Community. *Plos One, 5*(4), e10209. doi:10.1371/journal.pone.0010209

Nielsen, K. M., Johnsen, P. J., Bensasson, D., & Daffonchio, D. (2007). Release and persistence of extracellular DNA in the environment. *Environmental Biosafety Research, 6*(1-2), 37-53. doi:10.1051/ebr:2007031

Parsons, T. R., Takahashi, M., & Hargrave, B. (1977). *Biological oceanographic processes* (2nd ed.). Oxford: Pergamon Press.

Pavlovska, M., Prekrasna, I., Parnikoza, I., & Dykyi, E. (2021). Soil Sample Preservation Strategy Affects the Microbial Community Structure. *Microbes and Environments, 36*(1), ME20134. doi:10.1264/jsme2.ME20134

Pearson, K., & Henrici, O. M. F. E. (1896). VII. Mathematical contributions to the theory of evolution; III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, 187*, 253-318. doi:10.1098/rsta.1896.0007

Polinski, J. M., Bucci, J. P., Gasser, M., & Bodnar, A. G. (2019). Metabarcoding assessment of prokaryotic and eukaryotic taxa in sediments from Stellwagen Bank National Marine Sanctuary. *Scientific Reports, 9*(1), e14820. doi:10.1038/s41598-019-51341-3

Rubin, B. E. R., Gibbons, S. M., Kennedy, S., Hampton-Marcell, J., Owens, S., & Gilbert, J. A. (2013). Investigating the Impact of Storage Conditions on Microbial Community Composition in Soil Samples. *Plos One, 8*(7), e70460. doi:10.1371/journal.pone.0070460

Sokal, R., & Rohlf, F. (2012). *Biometry: the principles and practice of statistics in biological research* (2nd ed.). London: Macmillan Education.

Steyaert, M., Priestley, V., Osborne, O., Herraiz, A., Arnold, R., & Savolainen, V. (2020). Advances in metabarcoding techniques bring us closer to reliable monitoring of the marine benthos. *Journal of Applied Ecology, 57*(11), 2234-2245. doi:10.1111/1365-2664.13729

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Tatangelo, V., Franzetti, A., Gandolfi, I., Bestetti, G., & Ambrosini, R. (2014). Effect of preservation method on the assessment of bacterial community structure in soil and water samples. *FEMS Microbiology Letters, 356*(1), 32-38. doi:10.1111/1574-6968.12475

Thomsen, P. F., Kielgast, J. O. S., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology, 21*(11), 2565-2573. doi:10.1111/j.1365-294X.2011.05418.x

Thrush, S., Hewitt, J., Pilditch, C., & Norkko, A. (2021). *Ecology of Coastal Marine Sediments: Form, Function, and Change in the Anthropocene*. Oxford: Oxford University Press.

Publication:

Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems

Verena Dully[a], Giulia Rech[a], Thomas A. Wilding[b], Anders Lanzén[c,d], Kate MacKichan[e], Iain Berrill[f], Thorsten Stoeck[a,*]

[a] *Technische Universit¨at Kaiserslautern, Ecology, D-67663 Kaiserslautern, Germany*

[b] *Scottish Association for Marine Science, Scottish Marine Institute, Oban, Scotland, United Kingdom*

[c] *AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Pasaia, Gipuzkoa, Spain*

[d] *IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*

[e] *Scottish Sea Farms, Stirling, Scotland, United Kingdom*

[f] *Scottish Salmon Producers Organization, Edinburgh, Scotland, United Kingdom*

[*] *corresponding author*

From:

Dully, V., Rech, G., Wilding, T.A., Lanzén, A., MacKichan, K., Berrill, I, & Stoeck, T. (2021). *Marine Pollution Bulletin, 173*, e113129, doi:10.1016/j.marpolbul.2021.113129

Abstract

To avoid loss of genetic information in environmental DNA (eDNA) field samples, the preservation of nucleic acids during field sampling is a critical step. In the development of standard operating procedures (SOPs) for eDNA-based compliance monitoring, the effect of different routinely used sediment preservations on biological community structures serving as bioindicators has gone untested. We compared eDNA metabarcoding results of marine bacterial communities from sample aliquots that were treated with a nucleic acid preservation solution (treated samples) and aliquots that were frozen without further treatment (non-treated samples). Sediment samples were obtained from coastal locations subjected to different stressors (aquaculture, urbanization, industry). DNA extraction efficiency, bacterial community profiles, and measures of alpha- and beta-diversity were highly congruent between treated and non-treated samples. As both preservation methods provide the same relevant information to environmental managers and regulators, we recommend the inclusion of both methods into SOPs for biomonitoring in marine coastal environments.

1. Introduction

The advent of efficient, inexpensive, fast, up-scalable and fully automatable DNA metabarcoding protocols (Aylagas et al., 2016; Cordier et al., 2017; Keeley et al., 2018; Lanzén et al., 2020) and data analyses packages (Aylagas and Rodriguez-Ezpeleta, 2016; Macher et al., 2021; Zinger et al., 2021) have heralded in a new era in the field of environmental biomonitoring. Traditional methods for assessing marine ecosystem health usually rely on the microscopic identification of macroinvertebrates, which can be used as bioindicators due to their specific response to individual stressors or a combination of multiple stressors. Such stressors include for example aquaculture associated organic enrichment (Brown et al., 1987; Carroll et al., 2003), hydrocarbon pollution at production sites (Cordier et al., 2019; Lanzén et al., 2016; Laroche et al., 2016), or toxic chemicals originating in industry or urban wastewater (Chariton et al., 2010; Vaalgamaa et al., 2013; Yoon et al., 2020). Traditional taxonomic monitoring approaches rely on the observations of organisms that can be identified morphologically, such as macroinvertebrates (Bonada et al., 2006; Magurran et al., 2010) or algae (Reavie et al., 2010), and do not allow the exploitation of strong biomarkers that are not readily identifiable and quantifiable using microscopy, such as bacteria. Furthermore, they are limited by high costs, lengthy sample analysis time, often unverifiable taxonomic precision, and cannot easily be scaled up (Baird

and Hajibabaei, 2012). The necessity to increase the frequency and scale of environmental biomonitoring due to an increasing human impact on our planet in general (Chiang et al., 2021; Myers and Smith, 2018; Orr et al., 2005; Wu et al., 2013) and coastal ecosystems in particular (Harley et al., 2006; Puritz and Toonen, 2011; van de Velde et al., 2018), has fueled the development of alternative biomonitoring strategies. The solution that is widely accepted as having the highest potential for assessing the ecological status of marine ecosystems is the interrogation of DNA extracted from environmental samples. In these protocols, environmental DNA (eDNA) is extracted from e.g. sediment or water samples and specific taxonomic target genes are amplified using PCR prior to high-throughput sequencing (Taberlet et al., 2018). For marine biomonitoring, bacteria have emerged as a potentially powerful indicator group (Dowle et al., 2015; Keeley et al., 2018; Stoeck et al., 2018a; Verhoeven et al., 2018), alongside other microscopic organisms such as ciliates (Forster et al., 2019; Stoeck et al., 2018b), foraminifera (Pawlowski et al., 2016a; Pochon et al., 2015) and diatoms (Apothéloz-Perret-Gentil et al., 2017; Rivera et al., 2018).

Current efforts of the scientific, industrial and regulatory communities are to develop eDNA-standard protocols to exploit bacteria as bioindicators, which can be implemented in routine biomonitoring regulations and practice (Cordier et al., 2017; Laroche et al., 2018; Stoeck et al., 2018a). The establishment of standard protocols for sample collection, treatment and analysis, are critical for enabling results that can be compared between different monitoring studies, including in relation to sampling (Hestetun et al., 2021a), DNA extraction (Hestetun et al., 2021b; Pearman et al., 2021) and reproducibility of PCR and Illumina sequencing (Dully et al., 2021). One aspect that has so far gone untested, to the best of our knowledge, is the treatment of samples between collection and DNA extraction. In the first step after sample collection, it is necessary to preserve the sediment samples after sampling. In the ideal case, DNA should be extracted from sediment samples immediately after sample collection. However, this is often times not possible and samples will have to be stored several days to months until DNA extraction. Poorly controlled storage condition may then result in DNA degradation or microbial growth and severely bias downstream analyses of the DNA-derived biological community profiles (Rubin et al., 2013). The two most widely used options are preservation of samples using a nucleic acid preservation solution (Cordier et al., 2019; Laroche et al., 2017; Lejzerowicz et al., 2015) or freezing of the sample without any further sample treatment (Aylagas et al., 2014; Lanzén et al., 2020; Polinski et al., 2019; Steyaert et al., 2020).

Commercially available preservation solutions such as Lifeguard® prevent RNAse and DNAse activities, allowing for 16S rRNA profiling (metabarcoding) of bacterial communities on samples collected in the field under any conditions (MoBio, 2011). The advantage of using a preservation solution is that no additional equipment is required to freeze the sample after collection (such as a freezer, dry ice or a liquid nitrogen cooler), which is often challenging in remote locations or aboard small sampling vessels. Furthermore, an expensive cooled transport to the sample-processing laboratory by a courier service is not required because microbial community profiles in environmental samples are maintained with such a solution for at least one week and RNA integrity even for 30 days at room temperature (MoBio, 2011). On the downside, commercially available standardized nucleic acid preservation solutions can be relatively expensive (currently up to ca. €2000 per liter), which adds substantially to the costs of environmental monitoring. Furthermore, adding preservation buffer to each sediment sample is more time consuming, adding further increasing costs compared to freezing samples. Thus, freezing samples within a few hours of collection is a faster and less-expensive option for sample preservation. However, in this case the additional transportation costs of the frozen samples on blue or dry ice from the location of the sampling to the sample processing laboratory should also be considered. The major question, however, is the influence of the sediment preservation method on the resultant data.

To address this question, we here collected sediment samples from different marine coastal locations, namely two different locations at the west coast of Scotland subjected to organic enrichment resulting from aquafarming installations and estuarine sites at the Basque coast subjected to different degrees of urban and industrial impacts. We split all samples in two aliquots, one set of which was preserved with a commonly used nucleic acid preservation solution, and one set of which was preserved by freezing. We then extracted DNA from all samples and used a standard eDNA metabarcoding protocol to analyze the bacterial community composition. Statistical analyses were then conducted to compare the results obtained from frozen and solution-preserved samples to inform the development of standardized operating protocols for eDNA-based compliance monitoring.

2. Material and methods

2.1 Sample sites and sampling

Samples were collected from two salmon farm locations in Scotland, namely DUN located close to Oban and LIS located in Loch Linnhe. Further, samples were collected from several locations on the Basque coast (Bay of Biscay). Farms DUN and LIS were sampled during the mid-production and peak-production period, respectively, in December 2020. Sediment was collected at three sites along a transect from an outer cage edge (CE) to a reference site (REF) in the direction of the prevailing current flow, located ca. 800 m (DUN) and 525 m (LIS) distant from the CE. An intermediate impact zone (Allowable Zone of Effect, AZE) was located at ca. 100 m (DUN) and 109 m (LIS) distance from the cage edge. This sampling design followed a decreasing organic enrichment gradient from the CE towards the REF site, resulting from the deposits of feed and fish feces on the sea floor (Brown et al., 1987; Frühe et al., 2021; Keeley et al., 2012). At each site, two biological replicates were taken with a van Veen grab (0.1 m2 area, DUN; 0.045 m2 area, LIS). From each replicate we sampled approximately 20–25 g of surface sediment (upper few millimeters) into a sterile 50 ml plastic tube using disposable sterile plastic spatulas. Immediately following collection, we homogenized the samples in the 50-ml collection tube with a spatula and divided each sample in two nearly equal aliquots by transferring approximately half of the sediment from the original 50-ml collection tube into a fresh 50-ml collection tube. One set of aliquots was frozen within a few hours of collection (=non-treated samples). The second set of aliquots was incubated with an equal volume (10 ml) of Qiagen LifeGuard nucleic acid preservation solution (formerly MoBio's LifeGuard solution) (=treated samples) and stored in the fridge at 4 ◦C upon arrival of the samples in the lab (within the same day of sample collection).

Non-treated samples were shipped (<48 h) from Scotland to our laboratory in Kaiserslautern (Germany) on dry-ice; treated samples were shipped at ambient temperature and both sample sets were then frozen at -20 °C for one month prior to further processing. Samples from farm LIS were taken during routine compliance monitoring and are accompanied by macroinvertebrate inventories which were used to calculate the AMBI Index (Borja et al., 2000) and ecological quality. We used the macrofaunal data matrix as a reference to determine the relative similarity of macrofaunal communities at the LIS sampling sites among replicates to compare with the here obtained eDNA metabarcoding profiles of bacterial communities (see below). The macrofauna was obtained from

van Veen grabs after subsampling for DNA analyses. Therefore, the remaining sediment was washed through a 1 mm sieve, and the residue fixed in 4% borax-buffered formaldehyde prior to macrobenthic sorting and counting. The sieve-retained fauna was identified to species level under the National Marine Biological Quality Control Scheme (NMBAQCS) by APEM Ltd., Hertfordshire.

Samples from the Basque coast (Bay of Biscay) were collected from four different sites, all from tidal flats of three different estuaries of the rivers Oka, Urola and Bidasoa. The four sites were chosen for having contrasting environmental statuses. Two sites were sampled directly downstream of wastewater treatment plants: EOK05 in Oka, in the town Gernika, and EU08 in Urola, outside Zumaia. EU08 is considered to be moderately impacted while EOK05 has nearly anoxic sediments and is heavily affected by organic enrichment (Lanzén et al., 2020). The other site from Oka was located near the mouth of the river in Sukarrieta, from a sandy beach, with very good status (EOK20). Finally, EBI20 is located in a small sandy beach in the town Hondarribia near the mouth of Bidasoa with relatively little impact (good status). Samples were collected manually and followed the exact same procedure as described above for the aquaculture sites.

2.2 DNA extraction, amplification and Illumina sequencing of Scottish salmon farm samples

Following our previously described protocol (Frühe et al., 2020), environmental DNA was obtained from sediment samples using the PowerSoil DNA kit (Qiagen, Hilden, Germany) according to the manufacturer's manual. As DNA metabarcodes, we obtained the ca. 450 bp long hypervariable V3-V4 region of the bacterial 16S rRNA gene. The PCR protocol with the Bakt_341F (CCTACGGGNGGCWGCAG) and the Bakt_805R (GACTACHVGGG-TATCTAATCC) primer pair employed an initial activation step of NEB's Phusion High-Fidelity DNA polymerase at 98 ◦C for 30 s, followed by 27 identical three-step cycles consisting of 98 ◦C for 10 s, 62 ◦C for 30 s, and 72 ◦C for 30 s; then a final 5-min extension at 72 ◦C (Herlemann et al., 2011). From the resulting PCR products, sequencing libraries were constructed using the NEB Next® Ultra™ DNA Library Prep Kit for Illumina (NEB, USA). The quality of the libraries was assessed with an Agilent Bioanalyzer 2100 system. V3-V4 libraries were sequenced on an Illumina MiSeq platform, generating 2 × 300-bp paired-end reads. All sequences can be accessed from NCBI's SRA database under BioProject ID PRJNA768445.

2.3 Sequence data processing

Raw sequence reads were quality filtered and trimmed by executing the DADA2 workflow (Callahan et al., 2016) in R Studio 3.5.1 to obtain ASVs (Amplicon Sequence Variants). Truncation length was set to 255 bp (Dully et al., 2021) so the phred quality score reached >30 for at least 51% of all reads corresponding to 99.9% base call accuracy (Ewing et al., 1998). For maxEE we chose 1 to maximize downstream sequence quality. The paired end sequences were merged using minimum 20 bp overlap and a mismatch of two bases was allowed (Frühe et al., 2021). Before the construction of ASV-to-sample matrices, the sequences were checked for chimeras using the *uchime_denovo* function of *vsearch* (Rognes et al., 2016). Taxonomic assignment was conducted using *vsearch*'s *syntax* function based on the greengenes database (McDonald et al., 2012). To analyze sequencing depth, saturation curves for each dataset were constructed using the *rarecurve* function of the *vegan* package (Oksanen et al., 2020). Subsequently, normalization of read counts using the *rrarefy* function to the minimum sequence number sample of the DUN, LIS and BASQUE dataset was applied to account for differences in sequencing depth. ASVs contributing to less than 0.1% of the total reads per dataset were discarded to reduce uninformative noise as in similar studies (Lanzén et al., 2020).

2.4 Statistical analysis

a) Influence of sample treatment on sequence quality

To determine whether the treatment strategy has an influence on the obtainable number of high-quality sequences, we compared the number of raw sequences after initial quality filtering, sequence merging, and chimera check between treated and non-treated samples. For each dataset, a two-way analysis of variance (ANOVA; Chambers and Hastie, 1992) was conducted modelling the downstream sequence number as a function of the sample treatment and the quality filtering stage using *aov* function. Prior to this, sequence numbers were tested for normal distribution using Shapiro Wilk test (Shapiro and Wilk, 1965) implemented in the *shapiro.test* function.

b) Alpha diversity

For each individual replicate sample, the alpha diversity measures ASV richness and Shannon Index (Shannon and Weaver, 1949) were calculated using the *diversity* function of the *vegan* package. Treated and non-treated samples from the same spatial replicate were plotted against each other, and a linear Model II (Sokal and Rohlf, 2012)

was fit. Further, a correlation analyses to compare alpha-diversity measures was conducted. First, the alpha diversity measures were tested for normal distribution using Shapiro-Wilk test of normality using the *shapiro.test* function. Correlation between treated and non-treated replicates was then tested using Pearson correlation (Pearson and Henrici, 1896) for normally distributed data or Spearman rank correlation (Kendall, 1948) for non-normally distributed data. Therefore, we used the *cor.test* function. The ggplot2 package was used for graphical representations (Wickham, 2009).

c) Beta diversity

Bray Curtis (BC) dissimilarity matrices were calculated for each of the rarefied datasets (DUN, LIS, BASQUE) using the *vegdist* function of the R *vegan* package. Dendrograms were constructed for each dataset to visualize bacterial community similarities between replicates and between treated and non-treated samples.

We ran a permutational multivariate analysis of variance using distance matrices ADONIS (Anderson, 2001) using the *adonis* function implemented in the *vegan* package. It uses a permutation test for the partition of variation calculated directly from the dissimilarity matrix. For the LIS salmon farm samples, we also obtained a faunal abundance matrix of the macroinvertebrate inventory of each LIS replicate sample. Using the same community analyses as described above, we calculated a BC dissimilarity matrix for the macroinvertebrate communities and constructed a dendrogram. This dendrogram served as a reference from traditional compliance monitoring to which we compared the obtained beta diversity pattern of the LIS bacterial communities obtained from the treated and non-treated samples.

d) ASV and taxa distribution in treated and non-treated samples

To compare the proportion of shared ASVs and taxa between biological replicates and preservation replicates, incidence-based Venn Diagrams were constructed in three ways, namely based on ASV composition, based on ASV composition without rare ASVs and based on detected bacterial families. Family level was considered the optimal balance between taxonomic resolution and ASV assignment (Supplementary File 1[*]). To evaluate the ASV composition without rare ASVs, only ASVs accounting for >1% of the total reads in a dataset were taken into account.

[*] All supplementary files are additionally available at the appendix of this dissertation

3. Results

3.1 Sequence data overview

After cleaning of the obtained raw sequence datasets, we retained 211,293, 275,911 and 371,684 high quality (HQ) sequences for the DUN, LIS and BASQUE dataset, respectively. Subsequently, normalization of read counts to the minimum sequence number sample of the DUN, LIS and BASQUE dataset was applied to account for differences in sequencing depth, resulting in 12,700, 9542 and 7638 reads per sample respectively. After discarding ASVs accounting for <0.1% of the total reads per dataset to reduce uninformative noise, the final number of ASVs for downstream analyses were 3076 ASVs for the DUN dataset, 2575 ASVs for the LIS dataset and 2502 ASVs for the BASQUE dataset. Rarefaction profiles showed that all samples were sequenced to near saturation (Supplementary File 2[*]).

3.2 Statistical analysis

a) Sequence quality of treated and non-treated samples

To reveal potential differences in sequence quality between treated and non-treated samples, ANOVA was applied for the individual datasets comparing sequence numbers at three different quality filtering steps. We found that sequence loss after initial quality and length filtering, loss of non-mergeable reads, and loss of sequences due to chimera formation was insignificantly different between treated and non-treated samples in all datasets (DUN: $p = 0.49$, LIS: $p = 0.07$, BASQUE: $p = 0.09$).

b) Alpha diversity

The normalized ASV richness as well as the Shannon Index as a representative index measure of alpha diversity was highly similar when comparing results obtained from non-treated and from LifeGuard-treated samples (Fig. 1).

Using Model II regression, we compared treated against non-treated samples for both measures with a resulting R2 of 0.95 and a slope of 0.8 for the Shannon Index and a R2 of 0.85 with a slope of 0.96 for the ASV richness. Pearson correlation for the normally distributed ASV richness revealed a highly significant p of <0.001 and a correlation coefficient of 0.92. Spearman rank for the non-normally distributed Shannon Index also resulted in a highly significant correlation ($p < 0.001$) with a correlation coefficient (rho) of 0.9.

[*] All supplementary files are additionally available at the appendix of this dissertation

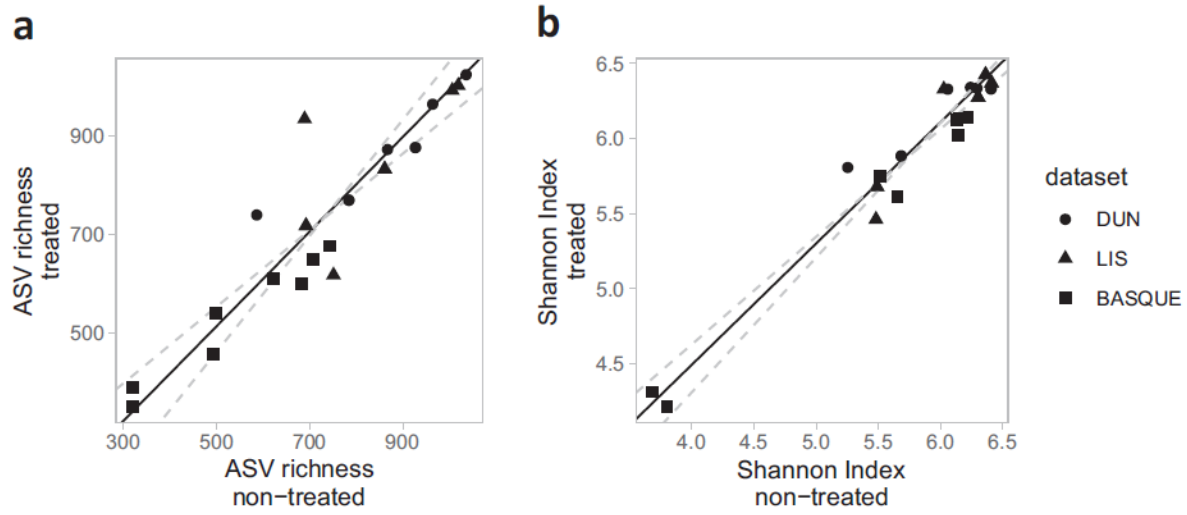


Fig. 1. *Normalized ASV richness (a) and Shannon Index (b) as a representative index measure of alpha diversity obtained from non-treated samples and from LifeGuard- treated samples. The solid black line represents the linear regression (Model II). The gray, dashed lines indicate the 95% confidence interval. The shapes indicate the sample location (DUN = circle; LIS = triangle; BASQUE = square).*

c) Beta diversity

For the DUN aquaculture samples, we find a clear treatment effect at all three sampling sites (CE, AZE, REF): Non-treated replicate samples have a higher similarity in bacterial community structure to each other than do treated samples and vice versa (Fig. 2a). This observation is, however, not consistent with all sampling sites at the LIS farm (Fig. 2b) and the BASQUE coast (Fig. 2c). At AZE and REF sites of the LIS samples, treated samples are more similar to non-treated samples than they are to their respective replicate samples. A treatment effect was observed for the cage edge site CE. For the BASQUE dataset, non-treated samples are more similar to treated samples as to the samples using the same treatment strategy for the sampling sites EBI20 and EU08, while we observed a treatment effect for sites EOK20 and EOK05.

However, the order of magnitude of treatment effects is notably smaller compared to the sampling site effects. In all datasets (DUN, LIS, BASQUE), bacterial community structures are notably more dissimilar to each other at the different sampling sites compared to the relatively small differences caused by the treatment effect. An ADONIS analyses confirmed highly significant effects of the sampling sites on bacterial community structure (for all three datasets: $p < 0.001$, $R^2 = 0.66$ for DUN, 0.72 for LIS and 0.88 for BASQUE), while treatment effects were insignificant ($p = 0.07$ and $R^2 = 0.08$ for DUN, $p = 0.53$ and $R^2 = 0.03$ for LIS, and $p = 0.27$ and $R^2 = 0.01$ for BASQUE).

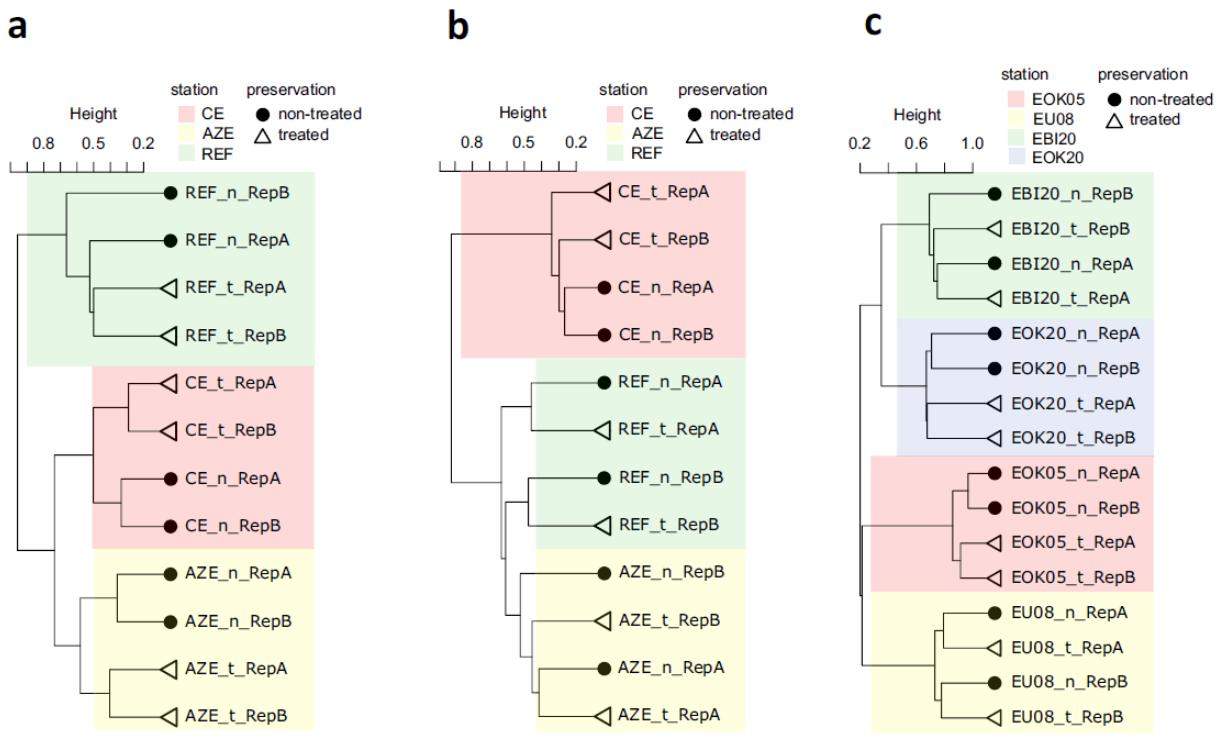


Fig. 2. *Beta-diversity dendrogram of bacterial community structures based on Bray-Curtis dissimilarities: (a) the DUN salmon farm; (b) the LIS salmon farm; (c) sampling sites at the Basque coast (BASQUE). At both salmon farms, the cage edge (CE), the allowable zone of effect (AZE) and a reference site (REF) were sampled and analyzed. At the Basque coast, four different estuarine locations subject to different kinds and degrees of industrial and urban stressors were included in our study. Sampling stations are indicated by color. The sample treatment strategy is indicated by shape: circles indicate non-treated samples while triangles indicate treated samples.*

d) ASV and taxa distribution in treated and non-treated samples

The proportion of bacterial ASV that were shared among non-treated biological replicates and LifeGuard-treated biological replicates was (i) highly consistent and (ii) in the same order of magnitude compared to the proportion of ASVs that were shared among non-treated and treated samples (Fig. 3). This was regardless of the sampling location (DUN – Fig. 3a, LIS – Fig. 3b, BASQUE – Fig. 3c) and regardless of whether all low-abundant ASVs were included in the analyses (middle panel in each figure) or excluded (i.e. using ASV accounting for >1% of the reads in a dataset, right panel in each figure). To mention one example for the consistency of shared ASVs between the different treated and non-treated samples: in the BASQUE dataset, the frozen biological replicates shared on average 68.8% of their ASVs and LifeGuard-treated sample shared 64.1% of their ASVs on average (Fig. 3c). When comparing the different treatments (treated vs. non-treated) in the same BASQUE dataset with each other, the number of shared ASVs even exceeds the proportion in the biological replicate comparisons and was as high as 79.7%.

When reducing noise of low-abundant ASVs even further (considering ASVs with at least 1% of read abundance within a dataset), the shared ASVs between non-treated and treated samples is up to 100% (LIS and BASQUE datasets, Fig. 3b and c). The same trend was found when we looked at the taxonomic rank of family instead of ASVs (Fig. 3a–c, left panel). In a comparison of the LIS microbial dataset with the LIS macrofauna community similarities among the same samples (Fig. 4), we find a high congruency in the partitioning of community diversity. Macrofaunal communities were more similar to each other at REF and AZE sites compared to the macrofaunal community at the CE site. The macrofaunal based biotic index AMBI was more similar between the REF (AMBI: 2.0 for replicate 1 and 1.8 for replicate 2) and AZE sites (AMBI: 2.3 for replicate 1 and 2.2 for replicate 2), compared to the CE site (AMBI: 5.7 for replicate 1 and 5.8 for replicate 2). The same picture was mirrored in the microbial communities, regardless of whether samples were non-treated or LifeGuard-treated (Fig. 2b).

4. Discussion

The preservation of nucleic acids during field sampling is a critical step in any eDNA-based study. To avoid the loss of genetic information from field samples, freezing or preservation of nucleic acid by a stabilizing solution are the two commonly used options (Bowers et al., 2021). However, this step has remained relatively unexplored by benchmarking studies so far and, to our knowledge, is not addressed in any of the efforts to develop a standardized protocol. Therefore, we compared two preservation strategies that are commonly used in this context, namely preservation by freezing (Dowle et al., 2015; Lejzerowicz et al., 2015) and by LifeGuard solution (Cordier et al., 2020; Laroche et al., 2018). We expected to see treatment effects when comparing bacterial community structures in marine coastal sediment samples. However, the magnitude of these effects was marginal compared to the effects of the sampling locations within each of the three datasets. We found that ecological gradients, which are reflected in the sampling conditions, are equally well mirrored in the benthic bacterial community structures that were obtained from both sediment preservation approaches.

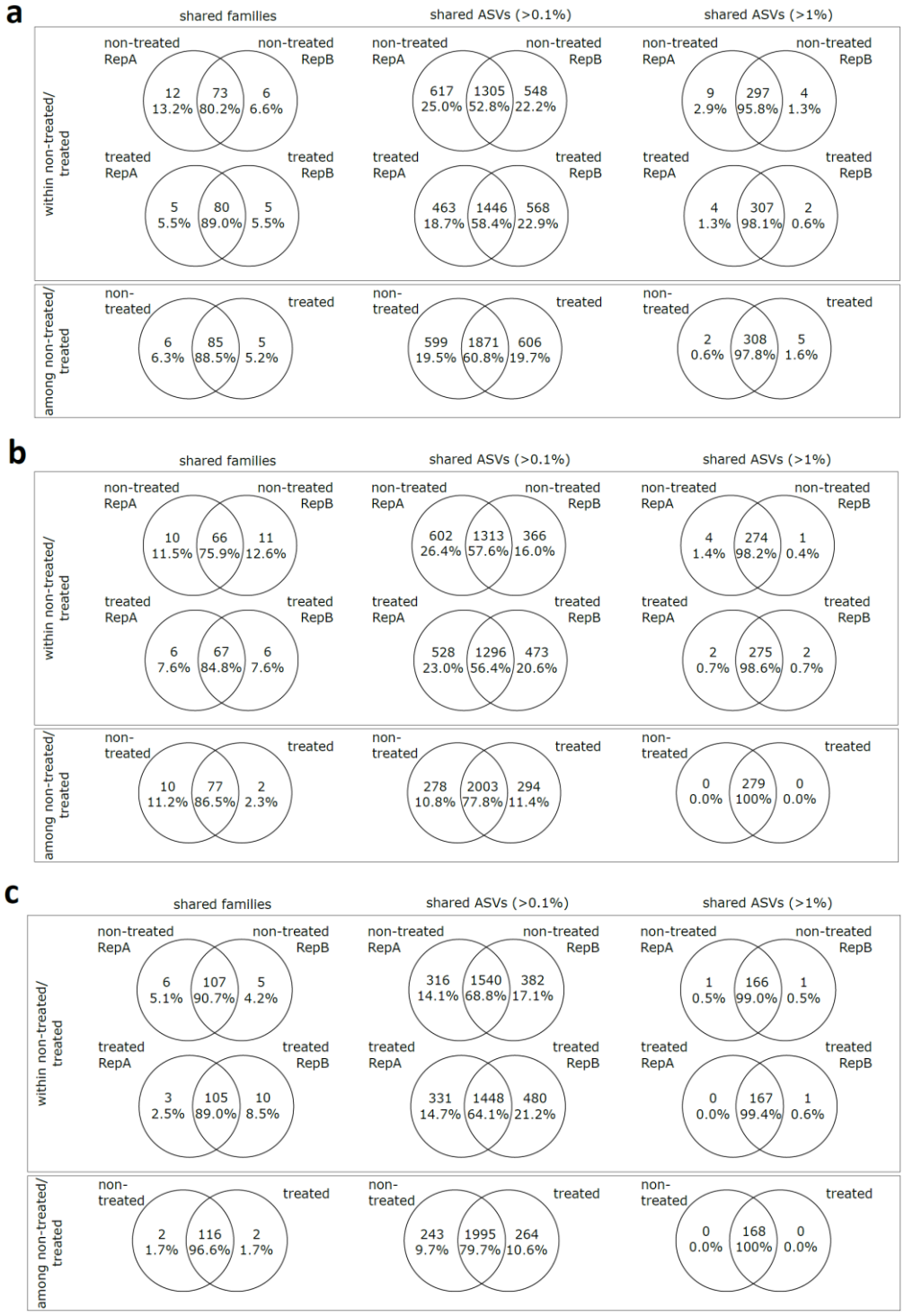


Fig. 3. *Venn diagrams to visualize the proportion of bacterial families and ASVs shared among true biological replicates of the same sediment sample preservation treatment (non-treated or LifeGuard-treated) and among samples of different preservation treatments (non-treated vs. treated). (a) DUN samples; (b) LIS samples; (c) Basque coast samples. For ASVs we used two different datasets, one in which all ASVs were considered that accounted for >0.1% of all sequence reads within a dataset (=inclusion of low-abundant ASVs, middle panel of figures) and one in which only ASVs were considered that accounted for >1% of all sequence reads within a dataset (=exclusion of low-abundant ASVs, left panel of figures). The proportion of shared ASVs and families in the among-treatment comparison (non-treated vs. treated) is in the same order of magnitude as the one shared within true biological replicates that were either non-treated or LifeGuard-treated. Furthermore, the figure illustrates that when low-abundant ASVs are removed, near-identical ASVs are recovered within replicates and among differently preserved samples.*

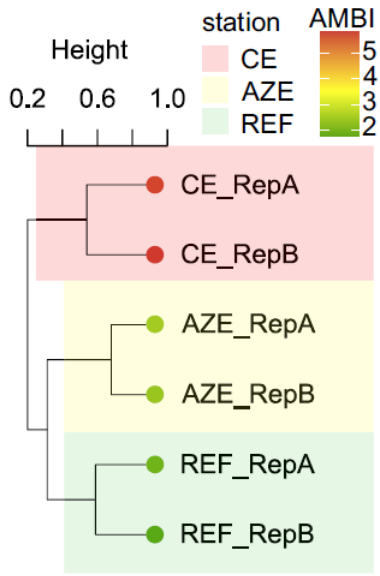


Fig. 4. *Beta-diversity dendrogram of macrofaunal community structures based on Bray-Curtis dissimilarities at the LIS salmon farm site. The colored dots of the samples in the dendrogram are coding the biotic AMBI index as calculated from the macrofaunal bioindicators.*

Therefore both, LifeGuard preservation as well as sample freezing without any nucleic acid stabilizing solution, are principally suitable sampling strategies as part of a standardized eDNA biomonitoring protocol. This conclusion is further supported by a highly congruent pattern matching between the microbial community structures of both, treated and non-treated LIS aquaculture samples and the macrofauna reference communities from the same aquaculture site, which are used as traditional bioindicators in compliance monitoring because their composition reflects the state of the environment (Pearson and Rosenberg, 1978). Based on macrofaunal communities, a biotic index, such as AMBI, is inferred to calculate the ecological quality of the samples under study (Borja et al., 2000; Muxika et al., 2005). Such an index can be equally well inferred from bacterial community structures to conduct environmental monitoring in marine coastal environments (Aylagas et al., 2021; Keeley et al., 2018; Lanzén et al., 2020).

More studies addressed potential bias of nucleic acid preservation solutions for soil samples (McCarthy et al., 2015; Pavlovska et al., 2021; Tatangelo et al., 2014), while few studies have investigated such effects on aquatic samples in general and to the best of our knowledge none for marine sediment samples. Our results are in agreement with the findings of Pavlovska et al. (2021) who investigated the effect of soil sample preservation strategies on microbial community structures. The authors reported that bacterial community richness estimated with the Shannon Index and evenness were not significantly different between soil samples preserved with freezing and soil samples preserved with a

nucleic acid preservation solution. Furthermore, qualitative data on taxonomic diversity obtained from the same samples were congruent among the differently preserved samples at the class, family and genus level. Pavlovska et al. (2021) also reported that a nucleic acid preservation solution better preserved rare taxa. However, we consider this of only minor relevance for the specific case of biomonitoring, as biomonitoring relies mostly on the presence of abundant bioindicators (Borja et al., 2000; Fortunato et al., 2013; Madoni, 1994). Pavlovska et al. (2021) also used a preservation solution from Zymo, while we used the solution offered by Qiagen. However, Gray et al. (2013) compared the effects of different preservation solutions on environmental bacterial community samples and came to the conclusion that no single solution outperformed any other.

For soil samples, Iturbe-Espinoza et al. (2021) found substantial differences in the DNA extraction efficiency for samples preserved with LifeGuard solution compared to freezing, leading to different ecological interpretations of the results obtained with the two different soil preservation strategies. Iturbe-Espinoza et al. (2021) showed the bacterial community profiles obtained from duplicate soil DNA samples were much more dissimilar among LifeGuard-preserved samples compared to freeze-preserved samples. Furthermore, community structures were remarkably incongruent between LifeGuard solution treated samples and frozen samples, with a notable higher relative abundance of Gram-positive species versus Gram-negative bacteria in the solution-preserved samples. Such an increased representation of Gram-positive bacteria and also of Chloroflexi and Alphaproteobacteria in the total soil bacterial community profile after use of a similar preservation solution (RNAlater) was also reported by Rissanen et al. (2010). For freshwater plankton samples, McCarthy et al. (2015) showed that preservation of filtered samples with Qiagen's RNAlater biased the bacterial community composition at the DNA-level (i.e. when investigating eDNA and not eRNA) relative to non-treated, frozen samples. However, despite this methodological bias introduced by sample preservation technique, the sample origin was the strongest determinant of community composition, corroborating the results of our own study.

Other studies in soil, water and also fecal samples reported bias in the bacterial community composition at the DNA level (Chen et al., 2019; Dominianni et al., 2014; Rissanen et al., 2010; Tatangelo et al., 2014), but rarely, if at all, in the dominant bacterial taxon groups detected in these individual samples. In their search of an explanation for a differential preference of individual bacterial taxon groups in LifeGuard-preserved samples, Iturbe-Espinoza et al. (2021) incubated individual bacterial cultures within the

LifeGuard solution. Unexpectedly, the authors found that this preservation solution (50% vol:vol solution to sample, as recommended from the supplier for sample preservation) supported the growth of individual bacterial taxa rather than maintaining them in stasis, which they explained by the utilization of carbon and energy sources from the preservation solution at ambient temperature. These authors found predominantly a bias for the enrichment of Gram-positive bacteria in nucleic acid preservation solution treated soil samples. The authors argued that this may come from a decreased membrane integrity resulting in enhanced efficiency of DNA extraction from mostly Gram-positive bacteria. Another possibility could be an increased porosity of the membranes of Gram-negative bacteria, leading to an enhanced release of DNA from these bacteria and its subsequent breakdown, also resulting in a higher relative abundance of Gram-positive taxa (Iturbe-Espinoza et al., 2021). Results of other studies do not corroborate with the findings of Iturbe-Espinoza et al. (2021), however, observed bias in bacterial community compositions was detected independently of gram properties of bacterial taxa (Chen et al., 2019; Dominianni et al., 2014; McCarthy et al., 2015; Rissanen et al., 2010; Tatangelo et al., 2014). Our results did not corroborate with these previous findings, and even the quantitative DNA extraction efficiency was not significantly different between treated and non-treated samples (Supplementary File 3[*]). It should also be noted that prolonged preservation of frozen samples, for more than one or two months at -20°C, can dramatically reduce DNA yield and potentially bias the perceived community structure, based on results from anaerobic sludge (Romanazzi et al., 2015).

4.1 Recommendation for a standardized sampling protocol

Our recommendation towards a standardized field protocol for eDNA-based marine biomonitoring as inferred from our results is that both sample preservation methods can be included as standards in official monitoring regulations, leaving the decision to the operators. The choice for an expensive commercially available preservation solution such as Qiagen LifeGuard is recommended if freezing is not available within a few hours of sample collection, or when chilled transportation of frozen samples to the laboratory would exceed the costs of a preservation solution. Further consideration may be if samples will be subjected to multiple freeze-thaw cycles. (Bowers et al., 2021) reported that freeze-thaw induced degradation of environmental DNA may occur already after one cycle of freeze-thawing. Therefore, for repeated sub-sampling of the same samples (which requires

[*] All supplementary files are additionally available at the appendix of this dissertation

thawing), a preservation solution should be the preferred choice. We note that our recommendation applies to samples that do not exceed the storage period of one month as tested in this study. Whether the same results will be obtained after notably longer periods of storage requires further investigations. Supplementary data to this article can be found online at https://doi. org/10.1016/j.marpolbul.2021.113129*.

CRediT authorship contribution statement

Verena Dully – Investigation, Formal analysis, Data curation, Visualization, Writing – Review & Editing. **Giulia Rech** – Investigation, Formal analysis, Data curation. **Thomas A. Wilding** – Sampling, Investigation, Validation, Resources, Writing – Review & Editing. **Anders Lanzén** – Sampling, Conceptualization, Validation, Resources, Writing – Review & Editing. **Iain Berrill** – Conceptualization, Resources, Writing – Review & Editing. **Thorsten Stoeck** – Conceptualization, Methodology, Validation, Writing – Original draft, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research leading to these results received funding from the Deutsche Forschungsgemeinschaft (DFG grants STO4141/15-1 and STO414/15-2) and the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984, ASSEMBLE Plus project. We would like to thank Leire Garate, Ion Abad (AZTI) and Jason Dobson (Scottish Sea Farms Limited) for sample collection and processing. Also thanks to Sheena Gallie (formerly Scottish Sea Farms Limited) for participating in this project and enabling the sampling of the LIS salmon farm. We also thank the crew of the R/V Seol Mara and Gail Twigg (SAMS) for technical support sampling of the DUN salmon farm.

* All supplementary files are additionally available at the appendix of this dissertation

References

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32-46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources, 17*(6), 1231-1242. doi:10.1111/1755-0998.12668

Aylagas, E., Atalah, J., Sánchez-Jerez, P., Pearman, J. K., Casado, N., Asensi, J., Toledo-Guedes, K., & Carvalho, S. (2021). A step towards the validation of bacteria biotic indices using DNA metabarcoding for benthic monitoring. *Molecular Ecology Resources, 21*(6), 1889-1903. doi:10.1111/1755-0998.13395

Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Frontiers in Marine Science, 3*(96). doi:10.3389/fmars.2016.00096

Aylagas, E., Borja, A., & Rodriguez-Ezpeleta, N. (2014). Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). *Plos One, 9*(3), e90529. doi:10.1371/journal.pone.0090529

Aylagas, E., & Rodriguez-Ezpeleta, N. (2016). Analysis of Illumina MiSeq Metabarcoding Data: Application to Benthic Indices for Environmental Monitoring. *Methods in molecular biology, 1452*, 237-249. doi:10.1007/978-1-4939-3774-5_16

Baird, D., & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology, 21*, 2039-2044. doi:10.1111/j.1365-294X.2012.05519.x

Bonada, N., Prat, N., Resh, V. H., & Statzner, B. (2006). Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology, 51*, 495-523. doi:10.1146/annurev.ento.51.110104.151124

Borja, A., Franco, J., & Pérez, V. (2000). A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments. *Marine Pollution Bulletin, 40*(12), 1100-1114. doi:10.1016/S0025-326X(00)00061-8

Bowers, H. A., Pochon, X., von Ammon, U., Gemmell, N., Stanton, J.-A. L., Jeunen, G.-J., Sherman, C. D. H., & Zaiko, A. (2021). Towards the Optimization of eDNA/eRNA Sampling Technologies for Marine Biosecurity Surveillance. *Water, 13*(8), 1113. doi:10.3390/w13081113

Brown, J. R., Gowen, R. J., & McLusky, D. S. (1987). The effect of salmon farming on the benthos of a Scottish sea loch. *Journal of Experimental Marine Biology and Ecology, 109*(1), 39-51. doi:10.1016/0022-0981(87)90184-5

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Carroll, M. L., Cochrane, S., Fieler, R., Velvin, R., & White, P. (2003). Organic enrichment of sediments from salmon farming in norway: environmental factors, management practices, and monitoring techniques. *Aquaculture, 226*, 165-180. doi:10.1016/S0044-8486(03)00475-7

Chambers, J. M., & Hastie, T. (1992). *Statistical Models in S*. New York: Chapman and Hall/CRC.

Chariton, A. A., Court, L. N., Hartley, D. M., Colloff, M. J., & Hardy, C. M. (2010). Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment, 8*(5), 233-238. doi:10.1890/090115

Chen, Z., Hui, P. C., Hui, M., Yeoh, Y. K., Wong, P. Y., Chan, M. C. W., et al. (2019). Impact of Preservation Method and 16S rRNA Hypervariable Region on Gut Microbiota Profiling. *mSystems, 4*(1), e00271-00218. doi:10.1128/mSystems.00271-18

Chiang, F., Mazdiyasni, O., & AghaKouchak, A. (2021). Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. *Nature Communications, 12*, e2754. doi:10.1038/s41467-021-22314-w

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology, 30*, 2937-2958. doi:10.1111/mec.15472

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01518

Cordier, T., Frontalini, F., Cermakova, K., Apothéloz-Perret-Gentil, L., Treglia, M., Scantamburlo, E., Bonamin, V., & Pawlowski, J. (2019). Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Marine Environmental Research, 146*, 24-34. doi:10.1016/j.marenvres.2018.12.009

Dominianni, C., Wu, J., Hayes, R. B., & Ahn, J. (2014). Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiolology, 14*(1), 103. doi:10.1186/1471-2180-14-103

Dowle, E., Pochon, X., Keeley, N., & Wood, S. A. (2015). Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *Fems Microbiology Ecology, 91*(8), fiv089. doi:10.1093/femsec/fiv089

Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., & Stoeck, T. (2021). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter-laboratory study. *Ecological Indicators, 121*, e107049. doi:10.1016/j.ecolind.2020.107049

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research, 8*(3), 175-185. doi:10.1101/gr.8.3.175

Forster, D., Filker, S., Kochems, R., Breiner, H.-W., Cordier, T., Pawlowski, J., & Stoeck, T. (2019). A Comparison of Different Ciliate Metabarcode Genes as Bioindicators for Environmental Impact Assessments of Salmon Aquaculture. *Journal of Eukaryotic Microbiology, 66*(2), 294-308. doi:10.1111/jeu.12670

Fortunato, C. S., Eiler, A., Herfort, L., Needoba, J. A., Peterson, T. D., & Crump, B. C. (2013). Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. *The ISME Journal, 7*(10), 1899-1911. doi:10.1038/ismej.2013.79

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

Frühe, L., Dully, V., Forster, D., Keeley, N. B., Laroche, O., Pochon, X., Robinson, S., Wilding, T. A., & Stoeck, T. (2021). Global Trends of Benthic Bacterial Diversity and Community Composition Along Organic Enrichment Gradients of Salmon Farms. *Frontiers in Microbiology, 12*, e637811. doi:10.3389/fmicb.2021.637811

Gray, M. A., Pratte, Z. A., & Kellogg, C. A. (2013). Comparison of DNA preservation methods for environmental bacterial community samples. *FEMS Microbiology Ecology, 83*(2), 468-477. doi:10.1111/1574-6941.12008

Harley, C. D., Randall Hughes, A., Hultgren, K. M., Miner, B. G., Sorte, C. J., Thornber, C. S., Rodriguez, L. F., Tomanek, L., & Williams, S. L. (2006). The impacts of climate change in coastal marine systems. *Ecology Letters, 9*(2), 228-241. doi:10.1111/j.1461-0248.2005.00871.x

Herlemann, D. P. R., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., & Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal, 5*(10), 1571-1579. doi:10.1038/ismej.2011.41

Hestetun, J. T., Lanzén, A., & Dahlgren, T. G. (2021a). Grab what you can—an evaluation of spatial replication to decrease heterogeneity in sediment eDNA metabarcoding. *PeerJ, 9*, e11619. doi:10.7717/peerj.11619

Hestetun, J. T., Lanzén, A., Skaar, K. S., & Dahlgren, T. G. (2021b). The impact of DNA extract homogenization and replication on marine sediment metabarcoding diversity and heterogeneity. *Environmental DNA, 3*, 997–1006. doi:10.1002/edn3.223

Iturbe-Espinoza, P., Brandt, B. W., Braster, M., Bonte, M., Brown, D. M., & van Spanning, R. J. M. (2021). Effects of DNA preservation solution and DNA extraction methods on microbial community profiling of soil. *Folia Microbiologica, 66*, 597–606. doi:10.1007/s12223-021-00866-0

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

Keeley, N. B., Forrest, B. M., Crawford, C., & Macleod, C. K. (2012). Exploiting salmon farm benthic enrichment gradients to evaluate the regional performance of biotic indices and environmental indicators. *Ecological Indicators, 23*, 453-466. doi:10.1016/j.ecolind.2012.04.028

Kendall, M. G. (1948). *Rank correlation methods*. Oxford: Griffin.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology, 25*(17), 4392-4406. doi:10.1111/mec.13761

Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology, 30*, 2969-2987. doi:10.1111/mec.15489

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lear, G., & Pochon, X. (2018). A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Marine Pollution Bulletin, 127*, 97-107. doi:10.1016/j.marpolbul.2017.11.042

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lejzerowicz, F., Pawlowski, J., Lear, G., Atalah, J., & Pochon, X. (2016). First evaluation of foraminiferal metabarcoding for monitoring environmental impact from an offshore oil drilling site. *Marine Environmental Research, 120*, 225-235. doi:10.1016/j.marenvres.2016.08.009

Laroche, O., Wood, S. A., Tremblay, L. A., Lear, G., Ellis, J. I., & Pochon, X. (2017). Metabarcoding monitoring analysis: the pros and cons of using co-extracted environmental DNA and RNA data to assess offshore oil production impacts on benthic communities. *PeerJ, 5*, e3347. doi:10.7717/peerj.3347

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D., & Pawlowski, J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports, 5*, e13932 doi:10.1038/srep13932.

Macher, T.-H., Beermann, A. J., & Leese, F. (2021). TaxonTableTools: A comprehensive, platform-independent graphical user interface software to explore and visualise DNA metabarcoding data. *Molecular Ecology Resources, 21*(5), 1705-1714. doi:10.1111/1755-0998.13358

Madoni, P. (1994). A sludge biotic index (SBI) for the evaluation of the biological performance of activated sludge plants based on the microfauna analysis. *Water Research, 28*, 67-75. doi:10.1016/0043-1354(94)90120-1

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution, 25*(10), 574-582. doi:10.1016/j.tree.2010.06.016

McCarthy, A., Chiang, E., Schmidt, M. L., & Denef, V. J. (2015). RNA Preservation Agents and Nucleic Acid Extraction Method Bias Perceived Bacterial Community Composition. *Plos One, 10*(3), e0121659. doi:10.1371/journal.pone.0121659

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal, 6*(3), 610-618. doi:10.1038/ismej.2011.139

MoBio (2011). LifeGuard Soil Preservation Solution - Instruction Manual. *Retrieved on 01.01.2022 from https://www.qiagen.com/us/resources/download.aspx?id=982fd584-9776-4dff-9324-587291cfe0fb&lang=en.*

Muxika, I., Borja, A., & Bonne, W. (2005). The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. *Ecological Indicators, 5*(1), 19-31. doi:10.1016/j.ecolind.2004.08.004

Myers, S., & Smith, M. (2018). Impact of anthropogenic CO2 emissions on global human nutrition. *Nature Climate Change, 8*. doi:10.1038/s41558-018-0253-3

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2020). R Package 'vegan': Community Ecology Package. *Version 2.5-7.*

Orr, J. C., Fabry, V. J., Aumont, O., Bopp, L., Doney, S. C., Feely, R. A., et al. (2005). Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature, 437*(7059), 681-686. doi:10.1038/nature04095

Pavlovska, M., Prekrasna, I., Parnikoza, I., & Dykyi, E. (2021). Soil Sample Preservation Strategy Affects the Microbial Community Structure. *Microbes and Environments, 36*(1), ME20134. doi:10.1264/jsme2.ME20134

Pawlowski, J., Esling, P., Lejzerowicz, F., Cordier, T., Visco, J. A., Martins, C. I. M., Kvalvik, A., Staven, K., & Cedhagen, T. (2016a). Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. *Aquaculture Environment Interactions, 8*, 371-386. doi:10.3354/aei00182

Pearman, J. K., Thomson-Laing, G., Howarth, J. D., Vandergoes, M. J., Thompson, L., Rees, A., & Wood, S. A. (2021). Investigating variability in microbial community composition in replicate environmental DNA samples down lake sediment cores. *Plos One, 16*(5), e0250783. doi:10.1371/journal.pone.0250783

Pearson, K., & Henrici, O. M. F. E. (1896). VII. Mathematical contributions to the theory of evolution; III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, 187*, 253-318. doi:10.1098/rsta.1896.0007

Pearson, T., & Rosenberg, R. (1978). Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology, 16*, 229-311. doi:10.2983/035.034.0121u1.10

Pochon, X., Wood, S. A., Keeley, N. B., Lejzerowicz, F., Esling, P., Drew, J., & Pawlowski, J. (2015). Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Marine Pollution Bulletin, 100*(1), 370-382. doi:10.1016/j.marpolbul.2015.08.022

Polinski, J. M., Bucci, J. P., Gasser, M., & Bodnar, A. G. (2019). Metabarcoding assessment of prokaryotic and eukaryotic taxa in sediments from Stellwagen Bank National Marine Sanctuary. *Scientific Reports, 9*(1), e14820. doi:10.1038/s41598-019-51341-3

Puritz, J. B., & Toonen, R. J. (2011). Coastal pollution limits pelagic larval dispersal. *Nature Communications, 2*(1), 226. doi:10.1038/ncomms1238

Reavie, E., Jicha, T., Angradi, T., Bolgrien, D., & Hill, B. (2010). Algal assemblages for large river monitoring: Comparison among biovolume, absolute and relative abundance metrics. *Ecological Indicators, 10*, 167-177. doi:10.1016/j.ecolind.2009.04.009

Rissanen, A. J., Kurhela, E., Aho, T., Oittinen, T., & Tiirola, M. (2010). Storage of environmental samples for guaranteeing nucleic acid yields for molecular microbiological studies. *Applied Microbiology and Biotechnology, 88*(4), 977-984. doi:10.1007/s00253-010-2838-2

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., & Rimet, F. (2018). Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia, 807*(1), 37-51. doi:10.1007/s10750-017-3381-2

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ, 4*, e2584. doi:10.7717/peerj.2584

Romanazzi, V., Traversi, D., Lorenzi, E., & Gilli, G. (2015). Effects of freezing storage on the DNA extraction and microbial evaluation from anaerobic digested sludges. *BMC research notes, 8*, 420. doi:10.1186/s13104-015-1407-2

Rubin, B. E. R., Gibbons, S. M., Kennedy, S., Hampton-Marcell, J., Owens, S., & Gilbert, J. A. (2013). Investigating the Impact of Storage Conditions on Microbial Community Composition in Soil Samples. *Plos One, 8*(7), e70460. doi:10.1371/journal.pone.0070460

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Champaign: University of Illinois Press.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), 591-611. doi:10.1093/biomet/52.3-4.591

Sokal, R., & Rohlf, F. (2012). *Biometry: the principles and practice of statistics in biological research* (2nd ed.). London: Macmillan Education.

Steyaert, M., Priestley, V., Osborne, O., Herraiz, A., Arnold, R., & Savolainen, V. (2020). Advances in metabarcoding techniques bring us closer to reliable monitoring of the marine benthos. *Journal of Applied Ecology, 57*(11), 2234-2245. doi:10.1111/1365-2664.13729

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., & Pawlowski, J. (2018b). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecological Indicators, 85*, 153-164. doi:10.1016/j.ecolind.2017.10.041

Taberlet, P., Bonin, A., Zinger, L., & Coissac, É. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press.

Tatangelo, V., Franzetti, A., Gandolfi, I., Bestetti, G., & Ambrosini, R. (2014). Effect of preservation method on the assessment of bacterial community structure in soil and water samples. *FEMS Microbiology Letters, 356*(1), 32-38. doi:10.1111/1574-6968.12475

Vaalgamaa, S., Sonninen, E., Korhola, A., & Weckström, K. (2013). Identifying recent sources of organic matter enrichment and eutrophication trends at coastal sites using stable nitrogen and carbon isotope ratios in sediment cores. *Journal of Paleolimnology, 50*(2), 191-206. doi:10.1007/s10933-013-9713-y

van de Velde, S., Van Lancker, V., Hidalgo-Martinez, S., Berelson, W. M., & Meysman, F. J. R. (2018). Anthropogenic disturbance keeps the coastal seafloor biogeochemistry in a transient state. *Scientific Reports, 8*(1), 5582. doi:10.1038/s41598-018-23925-y

Verhoeven, J. T. P., Salvo, F., Knight, R., Hamoutene, D., & Dufour, S. C. (2018). Temporal Bacterial Surveillance of Salmon Aquaculture Sites Indicates a Long Lasting Benthic Impact With Minimal Recovery. *Frontiers in Microbiology, 9*, e03054. doi:10.3389/fmicb.2018.03054

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wu, P., Christidis, N., & Stott, P. (2013). Anthropogenic impact on Earth's hydrological cycle. *Nature Climate Change, 3*(9), 807-810. doi:10.1038/nclimate1932

Yoon, S. J., Hong, S., Kim, S., Lee, J., Kim, T., Kim, B., et al. (2020). Large-scale monitoring and ecological risk assessment of persistent toxic substances in riverine, estuarine, and coastal sediments of the Yellow and Bohai seas. *Environment International, 137*, e105517. doi:10.1016/j.envint.2020.105517

Zinger, L., Lionnet, C., Benoiston, A.-S., Donald, J., Mercier, C., & Boyer, F. (2021). metabaR: An r package for the evaluation and improvement of DNA metabarcoding data quality. *Methods in Ecology and Evolution, 12*(4), 586-592. doi:10.1111/2041-210X.13552

CHAPTER II
Reproducibility of eDNA metabarcoding-based sample processing

Summary

Background

Robust methods are the key to providing a reliable assessment of the environment. Therefore, the reproducibility of DNA-based approaches is a prerequisite for implementing them in routine compliance monitoring (Goldberg et al., 2016; Helbing and Hobbs, 2019; Loeza-Quintana et al., 2020). Consequently, testing reproducibility is needed as a step towards verifying the NGS method for eDNA-based monitoring. For the eDNA analysis, PCR and sequencing represent essential molecular tools (Taberlet et al., 2018; Valentini et al., 2016). The objective of this chapter was to evaluate the robustness of those tools through separate sample handling by two independent laboratories and to compare the results to the traditional, macrofauna-based EQ. This way, biases introduced by PCR and sequencing, which potentially alter the ecological evaluation, can be identified (Berry et al., 2011; Frank et al., 2008; Kalle et al., 2014; Kennedy et al., 2014). Besides applications in ecology, NGS-based metabarcoding has emerged as a powerful tool to identify microbial community composition in other disciplines (Boers et al., 2019; Goodwin et al., 2016; Hajishengallis et al., 2012). It represents a time- and cost-effective, culture-independent approach, hence the implementation of the NGS method into diagnostic standard protocols is also being researched (Boers et al., 2019; Gohl et al., 2016; Hiergeist et al., 2016). Thus, potential PCR-induced biases are widely reviewed (Chandler et al., 1997; Haas et al., 2011; Smyth et al., 2010; Zylstra et al., 1998).

Technical biases such as selection and drift, which arise using PCR-based methods, have been reported, as well as an increased impact of PCR bias at low concentrations of template DNA (Chandler et al., 1997; Kennedy et al., 2014). Further, it was demonstrated that sequence length, GC content, and primer bias can lead to preferential amplification of specific sequences and therefore potentially alter the obtained bacterial community profiles (Frank et al., 2008; Kalle et al., 2014). Additionally, frequent PCR artifacts like chimeras, which can arise in a variety of ways, can also cause shifts in the acquired bacterial community composition (Haas et al., 2011; Odelberg et al., 1995; Zylstra et al., 1998).

Furthermore, additional technical biases regarding the eDNA method might be introduced by the sequencing of the amplicons itself. It has been described for forest soil and fecal samples that sequencing can create a representational bias resulting in over- or underrepresentation of specific taxa, or may not influence the obtained sequence data significantly (Kennedy et al., 2014). Additionally, common sequencing processes such as multiplexing can sometimes result in increased bias, although it was as well demonstrated to decrease bias by reducing the number of steps needed for analysis (Alon et al., 2011; Berry et al., 2011). Nevertheless, there are caveats concerning the results of other studies, given their potential inapplicability to distinct ecosystems (Loeza-Quintana et al., 2020). Hence, for the implementation of the eDNA method in SOPs for routine compliance monitoring, it must be investigated if potential PCR and sequencing biases can initiate misrepresentation of the ecological evaluation of marine coastal environments in particular. Therefore, we tested the robustness and reproducibility of the eDNA-based interpretations for such environments. Two independent laboratories received aliquots of the same benthic eDNA samples accompanied by molecular processing instructions. When the same environmental status is inferred by both laboratories independently, the method can be assumed to be sufficiently reproducible and robust against the influence of independent molecular lab processing. Only then, the eDNA method can be implemented into SOPs for routine compliance monitoring.

Methods

During compliance monitoring of a Scottish salmon farm, marine sediment has been examined. Macrofauna was analyzed for the traditional EQ inference using the IQI (Phillips et al., 2014). Additionally, eDNA sampling for molecular analysis has been conducted. Various sediments have been sampled along an expected pollution gradient, resulting in samples taken directly beneath the fish cages, samples taken in 50m, 60m, and 70m meters distance from the cage edge along with the prevailing current, and samples taken at two different reference sites representing pristine conditions. This schematic sampling has been conducted to capture the complete scope of EQ transition along the pollution gradient. Each sample was split into two aliquots which were processed independently. Aliquots deriving from the same sample but processed in different laboratories are referred to as technical replicates in the following section. The two independent laboratories are referred to as Eco-Lab and SeqIT-Lab in the following.

Independent V3-V4 PCR was conducted by the two laboratories using the same settings, followed by Illumina sequencing conducted on separate Illumina flow cells. Subsequently, the obtained sequences deriving from inter-laboratory handling of the technical replicates were bioinformatically processed and analyzed independently regarding alpha diversity, beta diversity, and taxonomic composition. A schematic representation of the conducted procedure can be found in *Figure 8*.

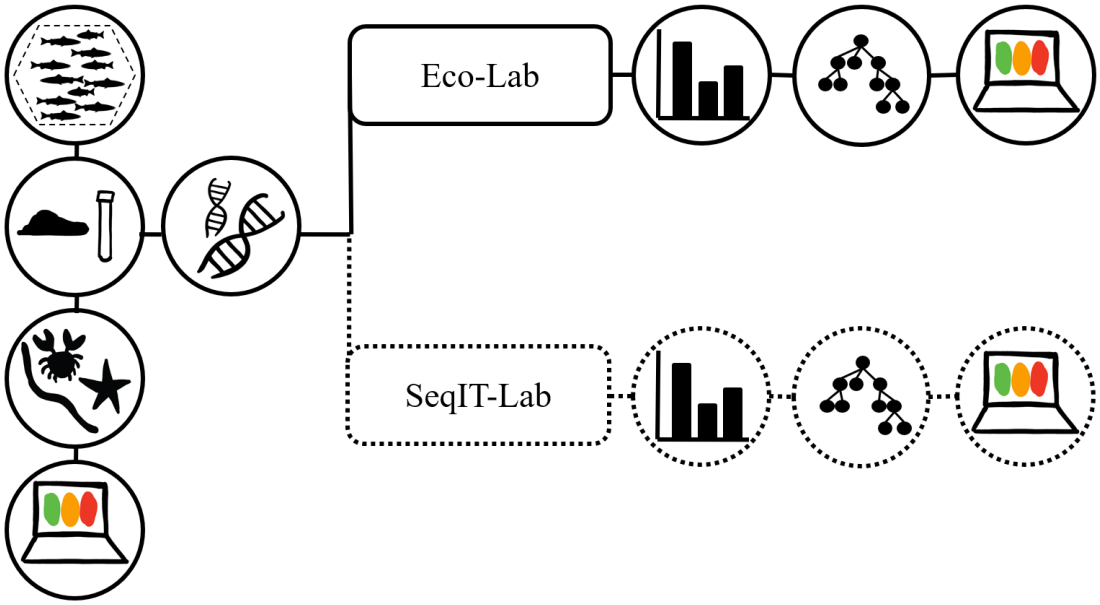


***Figure 8) Schematic representation of the conducted inter-laboratory study**. Sediment samples beneath the operating salmon farm were taken during compliance monitoring. Traditional macrofauna analysis was conducted for the ecological evaluation of the environment (left branch). Additionally, eDNA was sampled for a metabarcoding study. After extraction, eDNA was transferred to two independent laboratories (Eco-Lab or SeqIT-Lab). PCR and sequencing were conducted independently, as well as bioinformatic processing. Results obtained from alpha and beta diversity measures, taxonomic community composition, and SML were analyzed for pattern matching between the two sets of technical replicates derived from inter-laboratory processing (right branches).*

Additionally, all samples were tested for pattern matching among their technical replicates using an SML approach. RF was used to predict the IQI of the technical replicates processed in the respective opposite laboratory. Therefore, a model was trained on the data obtained by Eco-Lab, including the macrofauna-based IQI values. For model construction, a LOO approach was employed. Subsequently, the model based on Eco-Lab data was used to predict the IQI of samples processed in the SeqIT-Lab and *vice versa* (*Figure 9*).

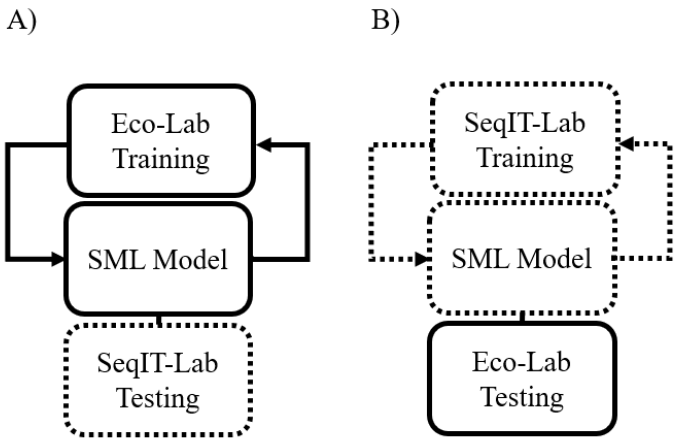


Figure 9) Schematic sketch of the procedure of the applied SML algorithm*. A) First, an SML model was created using all the replicates that have been processed in the Eco-Lab. The replicates which have been processed in the Eco-Lab thus served as a training dataset. Subsequently, this model was used to predict the IQI of the replicates which have been processed in the SeqIT-Lab. The SeqIT-Lab replicates served as a testing dataset. B) In a second step, the replicates which have been processed in the SeqIT-Lab were used for model construction while the Eco-Lab replicates functioned as the testing dataset.*

Results and Discussion

The number of obtained HQ ASVs among the two laboratories was congruent, as the Eco-Lab obtained 2232 ASVs and the independent SeqIT-Lab obtained 2230 ASVs. Additionally, alpha diversity analysis showed a high congruency among samples processed in the different laboratories, as the measures ASV richness, Shannon index, and Simpson index were highly congruent (*Figure 10*). Comparing ASVs and taxonomic inventories of both sets of technical replicates, which have been processed in independent laboratories, a high degree of similarity was observed. On average, technical replicates shared 78.2% of the ASVs. The natural variation represented by intra-laboratory processed biological replicates was higher, as 47.7% ASVs were shared (*Figure 11*). This means that the technical variance which was introduced by sample processing in independent laboratories remained within the naturally occurring variation.

Moreover, when looking at the beta diversity which represents similarities among samples, the two sets of data processed in different laboratories were highly congruent. This corroborated well with a recently published study focusing on marine biofouling metabarcoding (Zaiko et al., 2022). The authors have reported a constant beta diversity pattern when sample processing was conducted in 12 independent laboratories in six different countries.

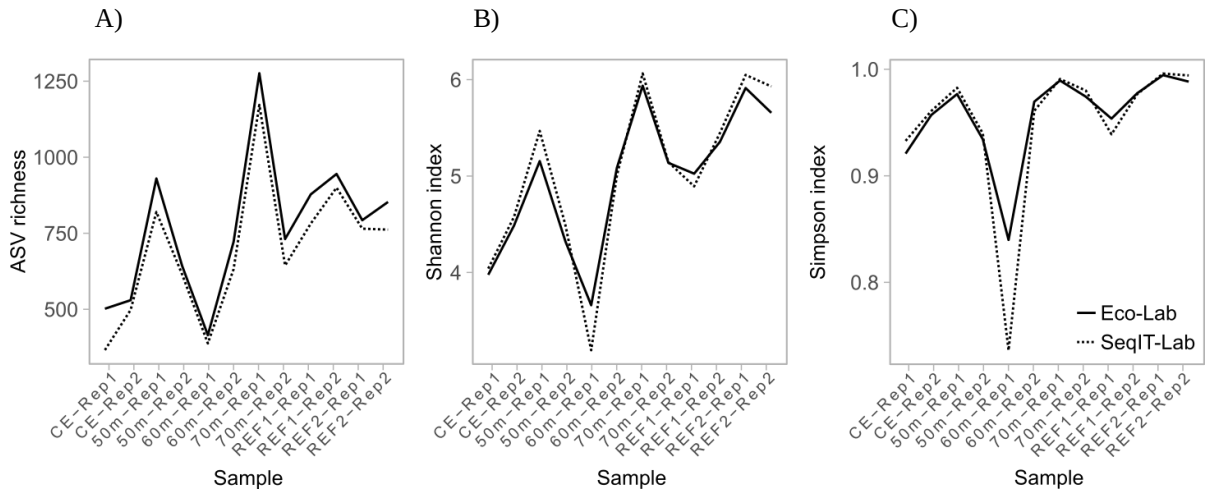


***Figure 10) Alpha diversity measures among different laboratories**. Three different alpha diversity measures, A) ASV richness, B) Shannon index, and C) Simpson index were compared between the Eco-Lab and SeqIT-Lab. Sample measures processed in the Eco-Lab are represented by the solid lines while samples processed in the SeqIT-Lab are represented by the dotted lines. Sample IDs are indicated by the X-axis in the order of the prevailing current.*

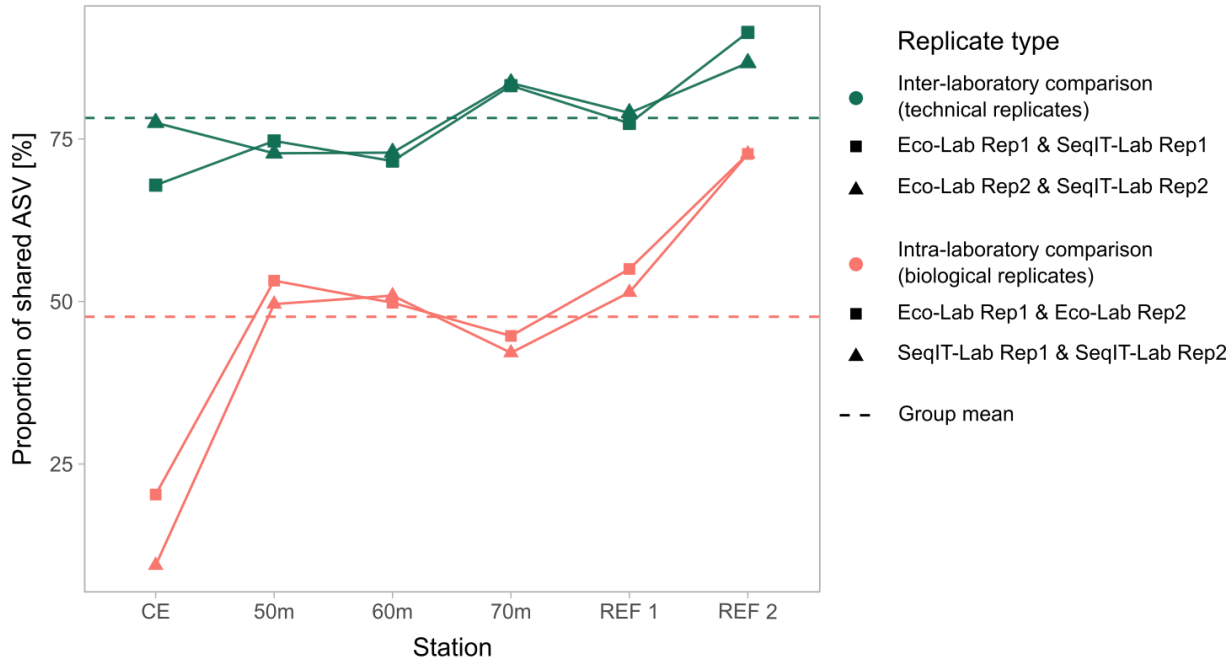


***Figure 11) Shared ASVs among technical and biological replicates**. The proportion of shared ASVs for the replicates sampled at distinct stations is presented. The proportion of shared ASVs within technical, inter-laboratory replicates (green color) which originated from different laboratories but from the same sample, and biological, intra-laboratory replicates (red color) which originated from the same laboratory but from a distinct sample, are presented. The colored shapes represent technical replicates processed in the different laboratories (Eco-Lab or SeqIT-Lab): The green square indicates replicate 1, the green triangle represents replicate 2. The red shapes represent the biological replicates 1 and 2 from a true biological sample but processed in the same laboratory: the red square represents replicates processed in the Eco-Lab while the red triangle represents samples processed in the SeqIT-Lab. The dashed lines indicate the mean of the shared ASVs among either technical replicates or biological replicates.*

They concluded that besides a high variation in raw results, the main conclusions were consistent and therefore, metabarcoding is sufficiently robust against laboratory-based effects. Regarding our dataset, two sample clusters could be identified using beta diversity statistics, one representing a high ecological status and one representing poor ecological status. This was highly congruent with the sample pattern based on the reference macrofauna, which served as a benchmark for the ecological assessment. Therefore, the eDNA-based molecular approach was able to mirror the macrofauna EQ, regardless of the laboratory the replicates have been processed in. Additionally, taxa were grouped according to their occurrence in stations that showed a 'high' ecological status (IQI≥0.64) or a 'poor' ecological status respectively (IQI<0.64), obtained by the macrofauna-based IQI inference. This was done for the taxa of each technical replicate dataset processed by the independent laboratories. Taxa occurring in high abundances either in the 'high' or in the 'poor' status group were referred to as indicators of the respective ecological status. This analysis showed a high concurrency among indicator taxa regardless of the sample processing laboratory.

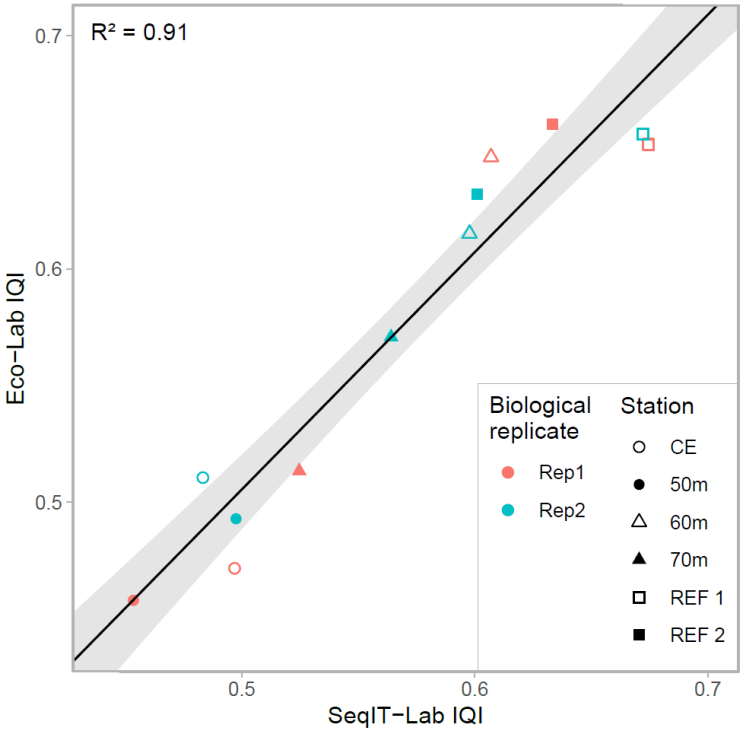


Figure 12) Predicted IQI using SML among independent laboratories. *Results of IQI prediction deriving from replicates processed in the respective opposite laboratory were plotted against each other. The shapes represent the sampled stations along the prevailing current ranging from cage edge (CE, unfilled circle) to the two reference sites REF 1 (unfilled square) and REF 2 (filled square). Intermediate stations are 50m (filled circle), 60m unfilled triangle), and 70m (filled triangle) away from the CE. The biological replicates are indicated by the color: red color indicates replicate 1 (REP 1) while a turquoise color indicates replicate 2 (REP 2). The black line indicates the linear model (R^2 = 0.91) while the grey color indicates the 95% confidence interval.*

To confirm the findings, an SML analysis was subsequently included. RF has been shown to provide reproducible classifications of samples according to their ecological status before (Cordier et al., 2017; Cordier et al., 2019b; Frühe et al., 2020), which was confirmed by applying both intra- and inter-laboratory approaches. For technical replicates, SML was still able to classify samples according to their correct EQ when the algorithm was initially trained with replicates which have been processed by the other laboratory. The fact that the model based on technical replicates from one laboratory could predict the IQI from the other laboratory emphasized a high congruency of the bacterial community structure (*Figure 12*). Additionally, when comparing RF-inferred indicator ASVs, a high-level agreement of the ASV inventories among the two independent laboratories was demonstrated. It was observed that the technical variance resulting from the sample preparation in two different laboratories was smaller than the biological variation. Thus, it can be assumed that the evaluation of the ecosystem was not distorted by independent processing conducted by independent laboratories. Because the seafloor is a highly complex and patchy ecosystem within a few centimeters, it can show variations in particulate matter deposition, oxygen availability, and particle structure on a small scale (Parsons et al., 1977; Thrush et al., 2021).

For the implementation of the eDNA method into SOPs, former recommendations which advocate sampling in spatial replicates and pooling of biological replicates before processing are therefore emphasized (Hestetun et al., 2021a; Hestetun et al., 2021b; Lanzén et al., 2017; Lejzerowicz et al., 2014). However, it was demonstrated that a reliable ecological evaluation was possible despite natural variations within biological replicates (Dowle et al., 2015; Hornick and Buschmann, 2018; Keeley et al., 2018; Stoeck et al., 2018a). In summary, alpha and beta diversity analysis, comparison of ASV and taxon inventories, and indicator inference showed a high congruency between technical replicates evaluated in two different laboratories. As all tested sample statistics suggested pattern matching among technical replicates, it was indicated that the used eDNA metabarcoding method is sufficiently robust and reproducible for environmental monitoring. As no misinterpretations of the ecological evaluation of the ecosystem were introduced by independent sample processing, implementation of the method in legislative SOPs for routine compliance monitoring can be suggested.

References

Alon, S., Vigneault, F., Eminaga, S., Christodoulou, D. C., Seidman, J. G., Church, G. M., & Eisenberg, E. (2011). Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Research, 21*(9), 1506-1511. doi:10.1101/gr.121715.111

Berry, D., Mahfoudh, K. B., Wagner, M., & Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology, 77*(21), 7846-7849. doi:10.1128/AEM.05220-11

Boers, S. A., Jansen, R., & Hays, J. P. (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *European Journal of Clinical Microbiology & Infectious Diseases, 38*(6), 1059-1070. doi:10.1007/s10096-019-03520-3

Chandler, D. P., Fredrickson, J. K., & Brockman, F. J. (1997). Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. *Molecular Ecology, 6*(5), 475-482. doi:10.1046/j.1365-294X.1997.00205.x

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01518

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019b). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology, 27*(5), 387-397. doi:10.1016/j.tim.2018.10.012

Dowle, E., Pochon, X., Keeley, N., & Wood, S. A. (2015). Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *Fems Microbiology Ecology, 91*(8), fiv089. doi:10.1093/femsec/fiv089

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., & Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Applied and Environmental Microbiology, 74*(8), 2461-2470. doi:10.1128/aem.02272-07

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., et al. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology, 34*(9), 942-949. doi:10.1038/nbt.3601

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307. doi:10.1111/2041-210X.12595

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351. doi:10.1038/nrg.2016.49

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research, 21*(3), 494-504. doi:10.1101/gr.112730.110

Hajishengallis, G., Darveau, R. P., & Curtis, M. A. (2012). The keystone-pathogen hypothesis. *Nature Reviews Microbiology, 10*(10), 717-725. doi:10.1038/nrmicro2873

Helbing, C. C., & Hobbs, J. (2019). Environmental DNA standardization needs for fish and wildlife population assessments and monitoring. Canadian Standards Association. *Retrieved on 10.02.2022 from https://www.csagroup.org/wp-content/uploads/CSA-Group-Research-Environmental-DNA.pdf.*

Hestetun, J. T., Lanzén, A., & Dahlgren, T. G. (2021a). Grab what you can—an evaluation of spatial replication to decrease heterogeneity in sediment eDNA metabarcoding. *PeerJ, 9*, e11619. doi:10.7717/peerj.11619

Hestetun, J. T., Lanzén, A., Skaar, K. S., & Dahlgren, T. G. (2021b). The impact of DNA extract homogenization and replication on marine sediment metabarcoding diversity and heterogeneity. *Environmental DNA, 3*, 997–1006. doi:10.1002/edn3.223

Hiergeist, A., Reischl, U., & Gessner, A. (2016). Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *International Journal of Medical Microbiology, 306*(5), 334-342. doi:10.1016/j.ijmm.2016.03.005

Hornick, K. M., & Buschmann, A. H. (2018). Insights into the diversity and metabolic function of bacterial communities in sediments from Chilean salmon aquaculture sites. *Annals of Microbiology, 68*(2), 63-77. doi:10.1007/s13213-017-1317-8

Kalle, E., Kubista, M., & Rensing, C. (2014). Multi-template polymerase chain reaction. *Biomolecular Detection and Quantification, 2*, 11-29. doi:10.1016/j.bdq.2014.11.002

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

Kennedy, K., Hall, M. W., Lynch, M. D., Moreno-Hagelsieb, G., & Neufeld, J. D. (2014). Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Applied and Environmental Microbiology, 80*(18), 5717-5722. doi:10.1128/aem.01451-14

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2017). DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *Plos One, 12*(6), e0179443. doi:10.1371/journal.pone.0179443

Lejzerowicz, F., Esling, P., & Pawlowski, J. (2014). Patchiness of deep-sea benthic Foraminifera across the Southern Ocean: Insights from high-throughput DNA sequencing. *Deep Sea Research Part II: Topical Studies in Oceanography, 108*, 17-26. doi:10.1016/j.dsr2.2014.07.018

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

Odelberg, S. J., Weiss, R. B., Hata, A., & White, R. (1995). Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. *Nucleic Acids Research, 23*(11), 2049-2057. doi:10.1093/nar/23.11.2049

Parsons, T. R., Takahashi, M., & Hargrave, B. (1977). *Biological oceanographic processes* (2nd ed.). Oxford: Pergamon Press.

Phillips, G. R., Anwar, A., Brooks, L., Martina, L. J., Miles, A. C., & Prior, A. (2014). Infaunal quality index: Water Framework Directive classification scheme for marine benthic invertebrates. *Retrieved on 01.01.2022 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/314673/Water_Framework_Directive_classification_scheme_for_marine_benthic_invertebrates_-_report.pdf*

Smyth, R. P., Schlub, T. E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., Davenport, M. P., & Mak, J. (2010). Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, *469*(1), 45-51. doi:10.1016/j.gene.2010.08.009

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Taberlet, P., Bonin, A., Zinger, L., & Coissac, É. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press.

Thrush, S., Hewitt, J., Pilditch, C., & Norkko, A. (2021). *Ecology of Coastal Marine Sediments: Form, Function, and Change in the Anthropocene*. Oxford: Oxford University Press.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., et al. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

Zaiko, A., Greenfield, P., Abbott, C., von Ammon, U., Bilewitch, J., Bunce, M., et al. (2022). Towards reproducible metabarcoding data: Lessons from an international cross-laboratory experiment. *Molecular Ecology Resources*, *22*(2), 519-538. doi:10.1111/1755-0998.13485

Zylstra, P., Rothenfluh, H. S., Weiller, G. F., Blanden, R. V., & Steele, E. J. (1998). PCR amplification of murine immunoglobulin germline V genes: strategies for minimization of recombination artefacts. *Immunology and Cell Biology, 76*(5), 395-405. doi:10.1046/j.1440-1711.1998.00772.x

Publication:

Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture -An inter-laboratory study

Verena Dully[a,] Heinrich Balliet[a], Larissa Frühe[a], Martin Däumer[b], Alexander Thielen[b], Sheena Gallie[c], Iain Berrill[d], Thorsten Stoeck[a,*]

[a]*Ecology Group, Technische Universität Kaiserslautern, Germany*

[b]*SeqIT, Laboratory for Molecular Diagnostics and Services, Kaiserslautern, Germany*

[c]*Scottish Sea Farms, Stirling, Scotland, United Kingdom*

[d]*Scottish Salmon Producers Organization, Edinburgh, Scotland, United Kingdom*

[*]*corresponding author*

From:

Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., Stoeck, T. (2021). *Ecological Indicators, 121*, e107049, doi:/10.1016/j.ecolind.2020.107049

Abstract

Environmental DNA metabarcoding of benthic bacterial communities emerged as a very powerful technology to assess environmental disturbance effects of coastal salmon aquaculture. A prerequisite for the implementation of this approach into compliance monitoring regulations is its robustness and reproducibility of obtained results. In the framework of regular compliance monitoring of a Scottish salmon farm, we therefore collected sediment samples along a transect from the salmon cages to reference sites in duplicates (biological replicates). Aliquots of both biological replicate samples were then processed by two different laboratories using independently the same eDNA metabarcoding protocol (technical replicates). Measures of alpha diversity and beta diversity, as well as taxonomic profiles of benthic bacterial communities were highly congruent among technical replicates, which even showed less variations than the biological replicates that were processed within each laboratory. Both technical replicate datasets identified the same bacterial indicator taxon groups and ASVs that are characteristic for the environmental quality (EQ) categories to which each of the samples was assigned based on traditional macroinvertebrate biomonitoring of the same samples. In a supervised machine learning (SML) approach, we could classify all individual samples from one technical replicate dataset into the correct EQ category using all samples of the other technical replicate dataset as a training dataset for the SML algorithm. We conclude that eDNA metabarcoding is sufficiently robust that different laboratories come to the same conclusions regarding officially regulated action criteria for environmental impact assessments in salmon aquaculture.

1. Introduction

Coastal marine ecosystems are subjected to a variety of industrial and recreational activities, which may compromise ecosystem function (ing) and the maintenance of ecosystem services. One example for industrial activities is aquaculture in coastal waters, which is a rapidly growing sector of seafood industry, with Atlantic salmon (*Salmo salar* L. 1758) being among the most important farmed finfish (FAO, 2018). The rapid growth of this industry has yielded significant socio-economic benefits but is accompanied by increasing environmental impact. Among others, this includes the seabed effects (benthic footprint) resulting from salmon aquaculture. Farmed salmon are maintained at up to 25 kg/m3 in surface-based sea-cage systems in coastal waters (Hvas et al., 2017) and fed with large amounts of a protein-rich diet. A proportion of this feed sinks down to the seafloor,

where it, together with fish faeces, results in an organic enrichment of the receiving benthic environment. This triggers a number of physical, chemical and biological processes on the seafloor and has lasting effects on the benthic ecosystem. These effects are reviewed in detail in the literature such as in Forrest et al. (2007), Gowen and Bradbury (1987) and (Keeley et al., 2013). In brief, when waste deposition exceeds the natural rate of organic material breakdown, a layer of fine-grained material with a high content of particulate organic material (POM) settles on top of the sediment. The seabed eventually becomes acidified and oxygen depleted because of microbial degradation processes. Toxic gases such as hydrogen sulfide and methane may be produced. Geochemical changes in the seabed structure are typically accompanied by changes in epifaunal and infaunal communities: less resistant benthic animals are wiped out and replaced by fewer, more tolerant organisms, which then become more abundant. In extreme cases, a layer of chemoautotrophic bacteria establishes on sediment surfaces and formerly species-rich sediments in the vicinity of farm cages become azoic (Brown et al., 1987; Holmer et al., 2005; Keeley et al., 2012). In addition to organic enrichment, also trace metals originating from antifouling paint for cages or from feed additives may accumulate in sediments beneath fish cages. Such chemical substances may also affect the composition of benthic communities (Burridge et al., 2010). As a rule, this depositional footprint of a salmon farm is mostly localized within the first 25–100 m from the point of discharge below the cages in direction of the prevailing current (Keeley et al., 2012). With increasing distance from the cages, the environmental effects decrease gradually until they have reached conditions as in an undeveloped control site.

To maintain coastal ecosystem function(ing) under such (and other) industrial and recreational activities, strict national and international monitoring regulations are in place. The backbone of these monitoring programmes is the biological component. Traditionally, benthic macroinvertebrates are collected from sediment samples, and microscopically identified. Based on the bioindicator qualities of individual species, an environmental quality (EQ) status for the sampling site is inferred from the macroinvertebrate inventory of this site. In case of salmon farms, this includes for example the Infaunal Trophic Index (ITI) or the Infaunal Quality Index (IQI) as defined in the Water Framework Directive classification scheme for marine benthic invertebrates (Phillips et al., 2014). Recently, the interrogation of taxonomic marker genes from environmental DNA (eDNA) extracted from sediment samples emerged as a very powerful, faster, less expensive and more efficient (as automatable for high-throughput) method to complement or possibly even replace the

traditional microscopy-based environmental monitoring, not only for aquaculture (Cordier et al., 2017; Forster et al., 2019a; Frühe et al., 2020; Keeley et al., 2018; Pawlowski et al., 2014; Stoeck et al., 2018a; 2018b), but also to assess the impact of other activities in marine (Aylagas et al., 2017; Borja, 2018; Lanzén et al., 2016) and freshwater (Apothéloz-Perret-Gentil et al., 2017; Bagley et al., 2019; Rivera et al., 2018) ecosystems. In particular, the analysis of benthic bacteria as bioindicators is the best option available for assessing the impact of aquaculture on the marine benthic environment, reflecting the results of EQ assessments based on traditional surveys of benthic macroinvertebrates on the same samples (Cordier et al., 2018; Frühe et al., 2020; Keeley et al., 2018; Stoeck et al., 2018a). Therefore, the next following logic step would be the implementation of eDNA-based biomonitoring into routine practice and compliance monitoring regulations.

A mandatory prerequisite towards this next step, however, is that eDNA-based biomonitoring is robust and reproducible. Different laboratories need to come to the same results and conclusions regarding the EQ class of sample sites under monitoring. Even when the same standardized laboratory protocol is applied, possible bias could be introduced during e.g. PCR amplification of the target gene or during library preparation and sequencing (Berry et al., 2011; Boers et al., 2019; Kennedy et al., 2014). This may lead to incongruent results and conclusions for the same samples. To test the robustness and reproducibility of eDNA metabarcoding to assess the EQ at a salmon aquaculture site, samples were processed independently in two different laboratories using the same standardized protocol. Fig. 1 provides an overview of the conceptual study design and the workflow from sampling to data analysis and interpretation. The results obtained from both laboratories (measures of alpha and beta diversity, community composition, discriminative bacterial ASVs, and prediction of EQ for each sample using supervised machine learning) were then compared among each other and with the corresponding results obtained from traditional macrofauna-based monitoring of the same samples.

2. Material and methods

2.1. Sampling

Benthic sampling followed standard protocols as described in e.g. Frühe et al. (2020). In brief, during a compliance monitoring survey of a Scottish salmon farm (Shetland) sediment was collected at six stations along a northwest (NW) transect from the cage edge to a reference site in the direction of the prevailing current flow. This transect included the northwesterly cage edge site (CE), 50 m distance from cage (Allowable Zone

of Effect, AZE, minus 10 m), 60 m (AZE), 70 m (AZE plus 10 m) and two reference sites (Ref1 and Ref 2) at about 735 m and 475 m distance from the CE. At each site, two van Veen grabs (0.045 m2 area) were taken as biological replicates for this study (Fig. 1). From these two grabs, ca. 10 g of surface sediment was collected from several areas within the grab. These two biological replicates were then preserved with Qiagen's LifeGuard solution (equal volume buffer to sediment) to preserve nucleic acids until further processing for eDNA metabarcoding. The remaining sediment of both grabs was used for macrofauna morphotaxonomic inventories (Fig. 1). Therefore, the sediment was washed through a 1-mm sieve and the residue fixed in 4% borax-buffered formaldehyde prior to macrobenthic sorting and counting. The sieve-retained fauna was identified to species level by a certified Scottish consulting company and provided in the official benthic report to Scottish regulators. As environmental quality index we calculated IQI version 2 (IQIv. II) from pooled macrofauna samples, which is based on Phillips et al. (2014) and which is commonly used in compliance monitoring of coastal waters in the United Kingdom.

2.2. Inter-laboratory comparisons

Laboratory work for eDNA metabarcoding was conducted independently by two different laboratories (Eco-Lab; SeqIT-Lab) using the same standardized molecular protocol, which was described previously (Frühe et al., 2020) and is given here in brief. The samples processed independently by Eco-Lab and SeqIT-Lab are considered as technical replicates. The comparative statistical analyses described below serve to evaluate the (in)congruencies among these technical replicates, in comparison to the (in)congruencies among true biological replicates obtained from sampling (Fig. 1).

2.3. Molecular lab work

Environmental DNA was obtained from sediment samples using the PowerSoil DNA kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. As DNA metabarcodes, we obtained the ca. 450 bp long hypervariable V3-V4 region of the bacterial SSU rRNA gene. The PCR mixture included the primer pair Bakt_341F (CCTACGGGNGGCWGCAG) and Bakt_805R (GACTACHVGGGTATCTAATCC), and the PCR protocol employed an initial activation step of NEB's Phusion High-Fidelity DNA polymerase at 98 ◦C for 30 s, followed by 27 identical three-step cycles consisting of 98 ◦C for 10 s, 62 ◦C for 30 s, and 72 ◦C for 30 s; then a final 5-min extension at 72 ◦C (Herlemann et al., 2011). From the resulting PCR products, sequencing libraries were constructed using the NEB Next® Ultra™ DNA Library Prep Kit for Illumina (NEB,

USA). The quality of the libraries was assessed with an Agilent Bioanalyzer 2100 system. V3-V4 libraries of both laboratories were sequenced on independent Illumina runs (Illumina MiSeq platform, generating 2x300-bp paired-end reads). Sequences (fastq files and metadata) are available at NCBI's BioProject database under project number PRJNA647362.

2.4. Sequence data processing

Sequences were processed using the Divisive Amplicon Denoising Algorithm (DADA2) (Callahan et al., 2016) as described for hypervariable taxonomic marker genes from metabarcoding studies (Forster et al., 2019b) with the model trained on the independent Illumina runs and the following criteria: bacterial V3-V4 sequences were filtered using *filterAndTrim* with truncLen = 225 and maxEE = 1. The truncation length criterion was determined by choosing the sequence position at which Phred assigned a quality score of ≥ 30 (Q3) for at least 51% of all reads in a dataset (=base call accuracy 99.9%, (Ewing and Green, 1998). For maxEE we chose the most stringent value. We chose these settings to maximize the quality of the final sequence reads used for downstream analyses. Bacterial V3-V4 sequences were merged using 20 base pairs overlap with an allowed mismatch of 2. To minimize ecologically uninformative noise, only amplicon sequence variants (ASVs) with at least 100 reads were maintained for downstream analyses.

2.5. Comparisons of datasets

To compare the agreement between the technical replicate datasets that were processed independently in two different laboratories, we used the following criteria that are relevant for the biomonitoring of aquaculture impacts: Alpha and beta diversity measures of bacterial community structures, taxonomic assignments of ASVs, potential bacterial indicator groups for environmental quality and ASVs discriminant for environmental quality. Finally, we determined a linear regression model, with which the environmental quality for a random sample of one of the two datasets could be predicted by using the full second dataset. To interpret the data obtained from these analyses, we then compared the obtained results from the technical replicate analyses with results obtained from the same analyses conducted with the biological replicates processed within each of the two laboratories (Fig. 1).

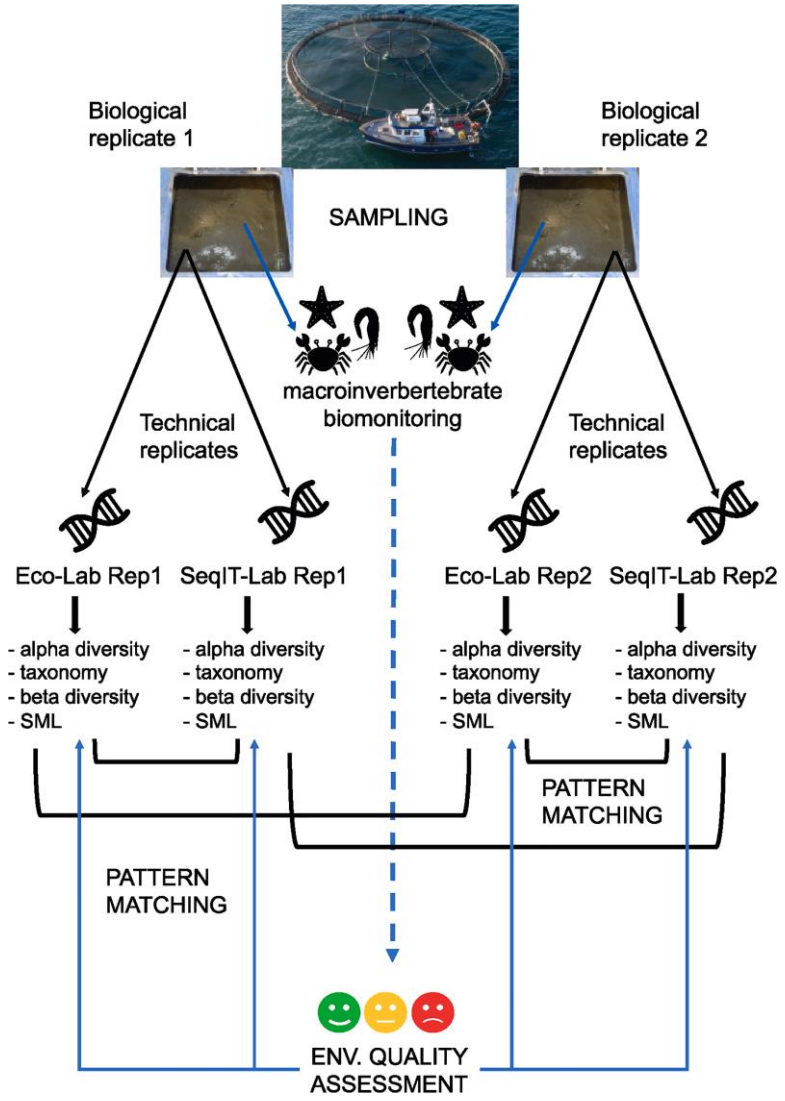


Fig. 1. *Conceptual study design and workflow. Two biological sediment replicate samples were taken at each of the six sampling locations for macroinvertebrate identification and eDNA extraction. Each biological eDNA replicate was then split up among two laboratories for eDNA metabarcoding of bacterial communities. Community information was obtained for each of the four replicates of each sampling location and results were compared within laboratories (biological replicates, intra-laboratory comparisons) and among laboratories (technical replicates, intra-laboratories comparisons). Macrofauna-obtained biomonitoring results were used as an external reference for pattern matching among all eDNA replicates. For details, please see Methods section.*

Analyses of bacterial community structures

All community structure analyses were conducted in R v. 3.5.3 using the program packages vegan (Oksanen et al., 2019), and ggplot2 (Wickham, 2016) for graphical visualization. Normalized (relative abundance) ASV richness, Shannon Index and Simpson Index were calculated for each sample as a measure of alpha diversity. Bray-Curtis (BC) Index was calculated as a measure of beta diversity. BC values were subjected to nonmetric multidimensional scaling (NMDS) using the *metaMDS* function with square root

transformation. Polygons were used to visualize separation of individual clusters. Using the *envfit* function, the macrofauna-derived IQI was fit to the NMDS ordination to evaluate the relation between bacterial community structures and macrofauna-derived environmental quality.

Taxonomic assignment

Taxonomy was assigned to resulting ASVs with the *assign_taxonomy* function (Galaxy v. 1.9.1.0) using the method UCLUST (Caporaso et al., 2010). Assignment was based on the Greengenes reference database (McDonald et al., 2012). Only ASVs with a sequence similarity >90% to any reference sequence were kept for further analyses. Non-target ASVs were excluded from further analyses. The same applies to ASVs which occurred only once and exclusively in one sample, and thus, may be artifactual sequences (Bokulich et al., 2013). The resulting ASV-to-sample matrix for each DNA marker was then used for all statistical downstream analyses, following normalization of the read counts using vegans *decostand* function. Currently, the Scottish Environmental Protection Agency (SEPA) is evaluating a new seabed monitoring approach to investigate the impacts of marine fish farms in Scottish coastal waters (SEPA, 2018). In this new approach, SEPA proposes a defined ellipse area around a fish farm, outside of which good ecological conditions (IQI > 0.64) need to be established and within of which severe to moderate disturbance (IQI < 0.64) is acceptable. Because of the relevance of this EQ threshold for biomonitoring, we analyzed the relative abundance of individual identified bacterial taxon groups (orders) below and above an IQI of 0.64 and compared the results of both technical replicate datasets.

Predicting environmental quality status from eDNA metabarcoding of bacterial communities using supervised machine learning

The aim of supervised machine learning (SML) was to evaluate the agreement between the predicted EQ from a randomly chosen individual technical replicate processed in one lab and in the other lab. Predictive models were trained using the random forest (RF) algorithm (Breiman, 2001) implemented in the randomForest v. 4.6.14 (Liaw and Wiener, 2002) package for classification and regression. IQI values were predicted independently for all the samples (testing datasets) using a predictive model trained on the technical replicates (training dataset) from two different laboratories and for biological replicates from within a laboratory.

Model construction was performed with the training datasets using the leave-one-out approach. This means, a training dataset consisted of all samples of a specific dataset with one sample being held out as test data. This was repeated until each sample was held out ten times. This cross-validation for each sample employed the default *m* hyperparameter and 2000 trees each. After predicting IQI values independently for each sample, the models were averaged to measure the overall importance of each ASV, which we then compared among the two technical replicate samples processed in the two different laboratories.
Variable importance was measured using percentage of increase of mean squared error (MSE). We then compared the agreement of each technical replicate prediction with results obtained from the same analyses conducted with the biological replicates processed within each of the two laboratories. Therefore, we applied a linear model of IQI prediction results comparing the Eco-Lab predictions with the SeqIT-Lab predictions adding 95% confidence intervals. The analyses yielding the highest agreement among replicate datasets was the one associated with the highest R2 value.

3. Results

3.1. Sequence data overview

We obtained 1,569,059 and 1,331,718 raw sequences for the Eco- Lab and the SeqIT-Lab dataset, respectively. After data cleaning, 672,767 (Eco-Lab) and 462,342 (SeqIT-Lab) high-quality target sequences remained for downstream analyses, which grouped into 2232 and 2230 bacterial ASVs, respectively. Rarefaction curves of both datasets indicated (near) saturation for all samples and were highly similar for both datasets (Fig. 2).

3.2. Macrofauna reference data

Based on microscopic identification of macroinvertebrate indicator species, the environmental quality ratio used to derive the ecological status in marine sediments in the UK following IQI (Infaunal Quality Index) values were obtained (Fig. 3): For the cage edge site, IQI was 0.38 (poor ecological status, major disturbance); for 50 m distance from cage, IQI was 0.41 (poor ecological status, major disturbance); for 60 m distance from cage, and for Reference sample 1 a good ecological status (slight disturbance, IQI 0.66 and 0.74, respectively) was attested; for 70 m distance from cage, IQI was 0.51 (moderate ecological status, moderate disturbance). For Reference sample 2, we obtained a high ecological status (negligible disturbance, IQI 0.75).

Thus, the three sampling sites CE, 50 m and 70 m were below the moderate environmental quality threshold of IQI 0.64, and the samples 60 m, Ref1 and Ref2 had better than moderate environmental quality (IQI > 0.64). A dendrogram illustrating sample grouping according to the Bray-Curtis dissimilarities in macroinvertebrate species composition between the sampling stations (Fig. 4a) shows two groups. The macroinvertebrate communities at sites CE and 50 m appear as most similar, with the 70 m sample distantly clustered. This group represents sampling sites with an IQI > 0.64, whereas the second group unites the samples with an IQI < 0.64 (Ref 1, Ref 2 and 60 m). In the ordination plot, this clustering was also evident (Fig. 4b). In addition, the ordination shows the variability and distribution of biological replicates from the same sampling sites, which will become relevant for the comparison of variation in biological eDNA replicates for bacterial community analyses.

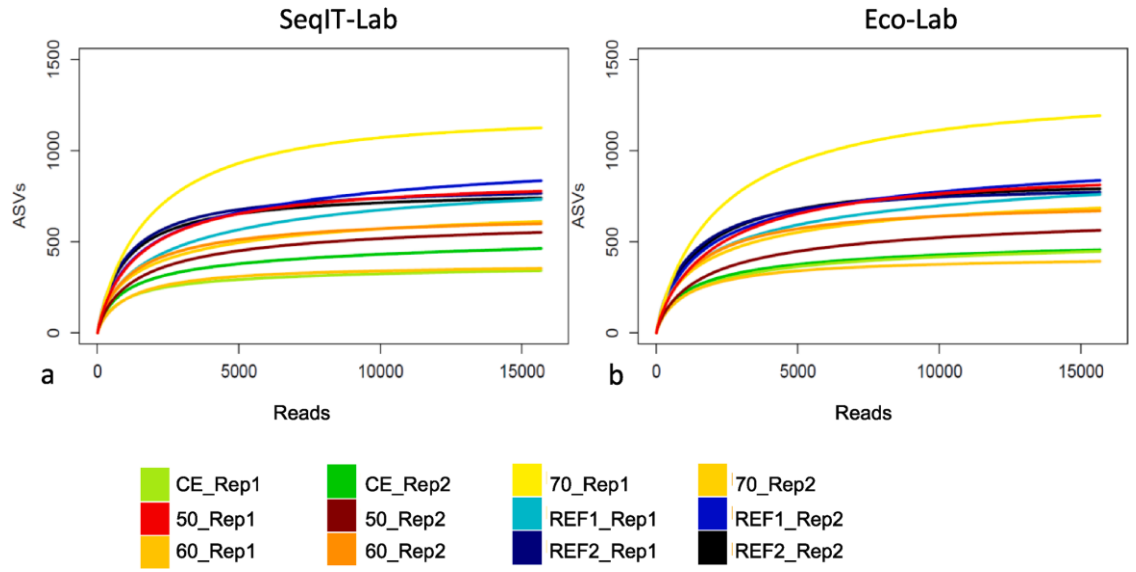


Fig. 2. *Rarefaction analyses of bacterial V3-V4 amplicons for all individual replicate samples for the technical replicate dataset processed at SeqIT-Lab (a) and the one processed at Eco-Lab (b).*

3.3. eDNA metabarcoding inter-laboratory comparisons

3.3.1. Alpha diversity of bacterial communities

In absolute and in relative values, bacterial ASV richness (Fig. 5a), Shannon Index (Fig. 5b) and Simpson Index (Fig. 5c) are highly congruent when comparing technical replicates processed by Eco-Lab and by SeqIT-Lab. All three measures had the general tendency to increase from cage edge samples towards reference sites. This increase was lowest for the Simpson Index, which showed only little sample-to-sample variation

compared to ASV richness and Shannon Index. As a rule, variations among technical replicates are smaller than variations among biological replicates that were processed in the same laboratory for all three alpha diversity measures. In all cases, the lowest diversity was calculated for the 60 m Rep1 sample.

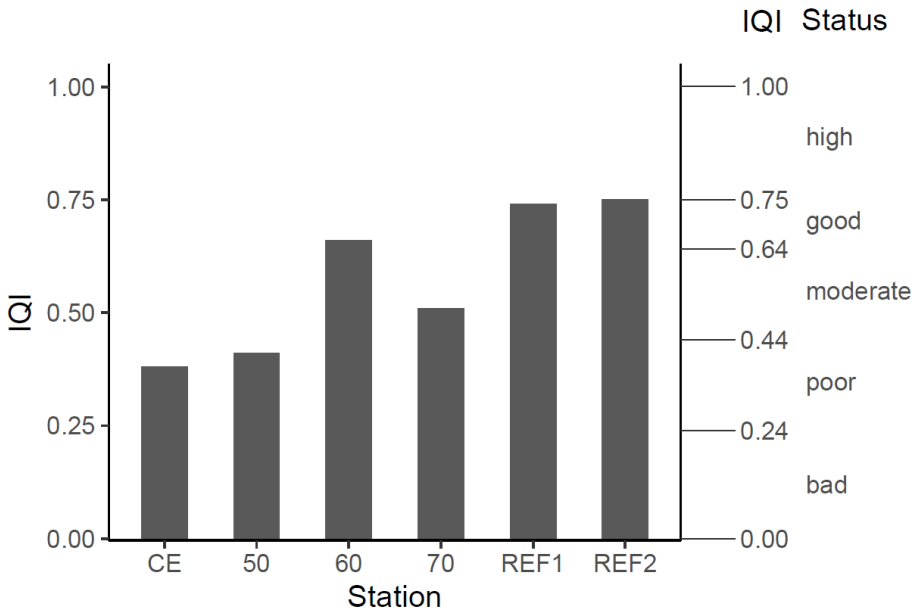
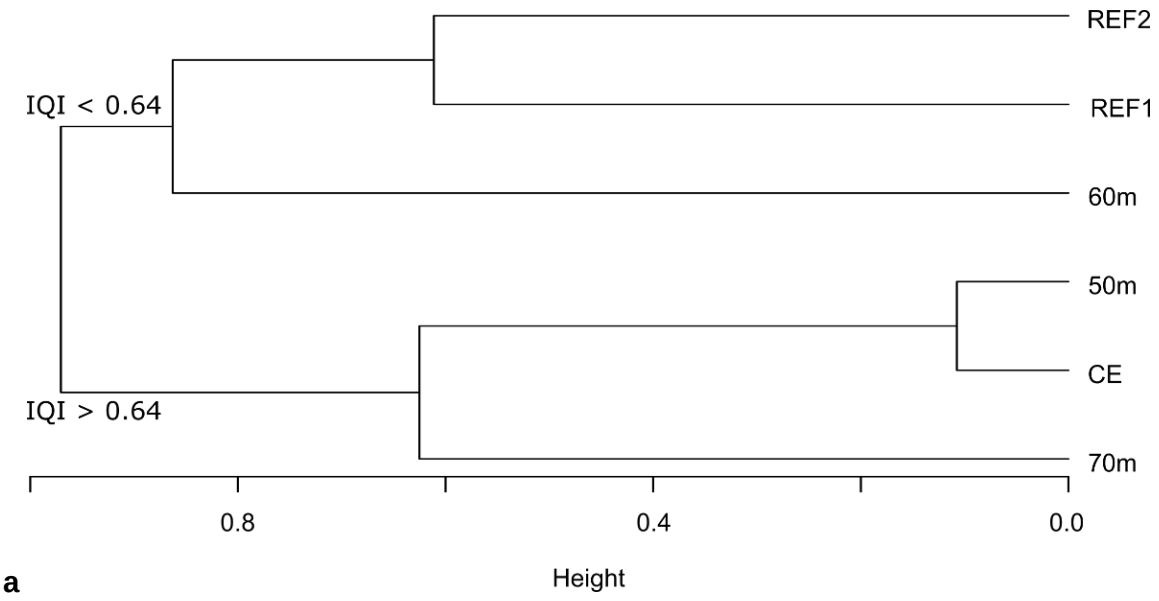


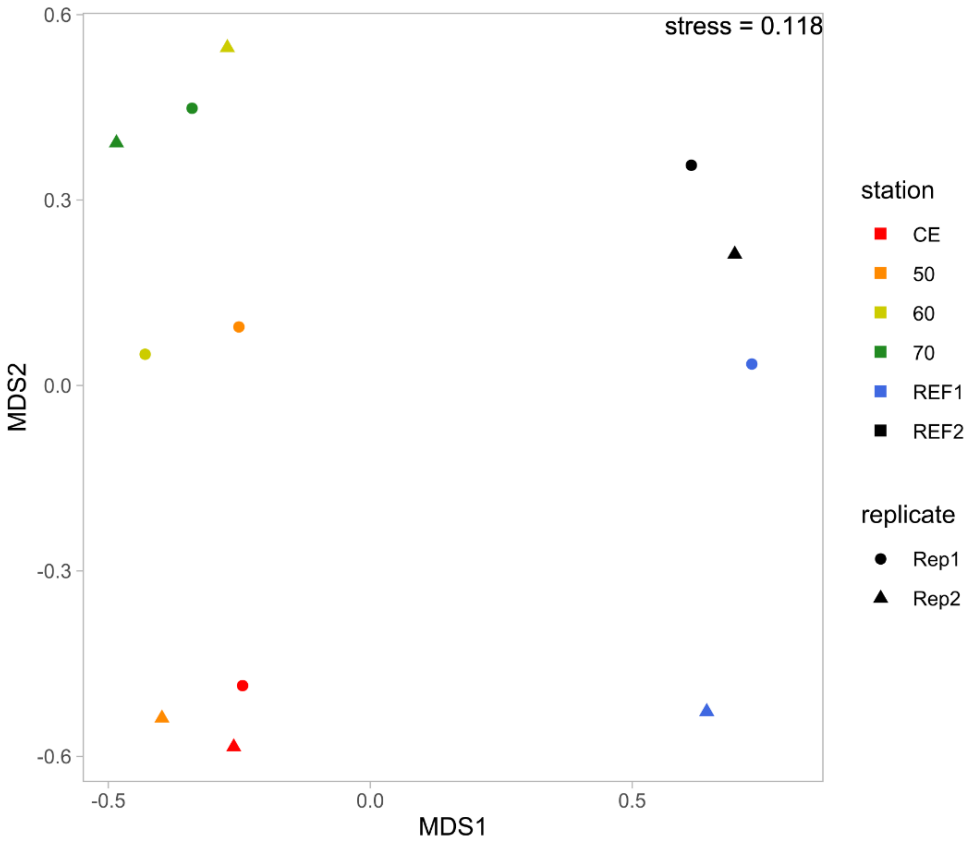
Fig. 3. *Infaunal Quality Index (IQI) for the locations along the sampled transect from the northwesterly cage edge (CE) of the salmon farm under study towards two reference (REF) sites. IQI was calculated based on macroinvertebrate bioindicators.*

3.3.2. Beta diversity of bacterial communities

Nonmetric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities (stress: 0.135) revealed that benthic bacterial community structures were more similar among technical replicates processed by two different labs than among biological replicates that were processed within each of these labs (Fig. 6). All four replicates from any sample within the distance gradient form clearly distinguishable clusters in the plots two-dimensional area, which do not overlap with the cluster formed by the four replicates of any other sample. Distribution of samples in the two-dimensional plot area followed a gradient along both NMDS axes. While axis 2 correlated significantly with the environmental quality index IQI, which was obtained from macrofauna reference data (Fig. 3) ($R^2 = 0.56$, $p = 0.001$), we lack an explanation explaining the gradient along axis 1.



a



b

Fig. 4*. Similarity of macroinvertebrate communities at the sampling locations based on a Bray-Curtis distance matrix. The dendrogram (a) shows two clusters, one of which represents sampling locations with an IQI < 0.64 (CE, 50 m, 70 m) and one representing sampling locations with an IQI > 0. 64 (60 m, Ref 1, Ref 2). The nonmetric multidimensional scaling ordination (b) illustrates the variability of biological replicates, but also confirms the partitioning of diversity among the sampling locations as represented in the dendrogram.*

In accordance with results of the NMDS analyses, the proportion of bacterial ASVs shared among technical replicates processed by two different laboratories was notably higher compared to the proportion of bacterial ASVs shared among biological replicates processed within each of the two laboratories (Fig. 7). On average, the technical replicates shared 78% and 79% of all bacterial ASVs obtained for all replicates from dataset Rep1 and from dataset Rep2, respectively. Whereas, the biological replicates (Rep1 and Rep2) within a laboratory shared only 49% (Eco-Lab) and 46% (SeqIT-Lab) of bacterial ASVs found in the respective datasets. The vast majority of exclusive (non-shared) ASVs are low-abundant ASVs (<0.1% relative contribution).

3.3.3. Taxonomic assignments of bacterial ASVs

To compare biological and technical replicates on higher and lower taxonomic levels, we, in detail, analyzed the composition of bacterial communities across three taxonomic levels: Phylum-, order- and genus-level (Fig. 8). Taxonomies on class and family level are available as supplementary online material[*]. In all three taxonomic categories (Fig. 8a-c), the taxonomic assignment of bacterial ASVs was more consistent among technical replicates obtained from two different laboratories (Eco-Lab and SeqIT-Lab) than among biological replicates processed within each of the two laboratories. On phylum-level (Fig. 8a), most ASVs were assigned to Firmicutes (33.3%) and/or Proteobacteria (50.6%) in all samples. At sites close to the salmon farm (cage edge, 50 m and 60 m samples) Actinobacteria and Bacteroidetes contributed notably to the bacterial community composition (averaged contribution at these sites: 6% and 9.8%, respectively). With sites more distant from the salmon cages, the relative abundance of these taxon groups decreased, whereas the relative abundance of Proteobacteria increased notably. It is conspicuous that sample “cage edge Rep2” included a high proportion (23%) of ASVs assigned to the phylum Fusobacteria. This phylum occurred also at other near-cage sites, far-cage sites (70 m, Ref1 and Ref2).

[*] All supplementary files are additionally available at the appendix of this dissertation

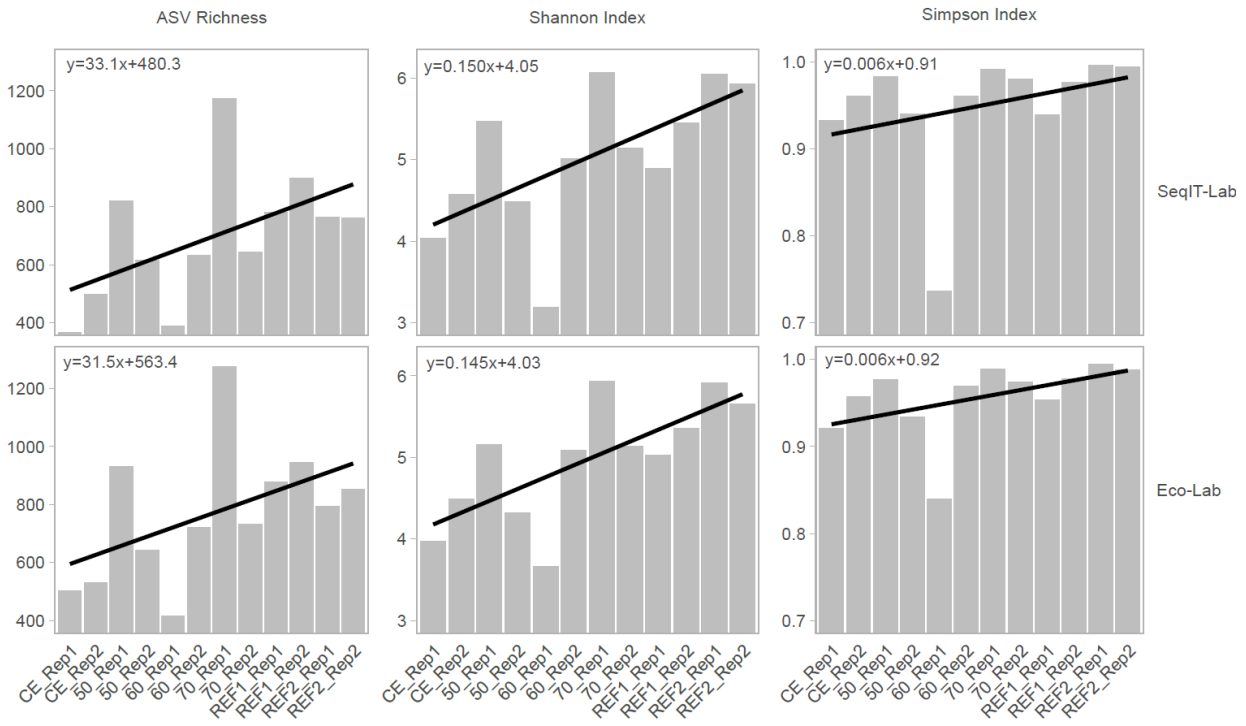


Fig. 5. *Alpha diversity measures of bacterial communities (eDNA metabarcoding) for biological replicates (Rep1 and Rep2) and their technical replication processed in Eco-Lab and SeqIT-Lab. Lines in plots represent linear regression model for each dataset with model equation shown in each plot.*

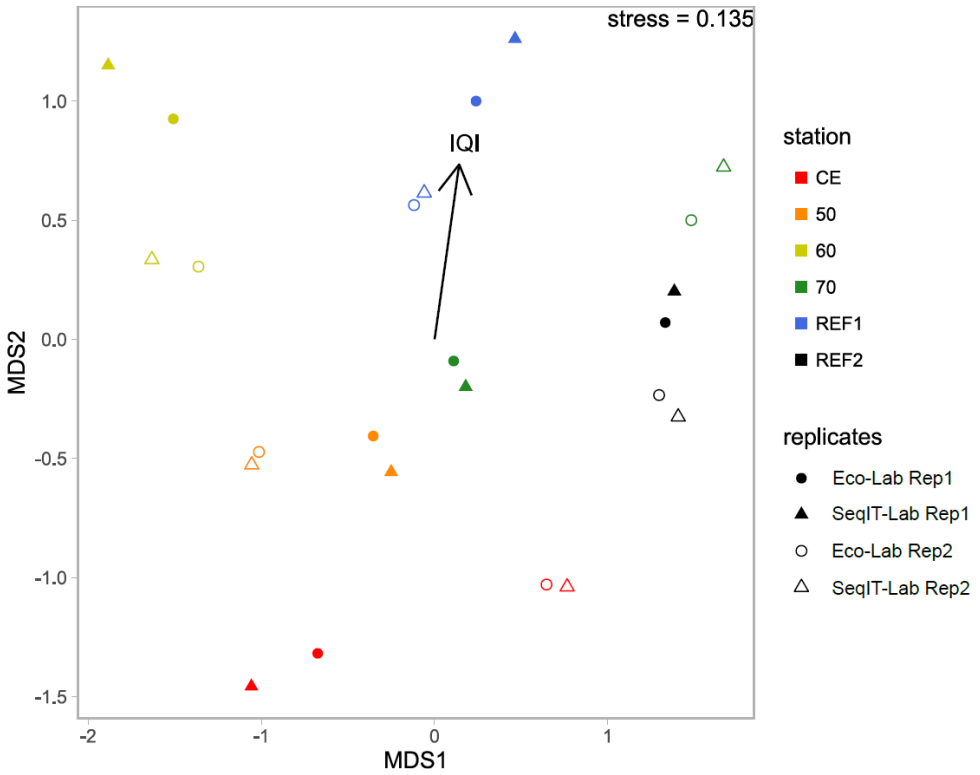


Fig. 6. *Nonmetric multidimensional scaling (NMDS) ordination of bacterial communities (eDNA metabarcoding) for biological and technical replicates. The macrofauna-obtained environmental quality measure IQI (Fig. 3) correlated significantly with axis 2.*

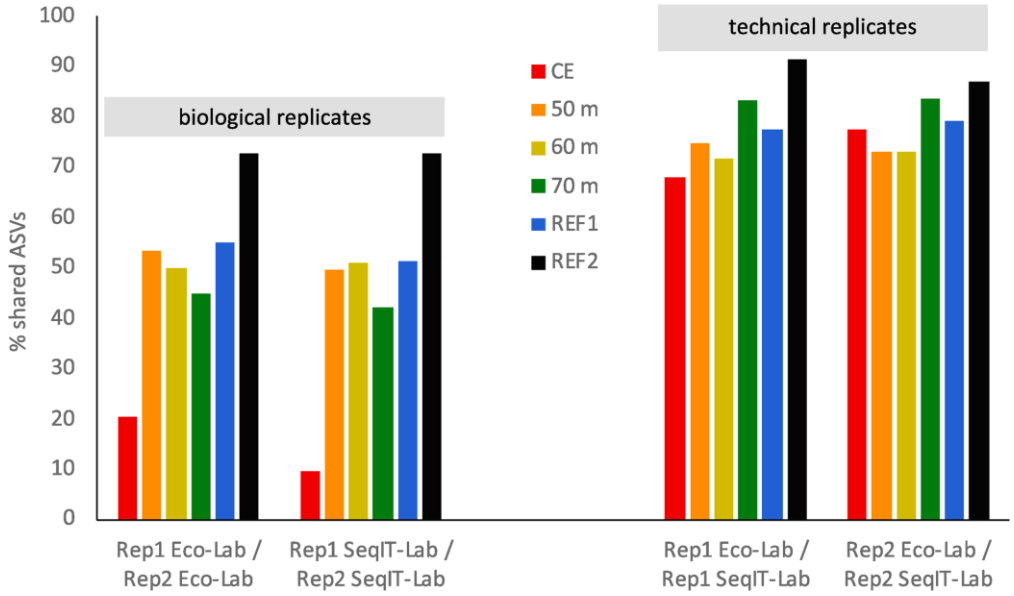


Fig. 7. *Proportion of bacterial ASVs shared among biological replicates (Rep1 and Rep2) processed within each of the two laboratories (Eco-Lab and SeqIT-Lab) and among technical replicates of Rep1 and Rep2 processed by each of the two laboratories.*

On order-level (Fig. 8b), we here only show the 15 most diverse (=highest relative proportion of ASVs assigned to these classes). Further families are categorized as "others". The general trend was a higher proportion of ASVs assigned to Clostridiales (50.2%, mainly Clostridiaceaea and Lachnospiraceae), Coriobacteriales (5%, mainly Coriobacteriaceae), Bacterioidales and Campylobacterales (6.6%) at near-cage sites (CE, 50 m and 60 m, given %-values are averaged across the three sites). At far-cage sites (70 m, Ref1 and Ref2 samples) ASVs assigned to Alteromonadales were notably more abundant compared to near-cage sites (averaged 9.2% vs. averaged 1.2%). ASVs assigned to the order Vibrionales did not show a clear pattern. ASVs assigned to this order were numerous at the CE site and in the 50 m sample (20.3%), notably rare in the 60 m sample (1.3%) and highly abundant at far-cage sites (averaged proportion of 58.4% at sites 70 m, Ref1 and Ref2 samples).

On genus level (Fig. 8c), we exemplarily focus on genera assigned to the Vibrionacea (order Vibrionales). Genera within the family Vibrionaceae showed a clear trend with increasing distance from the cage edge (Fig. 8c). *Allivibrio* was characteristic of near-cage sites (averaged 0.7%, CE, 50 m and 60 m), while *Vibrio* and *Photobacterium* had higher ASV proportions (averaged 12% and 10%, respectively) at far cage sites (70 m, Ref1, Ref2). It is noteworthy that a very high proportion of ASVs within the family Vibrionaceae remained taxonomically unassigned.

3.3.4. Bacterial indicator taxa for environmental quality assessment

To assess whether sediment samples processed independently by two different laboratories identify the same indicator taxa for an EQ threshold of 0.64, we analyzed the relative abundance distribution of bacterial families above and below the macrofauna-derived IQI 0.64 (Fig. 9). Biological replicates were combined. The indicator groups obtained by both labs (Fig. 9a and b) are highly congruent. In both datasets, >80% of all ASVs within the phyla Acidobacteria, Caldilineales, Campylobacterales, Chromatiales, Clostridiales, Desulfobacterales, Flavobacteriales, Fusobacteriales, Myxococcales, Sphaerochaetales, Thiohalorhabdales and Thiotrichales occurred at an IQI < 0.64 (poor to moderate environmental quality). The following orders were represented in notably higher abundances at an IQI > 0.64: Alteromonadales, Enterobacteriales, Ocenosprillales and Rhodobacterales. Vibrionales were nearly equally distributed at both sides of the IQI most of the Vibrionales at an IQI < 0.64 and Vibrio and Photobacterium accounting for most of the Vibrionales at an IQI > 0.64.

The Random Forest (RF) analyses identified ASVs that were discriminant of EQ category (Table 1). When comparing the ten strongest discriminant ASVs among the technical replicates from Eco-Lab and SeqIT-Lab, seven ASVs were assigned to the same lowest possible taxonomic rank: 5 ASVs belonging to Vibrionaceae (Vibrionales), one Sedimentibacter and one Lachnospiraceae (both Clostridiales). One further ASV assigned to Sphaerochaetaceae (Sphaerochaetales), which ranked on position 9 in the SeqIT-Lab dataset, ranked number 11 in the Eco-Lab dataset. Another ASV assigned to Photobacterium (Vibrionales) ranked number 9 in the Eco-Lab dataset, was found on position 21 in the SeqIT-Lab dataset. The cumulative variable importance is highly similar for both datasets: 50.89 for Eco-Lab and 50.09 for SeqIT-Lab.

3.3.5. Prediction accuracies for technical and biological replicates

To further assess the agreement among technical replicate samples that were processed independently by the two laboratories, we used RF to predict the IQI for individual samples (bacterial community structures) processed in one laboratory using the results obtained from the other laboratory as a reference (Fig. 10a). Linear regression modelling revealed a very high congruency of the two independently processed datasets ($R^2 = 0.91$, $p < 0.001$). Comparing the biological replicates within a laboratory (Fig. 10b), congruency was also highly accurate ($R^2 = 0.90$ and $R^2 = 0.86$ for datasets obtained from

Eco-Lab and from SeqIT-Lab, respectively, both $p < 0.005$), prediction agreement was not as accurate as for technical replicates processed independently in different laboratories.

4. Discussion

A central prerequisite for the implementation of novel methods in environmental compliance monitoring regulations includes the robustness and reproducibility of these methods (Golebiewski and Tretyn, 2020; Loeza-Quintana et al., 2020; Nicholson et al., 2020): it is mandatory that different laboratories analyzing samples for compliance monitoring come to the same objective conclusions when processing and interpreting the same samples. In the best-case scenario, this criterion is met, when variability of results obtained from the same samples processed independently in different laboratories is lower or equal to the variability of true biological replicates taken from the same sampling location (e.g. edge of a salmon cage) under monitoring. Our study demonstrated that this criterion is met using the described eDNA metabarcoding and data analysis protocol.

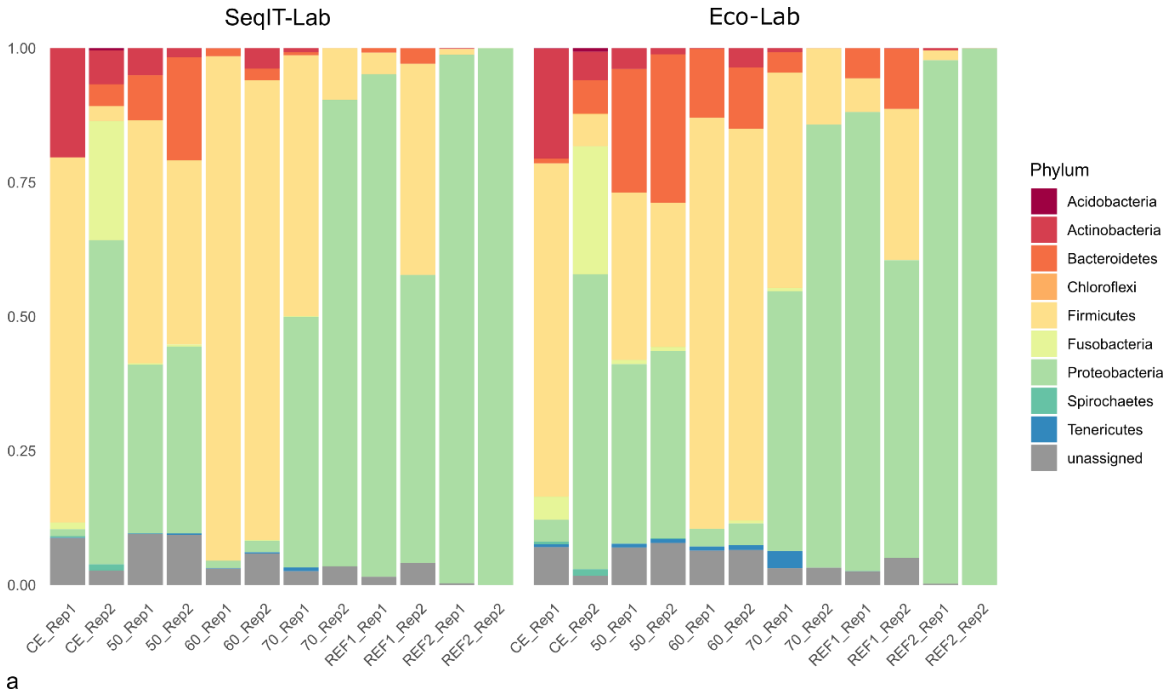


Fig. 8. *Taxonomic assignment of bacterial ASVs. The bars show the relative proportion of ASVs assigned to each of the different taxonomic entities. (a) Phylum-level taxonomy, (b) class-level taxonomy and (c) taxonomic assignment of genera within the family Vibrionaceae. Class- and order-level taxonomic bar charts are available as supplementary online material*[*].

[*] All supplementary files are additionally available at the appendix of this dissertation

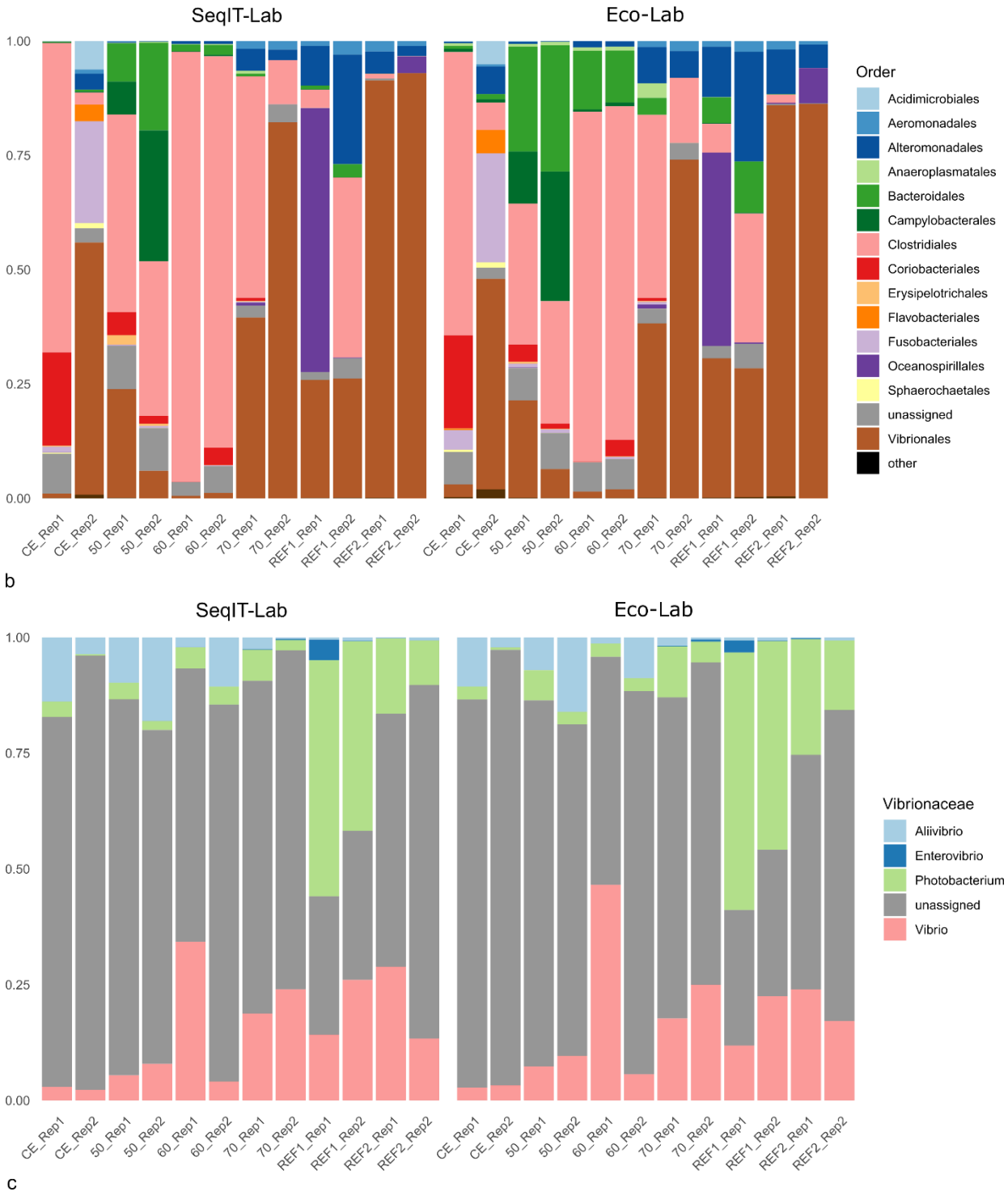


***Fig. 8** (continued)*

Sample-to-sample variation could have several technical reasons. Sources of technical biases in eDNA metabarcoding studies were analyzed previously and reviewed in detail (e.g. Golebiewski and Tretyn, 2020). They include sample collection, transport and storage, nucleic acid (NA) preparation and isolation, PCR amplification and sequencing, and data processing. In our study, we applied the currently best possible and

recommended practice to minimize technical bias. This includes for example: Safeguarding collected samples in a preservation buffer (Qiagen's LifeGuard solution); using a bead-beating NA extraction approach with a standardized kit, which in comparison to non-bead-beating NA extraction methods also obtains DNA from "difficult" taxon groups such Actinobacteria (Filippidou et al., 2015); using an engineered polymerase of high-fidelity for PCR amplification (Monroe et al., 2013); using the most common and best characterized taxonomic barcode (16S rRNA V3-V4 gene region) and primer set for bacterial communities (Golebiewski and Tretyn, 2020); using low cycle numbers and optimal annealing temperatures in PCR reactions (Smyth et al., 2010); using Illumina approved MID (Molecular IDentifiers)/primer combinations for library constructions to minimize the formation of secondary structures compromising Illumina sequencing (Berry et al., 2011); and using MIDs of different lengths, lowering cluster density, adding diverse library (PhiX) and use heat denaturation (96 ◦C for 2 min) to optimize cluster registration (=template generation) during Illumina MiSeq sequencing (Illumina, 2013). Also, our computational data processing pipeline applies the state-of-the art knowledge to minimize technical bias. This includes for example a Phred30 quality threshold for amplicon data (i.e. an error probability of 0.001, which means one read in a million will bear two errors, (Golebiewski and Tretyn, 2020)), which are in downstream analyses accounted for in the denoising procedure using dada2 (Callahan et al., 2016).

The importance of standardization of these steps (sample collection, transport and storage, NA preparation and isolation, PCR amplification and sequencing, data processing) cannot be overestimated in environmental metabarcoding studies (Schloss, 2018), especially when it comes to environmental quality assessments. The latter requires robust and reliable quality metrics to allowing for direct comparisons of samples (Golebiewski and Tretyn, 2020). Details regarding such an eDNA metabarcoding standard yet have to be established for coastal environmental in general and aquaculture compliance monitoring in specific. Even though there may be details that could be improved to further reduce technical biases in our study, all samples were subjected to the same protocol (in two different laboratories) and, thus, potential technical biases should not account for the differences observed when comparing intra-biological replicate variances with intra-technical replicate variances. Also, an insufficient sequencing depth (=numbers of reads generated per sample) may compromise results obtained from statistical analyses and their interpretation. For example, Gihring et al. (2012) demonstrated that an insufficient number of individuals captured per sample may influence alpha- and beta-diversity indices.

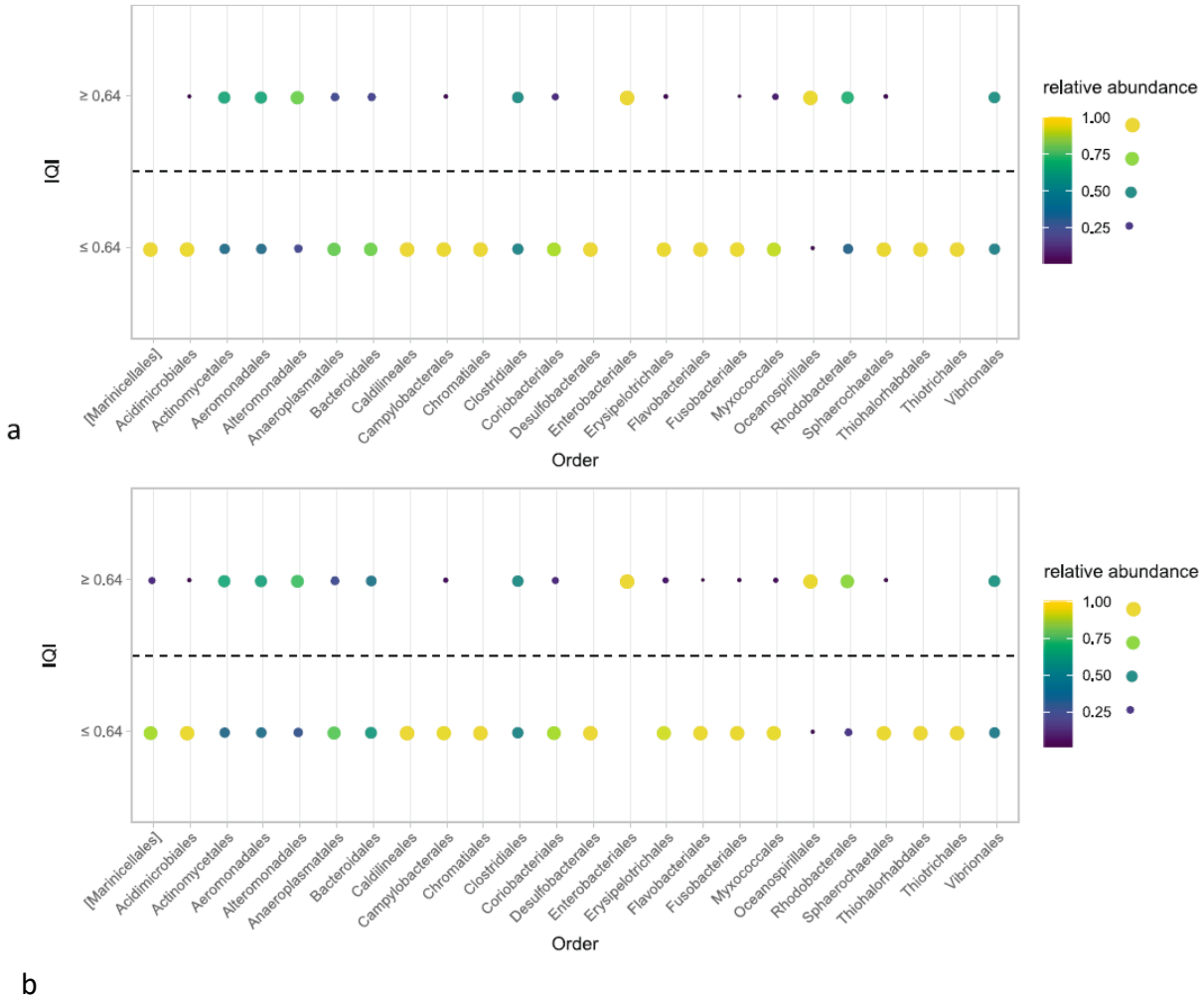


Fig. 9. *Cumulative relative ASV abundance within individual bacterial orders above and below the IQI threshold of 0.64. Size of nodes represents the proportion of the individual ASV populations within each class and together with the color code for relative abundances serves to enhance visualization of bacterial indicator orders. (a) SeqIT-Lab; (b) Eco-Lab.*

Table 1) *Ten most discriminant ASVs in Random Forest regression analyses to predict the environmental quality (EQ) category for each individual sample from one technical replicate dataset (Eco-Lab or SeqIT-Lab) using all samples of the other technical replicate dataset (SeqIT-Lab or Eco-Lab, respectively) as training data. Variable importance is the error increase in RF regression model when a specific ASV is removed from the analysis.*

Eco-Lab ASV	variable importance	taxonomic assignment	SeqIT-Lab ASV	variable importance	taxonomy assignment
1	6.67	Vibrionaceae	1	9.27	Vibrionaceae
2	6.60	Vibrionaceae	2	9.20	Vibrionaceae
3	6.09	Lachnospiraceae	3	7.07	Lachnospiraceae
4	5.88	Vibrionaceae	4	6.00	Sedimentibacter
5	5.41	Vibrionaceae	5	3.49	Vibrionaceae
6	5.08	Sedimentibacter	6	3.41	unid. bacterium
7	4.43	Vibrionaceae	7	3.41	Lachnospiraceae
8	3.79	Shewanella	8	3.16	Vibrionaceae
9	3.54	Photobacterium	9	2.62	Sphaerochaetaceae
10	3.40	Shewanella	10	2.46	Vibrionaceae

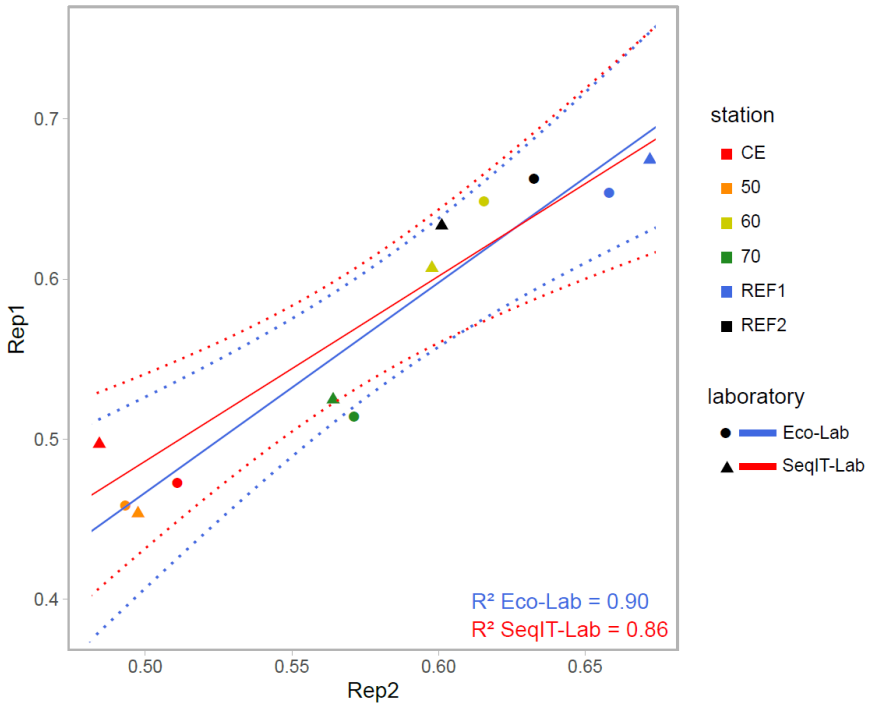
However, in our proof-of-concept study, all samples were sequenced to (near) saturation (Fig. 2). How many reads per sample is enough to obtain a reliable bacterial diversity measures and to infer environmental quality from these data still remains to be established. (Caporaso et al., 2010) recommended a deep as possible sequencing in case also rare (low abundant) taxa are important for specific ecological questions. For environmental monitoring, the importance of low-abundant taxa is presumably negligible. Therefore, to assess differences between samples (beta diversity), which mostly rely on the more abundant taxa, Lundin et al. (2012) showed that for bacterial communities ca. 1000 sequences may by sufficient for analyses. This number, however, is only a rough estimate as it may vary with the complexity of bacterial communities and also with the efficiency of data denoising tools.

Even though we found a higher variability in biological replicates compared to technical replicates (Fig. 6), predictions of IQI for individual samples of a biological replicate dataset based on a trained SML algorithm using all samples of the other biological replicate dataset were highly congruent (Fig. 10b). In our study, we used a relatively small dataset for a RF prediction of EQ. However, also other studies with predictions of EQ using a similar analysis (Cordier et al., 2017; Frühe et al., 2020). Variations in biological replicates from the same sampling site are expected for benthic microbial communities, but as a rule, these variations are not higher than the variation among different samples along an organic enrichment gradient of a salmon farm (Bissett et al., 2007; Dowle et al., 2015; Hornick and Buschmann, 2018; Stoeck et al., 2018a; 2018b; Verhoeven et al., 2018). This was also observed in our ordination analysis, in which biological and technical replicates of each individual sampling site form clusters of objects (shown as polygons in the plot) that are clearly separated along the first two gradients (axes) (Fig. 6). This allows for an unbiased distinction of each of the five objects (distance classes) in spite of the variation in biological replicates. Reasons for the observed variations in biological replicates as we found for both benthic macroinvertebrate (Fig. 4b) and bacterial communities (Fig. 6) are well known. The seafloor belongs to the most complex and patchy marine habitats. Sediments which appear to be of uniform texture are often composed of numerous microhabitats, especially in shallow coastal waters, notably varying for example in sediment particle topography, structure and diameter, sediment permeability and bulk water movement through sediment, oxygen and nutrient supply, redox potential, particulate matter deposition, and light availability on small spatial scales (Parsons et al., 1984).

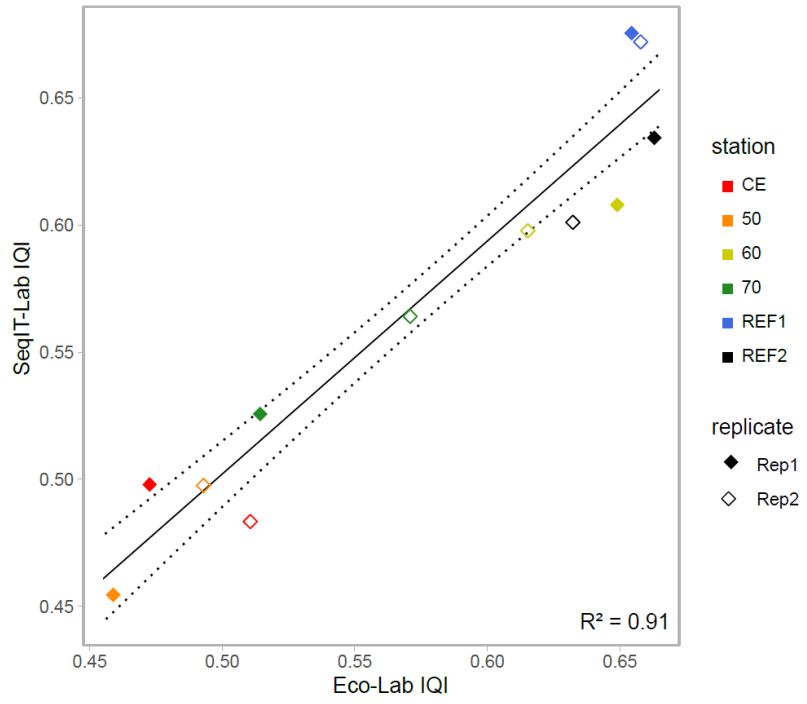
All these abiotic factors are highly selective for various taxa and may cause abrupt changes of benthic microbial community structures in spatial scales of millimeters to centimeters, and of benthic macroinvertebrates in scales of centimeters to meters (see examples reviewed in Parsons et al. 1984). Therefore, biological replicates are a necessity for benthic eDNA-based monitoring using bacterial marker genes. However, a possibility for practical "real-life" monitoring could be to unite biological replicates into a single sample for analyses to reduce possible noise due to sediment patchiness when interpreting microbial community shifts along enrichment gradients of salmon farms (see e.g. Keeley et al. (2018)).

Both technical replicate datasets are congruent regarding the qualitative and quantitative differences in bacterial community composition along the investigated distance gradient. The observed shifts in bacterial taxon groups agree largely with previously reported results (Bissett et al., 2007; Dowle et al., 2015; Hornick and Buschmann, 2018; Kawahara et al., 2009; Keeley et al., 2018; Stoeck et al., 2018a).

Therefore, we here discuss bacterial community shifts only briefly and refer to these previous studies for details. Common and well-described quantitative changes in bacterial taxon groups from cage edge to cage distant sites include for example an increase in chemoheterotrophic gammaproteobacteria and a decrease in anaerobic sulphate-reducing bacteria (SRB) (e.g. Dowle et al., 2015; Kawahara et al., 2008, 2009; Keeley et al., 2018; Stoeck et al., 2018a), as also observed in both technical replicate datasets of our study. Gammaproteobacteria is the most significant taxon group present in most marine sediments (Bowman and McCuaig, 2003; Inagaki et al., 2003; Li et al., 1999; Polymenakou et al., 2005), independent of pollution levels. SRBs are typically enriched in sediments immediately around marine fish farms, caused by the accumulation of organic matter (Bannister et al., 2014; Bissett et al., 2007; Carroll et al., 2003; Dowle et al., 2015; Holmer et al., 2005; Kawahara et al., 2008; Keeley et al., 2018; Kondo et al., 2008; 2012; Stoeck et al., 2018a; Verhoeven et al., 2018). In these sediments, they contribute to the anaerobic degradation of organic material via sulfate respiration. Recent evidence suggested that Desulfobacterales are also hydrogen scavengers (Dyksma et al., 2018). Therefore, it is not surprising that a recent study, identified Desulfobacterales as excellent environmental status indicators of organic enrichment and low redox potential in coastal habitats (Aylagas et al., 2017). Vibrionaceaea, Flavobacteriaceae, Clostridiaceae and Lachnospiraceae include representatives of the typical salmon gut microbiome (Fogarty et al., 2019). Thus, these taxa are released into the environment with the fish faeces, and deposited on the sea



a



b

Fig. 10. *Random Forest (RF) regression analysis to predict environmental quality (IQI as obtained from macroinvertebrate analyses, Fig. 3). (a) For each randomly chosen sample of one of the two technical replicate datasets using the RF algorithm trained on the complete dataset of the second technical replicate dataset. (b) for each randomly chosen sample of one of the two biological replicate dataset processed within a lab using the RF algorithm trained on the complete dataset of the second biological replicate dataset within the same lab (for both laboratories: Eco-Lab, SeqIT-Lab). 95% confidence intervals as well as regression statistics show that prediction accuracies are higher when the RF algorithm is trained on technical replicates compared to biological replicates. This demonstrates the high congruency among technical replicates of the same samples processed by two different laboratories.and their interpretation.*

floor. It is reasonable to assume that the relative contribution of these taxa to the sediment bacterial community decreases with increasing distance fromthe farm site. As pointed out earlier (Stoeck et al., 2018a), such a good agreement of benthic bacterial community structures at impacted coastal marine environments across large geographic scales (Europe and New Zealand) probably results from the high dispersal capabilities and global distribution of numerous bacterial taxa (Finlay, 2002; Yamaguchi et al., 2012). The observed changes in relative abundances of individual bacterial taxon groups along the distance gradient from cage edge to reference stations were not linear. This was congruent with macroinvertebrate community patterns obtained from traditional compliance monitoring of the salmon farm under study. Macroinvertebrate bioindicators evidenced a patchy organic enrichment across the three AZE stations. While IQI scores classed stations 50 m and 70 m as impacted, station 60 m achieved a similar score as both reference stations (negligible disturbance). Similar non-linear bacterial community shifts along organic enrichment gradients of salmon farms were reported previously and are common (Bissett et al., 2007; Dowle et al., 2015; Hornick and Buschmann, 2018; Kawahara et al., 2009; Keeley et al., 2018; Stoeck et al., 2018a). This is explained by the non-linear organic matter depositions from the fish farm, to which benthic microbes and also macroinvertebrates react. In a recent study, Armstrong et al. (2020) showed that the morphological structures of the seafloor as well as dynamic hydrological conditions may cause a mosaic-like patchy deposition of organic material from fish farm.

One further noteworthy finding, from the taxonomic analyses is the high proportion of unassigned bacterial ASVs. This corroborates well with previous reports of eDNA metabarcoded benthic marine bacterial communities in general (e.g. (Aravindraja et al., 2013) and fish farm associated benthic bacterial communities in specific (Cordier et al., 2018; Dowle et al., 2015; Frühe et al., 2020; Keeley et al., 2018; Stoeck et al., 2018a; Verhoeven et al., 2018). The major reason for our inability to assign more bacterial ASVs to identified taxa is due to shortcomings of available public nucleotide databases. Only a minute fraction of existing bacteria in marine sediments (and also other environments) are known and deposited in nucleotide databases (Cordier et al., 2020; 2019). Even though the ASVs that are assignable on higher taxonomic levels may be sufficient for coastal environmental monitoring (Aylagas et al., 2017; Borja, 2018), a very large fraction of informative data (unassignable ASVs) remain unused (Cordier et al., 2019). Therefore, current efforts are to develop taxonomy-independent data analyses approaches, which seem

very promising and our most efficient option at hand to use bacterial ASVs as bioindicators for EQ assessments (Cordier et al., 2020; 2019; Frühe et al., 2020).

5. Conclusions

In conclusion, the concordance of results obtained from the same samples in two different laboratories suggest that in the long term a formal eDNA metabarcoding based proficiency protocol could be developed for the implementation of this technology in compliance monitoring regulations. In the first step, this eDNA protocol could amend the macroinvertebrate-based monitoring scheme through e.g. replacing one macrofauna replicate by eDNA samples. This would reduce costs and allow further refinement of the eDNA technology through the exploitation of the simultaneously obtained macrofauna data as reference for environmental quality. These refinements include for example the progressive development of indicator ASV databases, with focus of bacterial ASVs as universal bioindicators that react exclusively to salmon farm impacts, independent of natural seasonal or geospatial effects (Frühe et al., 2020; Keeley et al., 2018). In the long term, eDNA based metabarcoding could then fully replace the more expensive, laborious, time consuming and tedious macroinvertebrate-based monitoring. Our study showed that the laboratory protocol for eDNA analyses is sufficiently robust and reproducible to meet the quality standards defined by accreditation or quality assurance bodies for the implementation of eDNA metabarcoding into routine environmental compliance monitoring in the aquaculture sector.

CRediT authorship contribution statement

Verena Dully: Investigation, Formal analysis, Data curation, Visualization. **Heinrich Balliet:** Investigation, Formal analysis, Data curation. **Larissa Frühe:** Formal analysis, Data curation. **Martin Däumer:** Investigation, Validation, Supervision. **Alexander Thielen:** Formal analysis, Data curation. **Sheena Gallie:** Conceptualization, Resources. **Iain Berrill:** Conceptualization, Resources. **Thorsten Stoeck:** Conceptualization, Methodology, Validation, Writing - original draft, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was funded by grant STO414/15-1 from the German Research Foundation (DFG) to TS. We appreciate the support of the environmental sampling team and the crew of the sampling vessel at Scottish Sea Farms for their assistance with the collection of the samples and provision of the faunal data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi. org/10.1016/j.ecolind.2020.107049[*].

[*] All supplementary files are additionally available at the appendix of this dissertation

References

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources, 17*(6), 1231-1242. doi:10.1111/1755-0998.12668

Aravindraja, C., Viszwapriya, D., & Karutha Pandian, S. (2013). Ultradeep 16S rRNA Sequencing Analysis of Geographically Similar but Diverse Unexplored Marine Samples Reveal Varied Bacterial Community Composition. *Plos One, 8*(10), e76724. doi:10.1371/journal.pone.0076724

Armstrong, E. G., Mersereau, J., Salvo, F., Hamoutene, D., & Dufour, S. C. (2020). Temporal change in the spatial distribution of visual organic enrichment indicators at aquaculture sites in Newfoundland, Canada. *Aquaculture International, 28*(2), 569-586. doi:10.1007/s10499-019-00478-z

Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., Irigoien, X., Danovaro, R., & Rodríguez-Ezpeleta, N. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Marine Pollution Bulletin, 114*(2), 679-688. doi:10.1016/j.marpolbul.2016.10.050

Bagley, M., Pilgrim, E., Knapp, M., Yoder, C., Santo Domingo, J., & Banerji, A. (2019). High-throughput environmental DNA analysis informs a biological assessment of an urban stream. *Ecological Indicators, 104*, 378-389. doi:10.1016/j.ecolind.2019.04.088

Bannister, R. J., Valdemarsen, T., Hansen, P. K., Holmer, M., & Ervik, A. (2014). Changes in benthic sediment conditions under an Atlantic salmon farm at a deep, well-flushed coastal site. *Aquaculture Environment Interactions, 5*(1), 29-47. doi:10.3354/aei00092

Berry, D., Mahfoudh, K. B., Wagner, M., & Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology, 77*(21), 7846-7849. doi:10.1128/AEM.05220-11

Bissett, A., Burke, C., Cook, P. L. M., & Bowman, J. P. (2007). Bacterial community shifts in organically perturbed sediments. *Environmental Microbiology, 9*(1), 46-60. doi:10.1111/j.1462-2920.2006.01110.x

Boers, S. A., Jansen, R., & Hays, J. P. (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *European Journal of Clinical Microbiology & Infectious Diseases, 38*(6), 1059-1070. doi:10.1007/s10096-019-03520-3

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods, 10*(1), 57. doi:10.1038/nmeth.2276

Borja, A. (2018). Testing the efficiency of a bacterial community-based index (microgAMBI) to assess distinct impact sources in six locations around the world. *Ecological Indicators, 85*, 594-602. doi:10.1016/j.ecolind.2017.11.018

Bowman, J., & McCuaig, R. (2003). Biodiversity, Community Structural Shifts, and Biogeography of Prokaryotes within Antarctic Continental Shelf Sediment. *Applied and Environmental Microbiology, 69*, 2463-2483. doi:10.1128/AEM.69.5.2463-2483.2003

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Brown, J. R., Gowen, R. J., & McLusky, D. S. (1987). The effect of salmon farming on the benthos of a Scottish sea loch. *Journal of Experimental Marine Biology and Ecology, 109*(1), 39-51. doi:10.1016/0022-0981(87)90184-5

Burridge, L., Weis, J. S., Cabello, F., Pizarro, J., & Bostick, K. (2010). Chemical use in salmon aquaculture: A review of current practices and possible environmental effects. *Aquaculture, 306*(1-4), 7-23. doi:10.1016/j.aquaculture.2010.05.020

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335-336. doi:10.1038/nmeth.f.303

Carroll, M. L., Cochrane, S., Fieler, R., Velvin, R., & White, P. (2003). Organic enrichment of sediments from salmon farming in norway: environmental factors, management practices, and monitoring techniques. *Aquaculture, 226*, 165-180. doi:10.1016/S0044-8486(03)00475-7

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology, 30*, 2937-2958. doi:10.1111/mec.15472

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01518

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources, 18*(6), 1381-1391. doi:10.1111/1755-0998.12926

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology, 27*(5), 387-397. doi:10.1016/j.tim.2018.10.012

Dowle, E., Pochon, X., Keeley, N., & Wood, S. A. (2015). Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *Fems Microbiology Ecology, 91*(8), fiv089. doi:10.1093/femsec/fiv089

Dyksma, S., Pjevac, P., Ovanesov, K., & Mussmann, M. (2018). Evidence for H2 consumption by uncultured Desulfobacterales in coastal sediments. *Environmental Microbiology, 20*(2), 450-461. doi:10.1111/1462-2920.13880

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilites. *Genome Research, 8*(3), 175-185. doi:10.1101/gr.8.3.175

FAO (2018). Food and Agriculture Organization of the United Nations. The State of World Fisheries and Aquaculture 2018. *Retreived on 05.01.2022 from https://www.fao.org/documents/card/en/c/I9540EN.*

Filippidou, S., Junier, T., Wunderlin, T., Lo, C.-C., Li, P.-E., Chain, P. S., & Junier, P. (2015). Under-detection of endospore-forming Firmicutes in metagenomic data. *Computational and Structural Biotechnology Journal, 13*, 299-306. doi:10.1016/j.csbj.2015.04.002

Finlay, B. J. (2002). Global dispersal of free-living microbial eukaryote species. *Science, 296*(5570), 1061-1063. doi:10.1126/science.1070710

Fogarty, C., Burgess, C. M., Cotter, P. D., Cabrera-Rubio, R., Whyte, P., Smyth, C., & Bolton, D. J. (2019). Diversity and composition of the gut microbiota of Atlantic salmon (Salmo salar) farmed in Irish waters. *Journal of Applied Microbiology, 127*(3), 648-657. doi:10.1111/jam.14291

Forrest, B. M., Keeley, N., Gillespie, P., Hopkins, G., Knight, B., & Govier, D. (2007). Review of the Ecological Effects of Marine Finfish Aquaculture: Final Report. The Ministry of Fisheries, New Zealand. *Retrieved on 18.01.2022 from https://www.yumpu.com/en/document/read/20615870/review-of-the-ecological-effects-of-marine-finfish-aquaculture-final-*.

Forster, D., Filker, S., Kochems, R., Breiner, H.-W., Cordier, T., Pawlowski, J., & Stoeck, T. (2019a). A Comparison of Different Ciliate Metabarcode Genes as Bioindicators for Environmental Impact Assessments of Salmon Aquaculture. *Journal of Eukaryotic Microbiology, 66*(2), 294-308. doi:10.1111/jeu.12670

Forster, D., Lentendu, G., Filker, S., Dubois, E., Wilding, T. A., & Stoeck, T. (2019b). Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environmental Microbiology, 21*(11), 4109-4124. doi:10.1111/1462-2920.14764

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

Gihring, T. M., Green, S. J., & Schadt, C. W. (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environmental Microbiology, 14*(2), 285-290. doi:10.1111/j.1462-2920.2011.02550.x

Golebiewski, M., & Tretyn, A. (2020). Generating amplicon reads for microbial community assessment with next-generation sequencing. *Journal of Applied Microbiology, 128*(2), 330-354. doi:10.1111/jam.14380

Gowen, R. J., & Bradbury, N. B. (1987). The Ecological Impact of Salmonid Farming in Coastal Waters - a Review. *Oceanography and Marine Biology, 25*, 563-575. doi:10.1016/0198-0254(88)92716-1

Herlemann, D. P. R., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., & Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal, 5*(10), 1571-1579. doi:10.1038/ismej.2011.41

Holmer, M., Wildish, D., & Hargrave, B. T. (2005). Organic Enrichment from Marine Finfish Aquaculture and Effects on Sediment Biogeochemical Processes. In B. T. Hargrave (Ed.), *Environmental Effects of marine Finfish Aquaculture* (pp. 182-206). New York: Springer.

Hornick, K. M., & Buschmann, A. H. (2018). Insights into the diversity and metabolic function of bacterial communities in sediments from Chilean salmon aquaculture sites. *Annals of Microbiology, 68*(2), 63-77. doi:10.1007/s13213-017-1317-8

Hvas, M., Folkedal, O., Solstorm, D., Vågseth, T., Fosse, J. O., Gansel, L. C., & Oppedal, F. (2017). Assessing swimming capacity and schooling behaviour in farmed Atlantic salmon Salmo salar with experimental push-cages. *Aquaculture, 473*, 423-429. doi:10.1016/j.aquaculture.2017.03.013

Illumina (2013). Metagenomic Sequencing Library Preparation. *Retrieved on 20.06.2020 from https://support.illumina.com/documents/documentation/chemistry_ documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.*

Inagaki, F., Suzuki, M., Takai, K., Oida, H., Sakamoto, T., Aoki, K., Nealson, K. H., & Horikoshi, K. (2003). Microbial Communities Associated with Geological Horizons in Coastal Subseafloor Sediments from the Sea of Okhotsk. *Applied and Environmental Microbiology, 69*(12), 7224-7235. doi:10.1128/AEM.69.12.7224-7235.2003

Kawahara, N., Shigematsu, K., Miura, S., Miyadai, T., & Kondo, R. (2008). Distribution of sulfate-reducing bacteria in fish farm sediments on the coast of southern Fukui Prefecture, Japan. *Plankton and Benthos Research, 3*(1), 42-45. doi:10.3800/pbr.3.42

Kawahara, N., Shigematsu, K., Miyadai, T., & Kondo, R. (2009). Comparison of bacterial communities in fish farm sediments along an organic enrichment gradient. *Aquaculture, 287*(1), 107-113. doi:10.1016/j.aquaculture.2008.10.003

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

Keeley, N. B., Forrest, B. M., Crawford, C., & Macleod, C. K. (2012). Exploiting salmon farm benthic enrichment gradients to evaluate the regional performance of biotic indices and environmental indicators. *Ecological Indicators, 23*, 453-466. doi:10.1016/j.ecolind.2012.04.028

Keeley, N. B., Forrest, B. M., & Macleod, C. K. (2013). Novel observations of benthic enrichment in contrasting flow regimes with implications for marine farm monitoring and management. *Marine Pollution Bulletin, 66*(1-2), 105-116. doi:10.1016/j.marpolbul.2012.10.024

Kennedy, K., Hall, M. W., Lynch, M. D., Moreno-Hagelsieb, G., & Neufeld, J. D. (2014). Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Applied and Environmental Microbiology, 80*(18), 5717-5722. doi:10.1128/aem.01451-14

Kondo, R., Shigematsu, K., & Butani, J. (2008). Rapid enumeration of sulphate-reducing bacteria from aquatic environments using real-time PCR. *Plankton and Benthos Research, 3*(3), 180-183. doi:10.3800/pbr.3.180

Kondo, R., Shigematsu, K., Kawahara, N., Okamura, T., Yoon, Y. H., Sakami, T., Yokoyama, H., & Koizumi, Y. (2012). Abundance of sulphate-reducing bacteria in fish farm sediments along the coast of Japan and South Korea. *Fisheries Science, 78*(1), 123-131. doi:10.1007/s12562-011-0439-3

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology, 25*(17), 4392-4406. doi:10.1111/mec.13761

Li, L., Kato, C., & Horikoshi, K. (1999). Microbial Diversity in Sediments Collected from the Deepest Cold-Seep Area, the Japan Trench. *Marine Biotechnology, 1*(4), 391-400. doi:10.1007/PL00011793

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R news, 2(3)*, 18-22.

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

Lundin, D., Severin, I., Logue, J. B., Östman, Ö., Andersson, A. F., & Lindström, E. S. (2012). Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity? *Environmental Microbiology Reports, 4*(3), 367-372. doi:10.1111/j.1758-2229.2012.00345.x

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal, 6*(3), 610-618. doi:10.1038/ismej.2011.139

Monroe, C., Grier, C., & Kemp, B. M. (2013). Evaluating the efficacy of various thermo-stable polymerases against co-extracted PCR inhibitors in ancient DNA samples. *Forensic Science International, 228*(1), 142-153. doi:10.1016/j.forsciint.2013.02.029

Nicholson, A., McIsaac, D., MacDonald, C., Gec, P., Mason, B. E., Rein, W., et al. (2020). An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines. *Environmental DNA, 2*(3), 343-349. doi:10.1002/edn3.81

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2019). R Package 'vegan': Community Ecology Package. *Version 2.5-6.*

Parsons, T. R., Takahashi, M., & Hargrave, B. (1984). *Biological oceanographic processes*: (3rd ed.). Oxford: Pergamon Press.

Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources, 14*(6), 1129-1140. doi:10.1111/1755-0998.12261

Phillips, G. R., Anwar, A., Brooks, L., Martina, L. J., Miles, A. C., & Prior, A. (2014). Infaunal quality index: Water Framework Directive classification scheme for marine benthic invertebrates. *Retrieved on 01.01.2022 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/314673/Water_Framework_Directive_classification_scheme_for_marine_benthic_invertebrates_-_report.pdf*

Polymenakou, P. N., Bertilsson, S., Tselepides, A., & Stephanou, E. G. (2005). Bacterial Community Composition in Different Sediments from the Eastern Mediterranean Sea: a Comparison of Four 16S Ribosomal DNA Clone Libraries. *Microbial Ecology, 50*(3), 447-462. doi:10.1007/s00248-005-0005-6

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., & Rimet, F. (2018). Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia, 807*(1), 37-51. doi:10.1007/s10750-017-3381-2

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio, 9*(3), e00525-00518. doi:10.1128/mBio.00525-18

SEPA (2018). Scottish Environmental Protection Agency. Fish farm survey report-evaluation of a new seabed monitoring approach to investigate the impacts of marine cage fish farms. *Retrived on 20.06.2020 from https://consultation.sepa.org.uk/sectorplan/finfishaquaculture/supporting_documents/Fish%20Farm%20Survey%20Report.pdf.*

Smyth, R. P., Schlub, T. E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., Davenport, M. P., & Mak, J. (2010). Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene, 469*(1), 45-51. doi:10.1016/j.gene.2010.08.009

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., & Pawlowski, J. (2018b). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecological Indicators, 85*, 153-164. doi:10.1016/j.ecolind.2017.10.041

Verhoeven, J. T. P., Salvo, F., Knight, R., Hamoutene, D., & Dufour, S. C. (2018). Temporal Bacterial Surveillance of Salmon Aquaculture Sites Indicates a Long Lasting Benthic Impact With Minimal Recovery. *Frontiers in Microbiology, 9*, e03054. doi:10.3389/fmicb.2018.03054

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). New York: Springer.

Yamaguchi, N., Ichijo, T., Sakotani, A., Baba, T., & Nasu, M. (2012). Global dispersion of bacterial cells on Asian dust. *Scientific Reports, 2*, 525. doi:10.1038/srep00525

CHAPTER III

Robustness and applicability of current approaches for EQ inference

Summary

In recent years, SML approaches as RF have been established as methods for inferring the ecological status of an ecosystem using eDNA metabarcoding-based datasets (Cordier et al., 2017; Cordier et al., 2018; Cordier et al., 2019b; Dully et al., 2021c; Frühe et al., 2020). SML is easy to use and can handle highly complex data with numerous variables as they arise in the bioinformatics process (Breiman, 2001; Fox et al., 2017; Freeman et al., 2015; Hastie et al., 2009a). It uses single features or combinations thereof, which makes it rather unintelligible for laymen, but results in high predictive power (Cordier, 2020; Cordier et al., 2018; Frühe et al., 2020). As an alternative to SML, a method based on QRS was recently established (Aylagas et al., 2021; Keeley et al., 2018; Lanzén et al., 2020). The QRS approach is based on the abundance distribution of taxa along an environmental gradient (Keeley et al., 2012). Taxa that occur at an EQ category at a certain rate are classified to a specific EG if specific quality characteristics are met. Based on the proportion of indicators assigned to EG of a novel sample, the EQ can be inferred based on an equation calculating a molecular index (Keeley et al., 2018).

To give a clear recommendation regarding the EQ inference method for the implementation in SOPs, the RF-based and the QRS-based approach were compared regarding coastal aquacultural sites. Such SOPs are the prerequisite for the implementation of the eDNA method into legislative regulations and are desired by regulators and scientists (Goldberg et al., 2016; Helbing and Hobbs, 2019; Kelly et al., 2014). This study demonstrated that the classification of the ecological status of an environment was consistent between the two methods and mirrored macrofauna-based evaluations. Therefore, the potential of both EQ-inference approaches for implementation in SOPs for marine coastal monitoring is emphasized. However, in the Scottish dataset used, certain environmental events are most likely not accurately represented due to the spatial and temporal dynamics of bacterial communities in marine sediment induced by biogeography patterns (Frühe et al., 2021), microhabitat patchiness (Lejzerowicz et al., 2014), seasonal variation (Delille, 1995) or aquafarm production phase (Dully et al., 2021c; Keeley et al., 2018).

This resulted in the underrepresentation of identified QRS-based EG indicators, which prevented the QRS approach from being applied to such datasets. For datasets mapping high degrees of environmental heterogeneity, SML is therefore recommended. However, for the construction of a universally applicable monitoring system in general, more samples are considered necessary to cover the whole spatiotemporal variation of benthic bacteria communities (Aylagas et al., 2021; Keeley et al., 2018; Lanzén et al., 2020). SML models are easier to augment than QRS-inferred EG databases, because manual intervention like filtering candidate indicators for special characteristics is obsolete. Additionally, SML harbors the potential of co-learning of environmental variables, which can help to disentangle background noise from the desired monitoring target (Cordier et al., 2020). Therefore, for the construction of a universal monitoring tool based on eDNA metabarcoding, SML for EQ inference is recommended.

Background

Routine compliance surveys are in place to monitor the ecological status of marine coastal environments affected by anthropogenic activities, such as organic enrichment originating from aquaculture installations. These monitoring programs are based on SOPs resulting in the EQ assessment of sites subjected to surveillance. Recently, a transition from using macrofaunal invertebrates towards the implementation of eDNA metabarcoding-based methods takes place. It has been demonstrated that specifically bacteria-based eDNA compositional analyses can mirror the traditionally inferred EQ (Aylagas et al., 2017; Stoeck et al., 2018a). As the eDNA approach for impact assessments can save time and money, it harbors a great potential for the implementation in the legislature (Kelly et al., 2014; Pawlowski et al., 2018). With that, calls for standardization of compliance monitoring protocols are being made by the scientific community, industrial stakeholders, and regulators (Goldberg et al., 2016; Helbing and Hobbs, 2019). For the successful implementation into SOPs, the applicability of the best practice has to be confirmed for different systems under investigation (Loeza-Quintana et al., 2020).

A major challenge of eDNA metabarcoding-based monitoring is the conversion of the bacterial community profiles into their respective EQ. Two different statistical approaches have successfully been developed for EQ inference in the current literature.

One approach uses SML algorithms (Armstrong and Verhoeven, 2020; Cordier et al., 2018; Dully et al., 2021a; Frühe et al., 2020) while others rely on the calculation of a biotic index, based on which an EQ category is then inferred (Aylagas et al., 2021; Keeley et al., 2018). As further detailed in a review by Cordier et al. (2019b), the SML strategy is to train a predictive model using a labeled dataset. This means that *a priori*, the desired label or response (e.g., EQ category) is known for each provided sample, in order to classify new samples with unknown labels. The training of such a model consists of identifying the features (e.g., ASVs), or combinations of them, that explain the response among a usually large number of features. This knowledge can then be used by the algorithm to make predictions on new samples with unknown labels. The SML approach is taxonomy-independent and therefore does not mind gaps in nucleic acid reference databases (Cordier et al., 2018). Furthermore, the approach is computationally fast and efficient despite large numbers of variables such as ASVs (Breiman, 2001). Association patterns within the full bacterial ASV dataset are automatically disentangled from the background noise allowing high predictive performance for complex datasets (Fox et al., 2017; Prasad et al., 2006). SML is easily upscalable and fully automatable, however, a high number of training data might be required for accurate predictions of new uncharacterized samples (Cordier et al., 2018; Dully et al., 2021c; Lanzén et al., 2021). Nevertheless, SML algorithms appear to be the most promising approach for establishing a new routine biomonitoring tool (Cordier et al., 2018; 2019b).

In contrast, other EQ inference methods have recently been proposed which rely on bioindicator inference and subsequent index calculation (Aylagas et al., 2021; Keeley et al., 2018). Such *de novo* approaches for the identification of bioindicators bypass limited availability of taxonomic information and therefore do not rely on any previous ecological knowledge about individual taxa or taxonomic units (Cordier et al., 2020). Therefore, they also allow detection of bioindicators that have not been investigated before (Chariton et al., 2015; Lejzerowicz et al., 2014; Pawlowski et al., 2018). Recently, several studies testing the applicability of *de novo* EQ inference approaches emerged (Frühe et al., 2020; Lanzén et al., 2021; Lanzén et al., 2020). In 2020, 152 sediment samples of Norwegian salmon farms were tested for their EQ inference performance regarding an AMBI classification system (Frühe et al., 2020). SML regression analysis was compared to a traditional Indicator Value (IndVal) approach following Dufrene and Legendre (1997), developed to identify indicators. The IndVal approach can be used for indicator inference, as it extracts indicators based on their distribution among distinct groups of samples.

The more conclusive the ASV abundance among sample group characteristics, the higher is the IndVal achieved. With those inferred values, a molecular index can be calculated subsequently. The main outcome of the study was that SML is superior to IndVal inference in terms of predictive power. Further improvement of the SML models is expected to be straightforward, as more samples can be easily added to the algorithm. Additionally, the authors suggested to account for natural community variation by including further environmental parameters for model construction (Frühe et al., 2020).

In 2018, Keeley et al. published a promising alternative *de novo* approach introducing QRS, which could be successfully applied for EQ inference of Spanish and New Zealand sediments (Anderson, 2008; Aylagas et al., 2021; Keeley et al., 2012). The principle of this approach is to compare abundance information of individual taxa or taxonomic units (e.g., ASVs) against an environmental parameter and statistically identify the abundance distribution along an environmental gradient while determining the peak of abundance along the gradient. Therefore, it is a well-established and frequently used approach for analyzing the ecological tolerances of little-known or undescribed organisms (Anderson, 2008; Keeley et al., 2018; Lanzén et al., 2020). Each taxon or ASV that meets specific bioindicator suitability requirements can then be allocated to a specific EG. Subsequently, using a biotic index, such as molecular AMBI (Aylagas et al., 2021; Keeley et al., 2018), the EQ category of a sample can be inferred based on the proportions of the assigned bioindicators in their respective EG. Therefore, the QRS follows a similar strategy as used for traditional monitoring approaches in coastal marine environments: the calculation of a biotic index based on identified bioindicators, with which the EQ category of a sample can be determined. Computationally, this approach is also upscalable and fully automatable after reliable bioindicators are found and confirmed. A possible disadvantage of an indicator-derived biotic index is the need to validate the response of indicators in different sites within an ecological gradient (e.g independent salmon farms) when the database is augmented (Keeley et al., 2018).

Because consensus of the scientific community regarding EQ inference methods which are based on eDNA metabarcoding data is still missing (Aylagas et al., 2021), Lanzén et al. (2020) conducted a study comparing the predictive power of the most promising approaches for EQ inference, namely SML and QRS. The authors analyzed 44 stations in the Cantabrian Sea which are influenced by various urban and industrial contaminations. In their study, both tested approaches performed similarly well for prokaryotic communities.

The authors reported very well-balanced predictions based on SML, resulting in a better linear correlation to reference values outperforming QRS in most cases. Regarding the classification accuracy in EQ categories, however, QRS-based predictions were at least equally accurate. Thus, the authors suggest that both approaches can be applied on samples deriving from the same habitat with a good predictive performance expected. For the inclusion of additional samples from novel geographical regions, the SML approach is suggested, as it is computationally less expensive and allows an easy re-training of the models (Lanzén et al., 2020).

However, when the spatial environment or the monitoring target is switched, the obtained results are no longer applicable. In a follow-up study focusing on the impact of oil and gas spills from deep sea wells, benthic eukaryotes and metazoans were investigated. The authors demonstrated that for the dataset under investigation, QRS could provide usable EQ assessments while the SML analysis was unsuccessful (Lanzén et al., 2021). The authors assume that this was due to an insufficient number of samples and therefore recommend the usage of QRS over SML for small datasets.

It must be assumed that neither the existent QRS-inferred indicators nor the SML models can be directly applied to different habitats under different environmental influences such as aquaculture-induced eutrophication (Aylagas et al., 2021; Keeley et al., 2018). As suggested by Lanzén et al. (2020), the methods need to be extended and validated to make statements about other habitats prior to the implementation in SOPs. Towards the development of an SOP for compliance monitoring of aquaculture effects on marine coastal environments, clear recommendations need to be given. Therefore, for the standardization of EQ inference for marine coastal environments that are under the influence of aquaculture, the approaches need to be validated. As such sediments are expected to be mainly characterized by the accumulation of organic carbon, a special ecological composition of the microbial communities is expected, mirrored by different indicators of the environmental status (Bissett et al., 2007; Frühe et al., 2021; Jansson et al., 2006; Kawahara et al., 2009).

Therefore, both recent approaches for EQ inference, SML and QRS, are compared to each other regarding their predictive performance based on the benthic bacterial community associated with seven Norwegian and seven Scottish salmon aquaculture installations (230 samples in total). The results of both approaches were compared to each other and to the results of a traditional compliance monitoring using benthic macroinvertebrate surveys as a benchmark.

Methods

In this study, metabarcoding bacterial data of a total of 230 sediment samples derived from salmon aquaculture installations were used for EQ inference. More specifically, 138 samples were analyzed from seven salmon farms in Norway and 92 samples from seven farms in Scotland. For the seven Norwegian farms, metabarcoding data was available at the sequence read archive (SRA) under the accession number PRJNA562304 (Frühe et al., 2020). Sequence data of two Scottish salmon farms (LIS, MAC) has also been published (Dully et al., 2021b; Frühe et al., 2021) and can be accessed under SRA accession numbers PRJNA666305 and PRJNA768445, respectively. The sequence data for the other five Scottish salmon farms (FIS, KIN, MAN, NOS, SCA) were newly produced during compliance monitoring following the workflow described in detail in Frühe et al. (2020). In short, sediment samples were collected Van-Veen grabs (0.045m² area). Samples were taken below the salmon farms at various distances from the salmon cage edge to represent the pollution gradient. Afterwards, eDNA was extracted using the DNeasy PowerSoil kit (Qiagen, Hildesheim, Germany). The hypervariable 16S V3-V4 region was amplified using the primers Bakt_341F and Bakt_805R (Herlemann et al., 2011). After Illumina sequencing, *cutadapt* v.2.11.0 (Martin, 2011) was used for primer truncation. Subsequently, the sequences were embedded into the DADA2 workflow (Callahan et al., 2016), which aggregates the sequences into ASVs. All accessed and obtained ASVs were merged into one single ASV-to-sample matrix. For each dataset (Norway and Scotland), the ASV-to-samples matrices were converted to proportional tables, showing the relative abundance of ASVs within a sample by dividing the number of reads per ASV by the total number of reads and then multiplying by 100 as in similar studies (Dully et al., 2021c; Lanzén et al., 2020).

EQ inference was based on the 250 most abundant ASVs, therefore excluding rare taxa to reduce uninformative noise and to reduce computational time similar to previous studies (Cordier et al., 2018; Dully et al., 2021a; Keeley et al., 2018; Lanzén et al., 2020). ASVs were taxonomically assigned using vsearch's *syntax* function (Rognes et al., 2016) based on the *greengenes* database (McDonald et al., 2012). Based on the resulting ASV-to-sample matrix, a molecular environmental index was either calculated via QRS or predicted using an SML regression model. The results of the two methods were compared to a macrofauna-inferred reference index to evaluate their potential of making accurate EQ predictions.

As an index of environmental quality, IQI was used. This index is used for compliance monitoring in the UK (Phillips et al., 2014; SEPA, 2018). It calculates the ratio between the observed value of the metric and the value in a reference unimpacted site, being an important EQ ratio tool for coastal and transitional water bodies (Kennedy et al., 2011). IQI values range from 0 to 1, where measures close to 1 indicate high status (e.g. at unimpacted reference sites) and values close to 0 indicate a low status of ecological quality (e.g. at the salmon cage edge), while the 'good/moderate' boundary is set to ≥0.64 (Phillips et al., 2014). We used this established threshold value for the classification into 'adequate" or 'inadequate' samples. Due to privacy guidelines, the full names of the Scottish farms are not available. All supplementary material can be found in the appendix of this thesis.

Quantile Regression Splines analysis

Bacterial indicators were identified across an EQ gradient following the analytical approaches of Keeley et al. (2018). First, QRS models for the 95th percentile were constructed to examine the abundance response of the ASVs to the EQ gradient as estimated by the IQI. For each ASV, the read abundance values in each farm (response variable) were plotted against IQI values (predictor variable). For this analysis, the R packages *quantreg* and *splines* were used (Bates and Venables, 2011; Koenker et al., 2018). After QRS, the peak of the abundance was predicted for each ASV across the EQ gradient. The abundance peaks of the ASVs were detected using the function *find_peaks* from *pracma* R package (Brown, 2012) and were further evaluated using two criteria: i) the number of farms with a peak and ii) the agreement of peak values among farms to eventually select the ones which show a consistent response to the EQ gradient as in similar studies (Aylagas et al., 2021; Keeley et al., 2018; Lanzén et al., 2020). More specifically, ASVs with one peak in at least five farms were selected. Then, for each ASV, the mean IQI peak value and the standard deviation were calculated across all farms of a dataset. Finally, ASVs with good quality scores, i.e., corresponding to standard deviation less than 0.2, were selected as bioindicators and were used for the calculation of the molecular biotic index, i.e., the molecular IQI. The selected indicator ASVs were assigned to an EG from EG I, corresponding to sensitive taxa, to EG V, corresponding to opportunistic taxa according to their mean IQI peak value following the WFD practitioners guidelines (*Table 1*, UKTAG, 2012).

Table 1: Classification of IQI values into EGs used for QRS-based indicator inference. *The table shows which EG was assigned for each IQI range and their corresponding classification into ecological status. The status reflects the degree of disturbance of external influences, such as organic enrichment. The color-coding scheme is set according to the UKTAG guidelines.*

IQI	Eco-Group	Status	Disturbance	
1-0.75	I	High	Nor / minor	
0.74-0.64	II	Good	Slight	
0.63-0.46	III	Moderate	Moderate	
0.45-0.25	IV	Poor	Major	
0.24-0	V	Bad	Severe	

After indicator inference, the molecular IQI was calculated. Therefore, based on the EG proportions within a sample, the AMBI, another molecular index for EQ inference, was firstly calculated using the traditional formula according to Borja et al. (2000, *Eq.1*). Additionally, diversity metrics were calculated using the R package *vegan* (Oksanen et al., 2020). Then, the molecular IQI was calculated by using the IQI version IV by Phillips et al. (2014) and applying modifications for the molecular data as follows (*Eq. 2*).

$$IQI_{v.IV} = \left(\left(0.38 \times \left(\frac{1-(AMBI/7)}{1-(AMBI_{\text{Ref}}/7)} \right) \right) + \left(0.08 \times \left(\frac{1-\lambda'}{1-\lambda'_{\text{Ref}}} \right) \right) + \left(0.54 \times \left(\frac{S}{S_{\text{Ref}}} \right)^{0.1} \right) - 0.4 \right) \Big/ 0.6 \quad \textit{(Eq.2)}$$

Where:

$IQI_{v.IV}$ represents the newly calculated molecular IQI

AMBI corresponds to the molecular AMBI

1-λ´ is Simpson´s evenness

S is the log_{10} number of ASVs per sample

Ref corresponds to reference values of unimpacted sites.

Supervised machine learning/ Random Forest algorithm

SML was conducted using the RF algorithm which emerged as a powerful tool for eDNA metabarcoding-based data (Cordier et al., 2017; 2018; Dully et al., 2021c; Frühe et al., 2020). The data basis for the RF regression analysis was formed by the proportional ASV-to-sample matrices containing the 250 most abundant ASVs and the macrofauna-IQI values representing the reference labels. All samples were subjected to a LOO-CV procedure (James et al., 2013). This means that in each run, one single observation, in our case one sample, is omitted to build a regression model with all remaining samples.

The previously omitted observation is then used to validate the regression model built on all remaining observations resulting in 138 independent models for the Norwegian dataset and 92 independent models for the Scottish dataset. The predicted continuous IQI is subsequently transformed into discrete categories, namely 'adequate' (IQI≥0.64) or 'inadequate' (IQI<0.64), indicating adequate or inadequate EQ. The categories are then compared to the benchmark macrofauna IQI categories. The LOO-CV approach was repeated 10 times for each sample individually. RF was executed using the R package *randomForest* for classification and regression (Liaw and Wiener, 2002) with default parameters to keep the influence of the operator as small as possible. Variable importance representing the participation of each ASV for a correct prediction was inferred from all constructed RF regression models and then averaged.

Statistical comparisons

The accuracy of each EQ inference method was calculated by counting the number of correct predictions in the respective category. The agreement between inferred categories between reference and predicted values was tested with Cohen's kappa statistics κ by using the *kappa2* function of R package *irr* (Gamer et al., 2012). Kappa values from 1.0-0.81 indicate an almost perfect agreement of predicted to reference standards while values from 0.8-0.61 indicate substantial agreement (Landis and Koch, 1977).

To further examine the relationship between the morphologically inferred IQI and molecular IQI, a regression analysis was performed (function *lm*). Regarding linear regression, the goodness of model fitting can be evaluated by different measures. If the predicted values correspond exactly to the reference values, the slope x of the linear model corresponds to $x = 1$. Additionally, R^2 values indicate the proportion of variance explained by the predictor variables, resulting in a maximal $R^2 = 1$ for a perfectly fitted model (Lepš and Šmilauer, 2020). Additionally, the root mean squared error (RMSE) was calculated for each model as a measure of agreement between actual and predicted values. The lower the RMSE, the closer are predicted values are to the reference data (Hastie et al., 2009b).

Results

Data Overview

The raw data of the seven Norwegian salmon farms sediments consisted of 21,877,920 raw reads in total.

They were bioinformatically filtered for HQ sequences, resulting in 3,541,124 HQ reads, corresponding to 83,662 ASVs. On average, Norwegian samples contained 1588 ASVs. For Scotland, seven farms in total were investigated, from which we obtained 24,306,286 raw reads. After DADA2 processing, we were able to retain 5,229,185 HQ reads, corresponding to 186,539 ASVs. Average ASV richness per sample amounted to 4215 ASVs for the Scottish dataset. For the novel samples that were taken at the farms FIS, KIN, MAN, NOS, and SCA, a detailed sequence overview per farm and the respective rarefaction curves are provided in *Supplementary File 3.1* and *Supplementary File 3.2*. For Norway, sample classification into EGs revealed 65 'adequate' samples representing a high or good EQ (IQI ≥0.64) and 73 'inadequate' samples representing a moderate, poor, or bad EQ (IQI<0.64). For Scotland, there were 40 'adequate' samples and 52 'inadequate' samples (*Figure 13*).

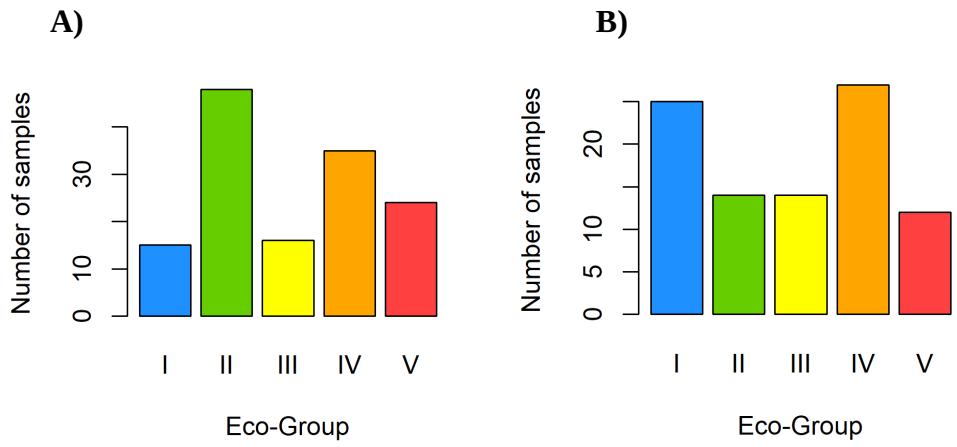


Figure 13) Sample EG distribution inferred by IQI values. *EG distribution of the investigated samples across A) Norway (n = 138 samples) and B) Scotland (n = 92 samples). The bars indicate the number of samples assigned for respective EGs according to their macrofauna-inferred IQI values. EGs I and II represent IQI ≥ 0.64 resulting in a classification into 'adequate' environmental status (blue and green color), while EGs III-V indicate an 'inadequate' environmental status (yellow, orange, and red color).*

Bacterial indicators

Inference of bacterial indicators inferred by QRS was performed on the 250 most abundant ASVs per dataset. EGs were assigned only to ASVs representing good quality splines, corresponding to a peak in at least five farms and a standard deviation of peak values less than 0.2. The QRS analysis resulted in 148 indicator ASVs for Norway, meaning 59.2% of the 250 ASVs showed a consistent response to the EQ gradient (*Supplementary File 3.3*).

For Scotland, the analysis resulted in 79 indicators, corresponding to 31.6% of the 250 ASVs used for analysis (*Supplementary File 3.4)*. The majority of the indicator ASVs for both Norway (43%, n = 63) and Scotland (76%, n = 60) belonged to EG IV (*Figure 15*). One EG V indicator ASV was found for each dataset. For Norway, 35% (n = 52) of indicators were assigned to EG II, and 17% (n = 25) were assigned to EG III. For Scotland, EG III indicators make up for 14% (n = 11) of the ASVs. EGs I and II contain 4% (n = 3) and 5% (n = 4) of indicators ASVs. Exemplary QRS plots showing the abundance response of selected indicator ASVs (one from each EG) across the ecological gradient are provided in *Supplementary File 3.5* for the Norwegian dataset and *Supplementary File 3.6* for the Scottish dataset. Comparing the inferred indicators among datasets, 25 indicators, corresponding to 12.4%, were shared between Norway and Scotland. 72% of the shared ASVs indicated poor ecological status (EG IV) in both countries. For 16% of the shared indicator ASVs, EG assignment was not congruent among Norway and Scotland (*Supplementary File 3.7*).

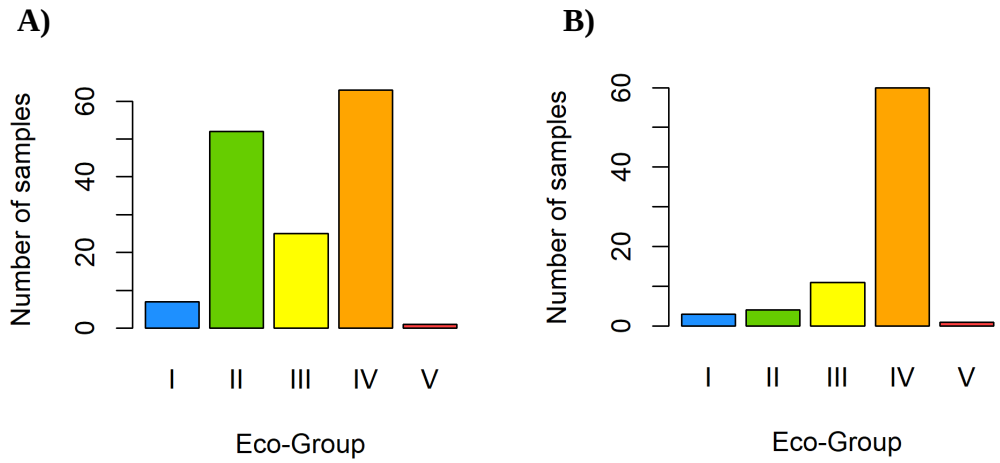
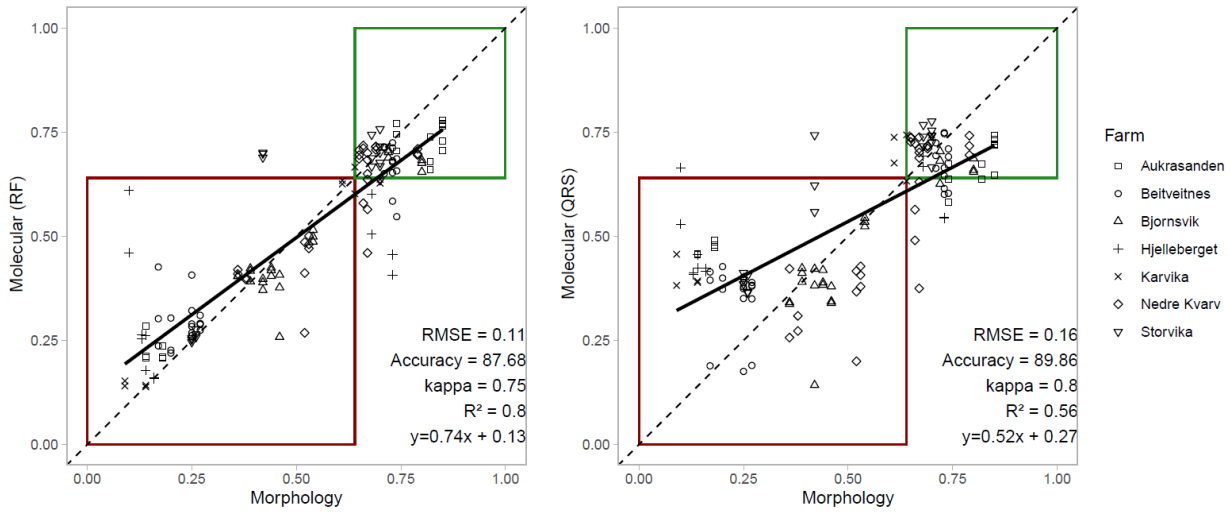


Figure 14) Indicator EG distribution inferred by QRS. *EG distribution of QRS indicator ASVs across samples in A) Norway (n = 148 indicators) and B) Scotland (n = 79 indicators). The bars indicate the number of indicators ASVs retrieved by QRS analyses among their assigned EGs. EGs I and II represent indicators for 'adequate' sample status (blue and green color) while an 'inadequate' environmental status is indicated by EGs III-V (yellow, orange, and red color).*

Molecular IQI as estimated by SML and QRS

Linear relationships between the macrofauna-based IQI and the molecular IQI as estimated by QRS and SML were significant ($p < 0.001$) for Norway and Scotland datasets. However, SML gave a higher correlation coefficient for both datasets compared to QRS, with a 0.24 difference among R^2 values for Norway and 0.05 for Scotland *(Figure 15)*.

A)



B)

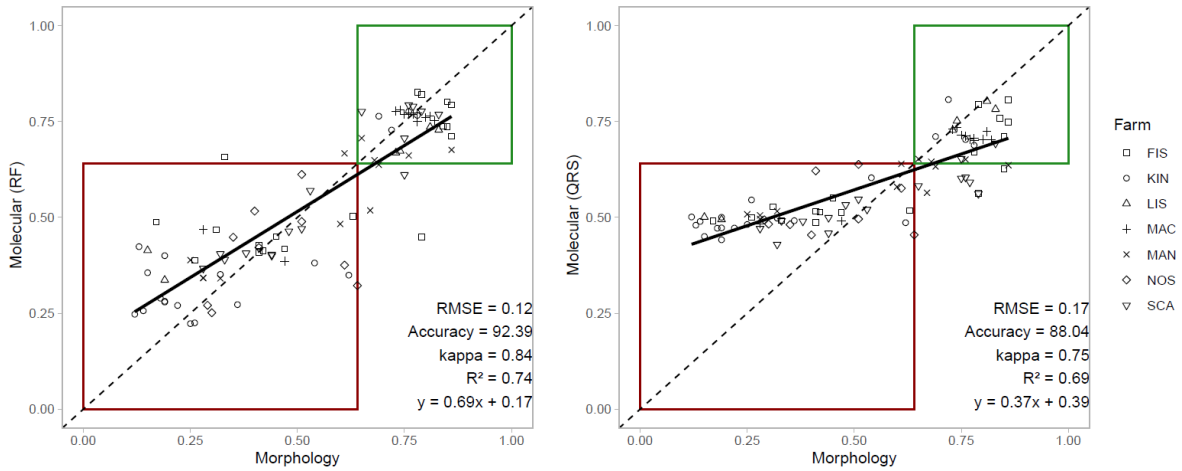


Figure 15) Linear regression plots comparing IQI values. *Linear regression plots showing the relationship between the morphologically inferred IQI and the molecular IQI as estimated by RF and QRS for A) Norway and B) Scotland salmon farms. The boxes indicate the two ecological quality categories that IQI assigns the samples (i.e. red for 'inadequate' (IQI<0.64) samples and green for 'adequate' (IQI≥0.64) samples). The corresponding R^2 values, slopes, and RMSE values are given for each regression plot.*

The classification into categories was successful for the Norwegian and the Scottish dataset using SML and QRS, as both approaches achieved κ values greater than 0.61 indicating substantial to almost perfect agreement (Landis and Koch, 1977). The agreement is almost perfect (κ > 0.8) for QRS predictions in Norway and SML predictions in Scotland. For Norway (n = 138 samples), the regression analysis of the traditionally inferred IQI compared to the molecular IQI resulted in R^2 = 0.56 for QRS and R^2 = 0.8 for SML. For the QRS analysis, the slope of the linear model was 0.52 while for SML analysis, the slope was 0.74. The RMSE of the QRS model was 0.16 while the RMSE of the SML model was 0.11.

All three measures evaluating the goodness of model fit indicate that SML produced superior regression IQI predictions than QRS. Regarding the accuracy of IQI categorization (‘adequate’ vs. ‘inadequate’), QRS and SML were coherent. The number of samples categorized correctly based on their IQI prediction was similar among SML (n = 121) and QRS (n = 124). SML prediction showed an accuracy of 88% (κ = 0.75) indicating a ‘good agreement’ between predicted and reference categories. The QRS model achieved an accuracy of 90% with κ = 0.8 indicating an ‘almost perfect agreement’. The 12% (n = 17) wrongly predicted samples using the SML were separated into 2% (n = 3) IQI overestimations and 10% (n = 14) underestimations. For QRS, 10% (n = 14) of the samples were misclassified, corresponding to 3% (n = 4) overestimations and 7% (n = 10) underestimations.

For Scotland (n = 91 samples), the regression analysis of the SML model also revealed a good correspondence of predicted IQI values compared to the reference values. The R^2 of the linear model was 0.74, RMSE accounted for 0.12 and the slope of the linear model was 0.69. For QRS-based IQI inference, R^2 was 0.69, RMSE was 0.17 and the slope of the linear model was 0.37. Regarding IQI classification (‘adequate’ vs. ‘inadequate’), QRS showed 88% accuracy (κ = 0.75) while SML showed an accuracy of 92% (κ = 0.84). This corresponded to 81 correctly classified samples based on QRS, and 84 correctly classified samples based on SML. The 8% (n = 7) wrongly predicted samples using the SML model were separated into 2% (n = 2) IQI overestimations and 6% (n = 5) underestimations. For QRS, 12% (n = 11) of the samples were misclassified, which corresponded to 2% (n = 2) overestimations and 10% (n = 9) underestimations.

In total, there were 26 incorrect predictions for Norway. Five of them, which corresponded to approximately 20%, were predicted erroneously by SML and QRS simultaneously. The SML approach misclassified 12 further samples, while QRS misclassified nine other samples. In total, 70% of the misclassified Norway samples were underestimations. This means that the molecular IQI inferred from SML or QRS was lower than the reference IQI-based on morphology to the degree of allowing for wrong conclusions of the category (*Figure 16A).* Regarding the Scottish dataset, the majority of wrongly classified samples (10 of 13) was also caused by IQI underestimations. Five samples (5.4%) were erroneously predicted by SML and QRS simultaneously. While QRS overestimated the IQI of six further samples, SML overestimated two samples exclusively. Notably, some wrong predictions were induced by a strong disagreement of the predicted to the reference IQI for QRS and SML *(Figure 16B)*.

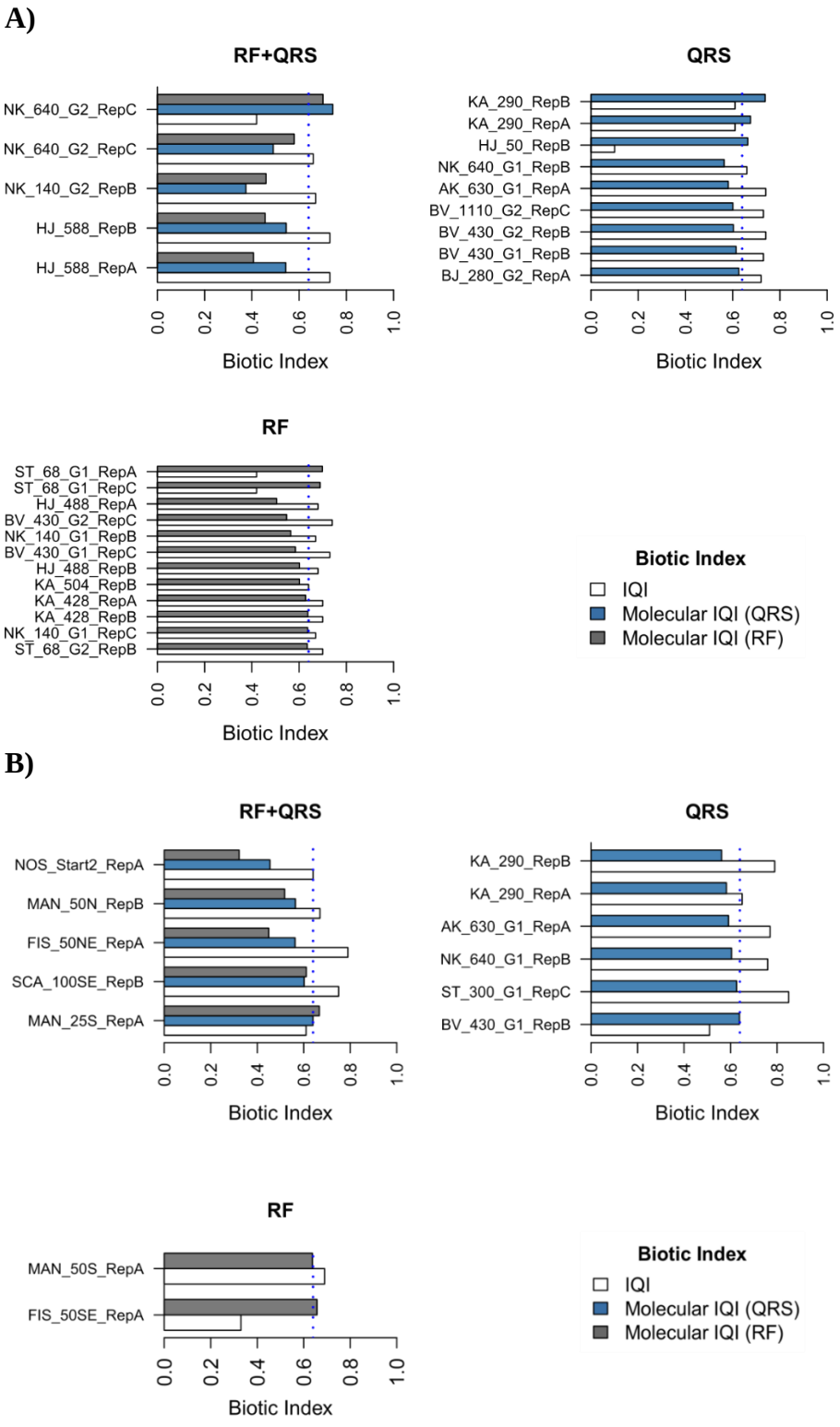


Figure 16) Erroneously classified samples and their respective biotic indices. *Samples were misclassified based on predicted IQI values using QRS and RF for A) Norway and B) Scotland salmon farms. The blue bars indicate QRS-inferred IQI values while the grey bars indicate RF-interred IQI values. The white bars indicate the reference macrofauna IQI. The blue, dashed line indicates the IQI threshold of 0.64 at which classification in 'adequate' or 'inadequate' samples is conducted. In the first section, misclassified samples using both approaches are indicated, following erroneously classified samples using either QRS or RF.*

Taxonomic assignment of RF and QRS bioindicators

Indicator ASVs inferred by QRS were compared to the 20 highest rated SML indicator ASVs according to their variable importance per dataset (*Figure 17*). A table of the ASVs, their taxonomic assignment, and the corresponding variable importance can be found in *Supplementary File 3.8*. Out of the 20 ASVs with the highest SML variable importance for Norway, 15 belonged to the indicators selected after the QRS analysis. The same amount of indicator ASVs was shared for the analysis of the Scottish dataset. In total, 40% of the ASVs identified as the top indicators for SML belonged to the bacterial family Helicobacteracea, the genus *Psychrilyobacter* (family: Fusobacteriaceae), and the genus *Lutimonas* (family Flavobacteriaceae). Another set of predictive ASVs accounting for 12.5% of the top SML indicators was considered as Bateriodales which could not be assigned to a deeper taxonomic level. Regarding QRS-based indicator inference, the family Helicobacteracea was represented by 7% of the Norwegian indicators and 20% of the Scottish indicators. Helicobacteracea were categorized as EG IV indicators (opportunistic taxa) in over 96% of the cases. Nine different indicator ASVs belonging to the genus *Lutimonas* were also identified by QRS analysis.

A)

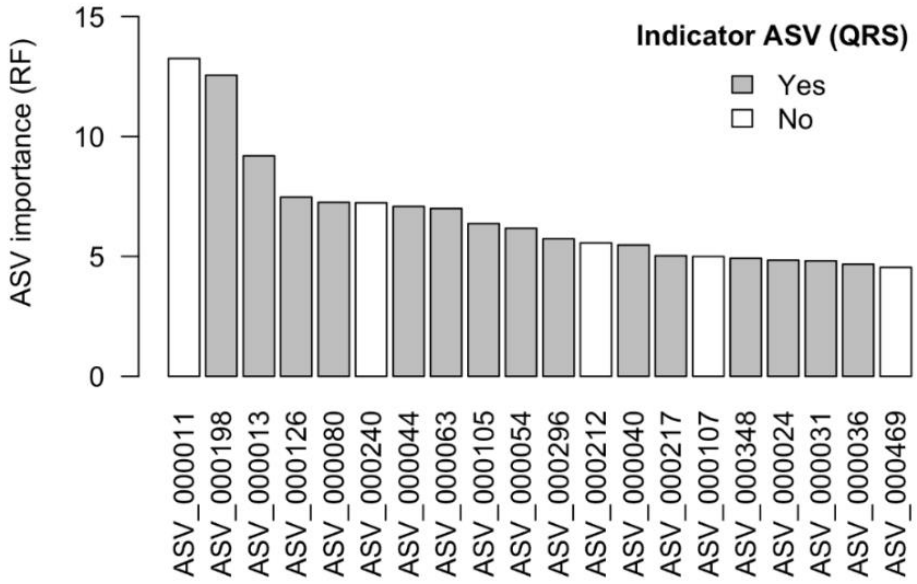


Figure 17) RF top 20 indicator ASVs per dataset. *The top 20 ASVs assigned the highest variable importance by RF for A) Norway and B) Scotland ASVs which were assigned as indicators by QRS analysis are indicated in grey color.*

B)

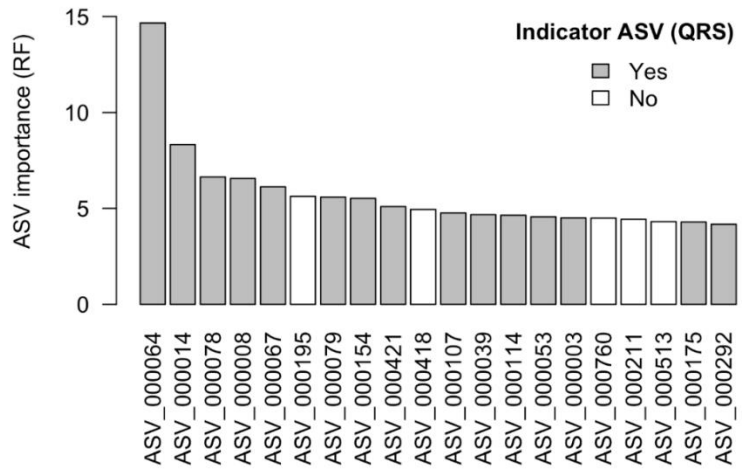


Figure 17) *continued*

Two *Lutimonas* ASVs (ASV_000053 and ASV_000064) were shared between Norway and Scotland. Those ASVs were also considered as indicator ASV by RF. Other common QRS-inferred indicators for Norway were Piscirickettsiaceae (EG II-III), Myxococcales (EG I-III), and Desulfobulbaceae (EG II-IV), Bacteroidales (EG III-V), and Alteromonadales (EG II-IV). For Scotland, Alteromonadales and Bacterioidales were also identified as indicators, indicating EG IV. Six ASVs belonging to the family Desulfobulbaceae were identified for Scotland, with three of them belonging to EG I and three of them belonging to EG IV.

Discussion

For both countries under investigation, RMSE was lowest for SML-inferred EQ prediction compared to QRS-based calculations. A minimal RMSE is desired, as a low error measure indicates high agreement between predicted and reference values. This observation was confirmed by higher R^2 values for SML, which represent a high proportion of predicted values explained by the reference. For the IQI inferred by SML, x was closer to $x = 1$ compared to the QRS prediction for both datasets. Combining those statistical measures, we could demonstrate that SML showed a superior predictive model regarding IQI regression prediction compared to the QRS-based bioindicator approach.

Subsequently, the inferred IQI values were used to categorize samples into ‘adequate’ (IQI≥0.64) or ‘inadequate’ (IQI<0.64) categories, which enabled the QRS approach to match the SML-inferred accuracies. For Scotland, classification based on SML outperformed classifications based on QRS, while for Norway, QRS-based classification accuracy was slightly increased compared to SML. These findings can be confirmed by a previous study conducted by Lanzén et al. (2020), in which the authors compared SML and QRS predictive models regarding various environmental exposures introduced by urbanization at the Basque coast. Corresponding to our findings, the authors reported fair to good agreements with reference ecological status groups for both tested approaches. They also reported a superior linear correlation of SML-inferred index values compared to QRS-based values. However, when using an alternative dataset focusing on another monitoring objective in a follow-up study, they contradicted these findings (Lanzén et al., 2021). Taking all of this together, it is indicated that for different monitoring locations and objectives, different EQ inference methods might yield the best possible result.

Misclassifications

For the Norwegian dataset under investigation, several samples were misclassified triggered by IQI prediction discrepancies closely to the adequate’ (IQI≥0.64) or ‘inadequate’ (IQI<0.64) threshold, accounting for seven and two misclassifications for SML and QRS, respectively. For example, if the IQI prediction resulted in an IQI of 0.63 instead of 0.67 for the sample NK_140_G1_RepC via SML, the sample was therefore misclassified as ‘inadequate’. A correct classification would have been made if the IQI value was predicted 0.01 higher, resulting in an IQI of 0.64, indicating ‘adequate’ status. Regarding SML, approx. 40% of the misclassified samples differed less than 0.1 from the reference IQI values. For QRS, approx. 20% of all misclassifications are induced by near-threshold misclassifications with less than 0.1 difference in IQI.

In the Scottish dataset, seven samples were misclassified by SML, from which two samples are near-threshold samples (MAN_25S_RepA and MAN_50S_RepA). For those samples, a discrepancy of the predicted IQI of 0.05 and 0.06 led to misclassification. For QRS, 11 samples were misclassified in total. Again, the sample MAN_25S_RepA was overestimated. An increased IQI of 0.03 lead to a classification into ‘adequate’ status, whereas the reference sample was categorized as ‘inadequate’ with a reference IQI of 0.61. For the sample SCA_57_RepB, an underestimation of 0.07 categorized the sample into ‘inadequate’ status, although the sample should have been classified as ‘adequate’.

As narrow over- and underestimation of samples took place, samples that were close to the threshold tended to be easily misclassified. Thus, a difficulty in classification of samples that show ecological statuses narrowly around the threshold became apparent. A similar phenomenon has been described by Lanzén et al. (2020), as SML regression analyses showed a better agreement to the reference than the actual classifications. Other studies have analyzed predictive power by classification into a higher number of categories (Aylagas et al., 2017; Cordier et al., 2019b; Frühe et al., 2020; Lanzén et al., 2021), which precluded an analysis of the samples close to the threshold. Therefore, estimations of the predictive performance using model regression is often preferred over evaluation based on sample classification (Keeley et al., 2018; Lanzén et al., 2021). To improve predictive performance for SML and QRS, more samples slightly above and slightly below the set threshold are required to increase group separability. This would lead to both better SML prediction performance around the threshold and an increase in bioindicators that can be found via the QRS approach, enabling more reliable IQI inferences (Dully et al., 2021c). To facilitate these improvements, targeted selection of near-boundary samples is recommended during compliance monitoring.

SML classification

Another potential solution to overcome this bias, though exclusively for SML approaches, consists of applicating an SML classification approach rather than an SML regression approach. In this case, it is advantageous that both alternatives are already implemented in the RF algorithm. For classification, the RF algorithm is no longer trained to predict the continuous molecular index but is trained for predictions into the two discrete categories ‘adequate’ or ‘inadequate’. The output consists of sample classifications rather than continuous values which subsequently need to be classified by a threshold. Interpretation of such classification approaches is straightforward, as they are reduced in dimensionality. As a result, only classification accuracy and kappa values can be inferred, while regression models to compare the prediction values to the reference must be waived. The RF classification approach was able to outperform the conducted RF regression and QRS analysis for the Norwegian dataset, but not for the Scottish dataset (*Supplementary Files 3.9 and 3.10*). An explanation might be that the classification approach better handled near-threshold samples in some cases, which are frequent in the Norwegian dataset but scarce in the Scottish dataset.

However, the SML classification approach in general does not add any value for comparative studies since the reduction of the dimensions also prevents comparability among gradients and approaches. Regarding SOP development for monitoring approaches, SML should be conducted as a regression analysis since the use of the SML classification would prevent comparability with QRS or other methods. QRS is not able to perform classification tasks directly and is only intended for the calculation of continuous values as molecular indexes, which are then subsequently categorized. As the SML classification output does not include continuous values, the model must be constructed again when the threshold or the monitoring goal is changed. Thus, with the existing predictions, it would not be possible to separate the samples into further categories without retraining. Therefore, SML approaches solely depending on classification are not recommended for monitoring SOPs, which require to be dynamically adaptable to new regulations including forward and backward compatibility with current indices (Cordier et al., 2020).

QRS-induced under- and overestimations in the Scottish dataset

For Scotland, QRS showed an underestimation bias for high-status samples for a variety of farms. This was likely caused by the fact that 76% of all Scottish indicator ASVs established by QRS belonged to EG IV, while indicators for 'adequate' samples were hardly found. Additionally, 'inadequate' status samples were heavily overestimated, which induced a pronounced upwards shift of the linear trend line (*Figure 15B, right panel*). The inferred IQI values ranged from 0.42 and up, whereas the original macrofauna-inferred IQI started at 0.12. For the sample with the lowest macrofauna-based IQI assigned, KIN_60T1_RepB with a reference IQI of 0.12, QRS-based IQI inference predicted a molecular IQI of 0.5. This overestimation pattern was likely introduced by a technical bias deriving from the IQI calculation formula. Scottish 'inadequate' samples were overestimated as they showed a high average number of ASV per sample which was a factor for IQI calculation (*Eq.2*). For the Norwegian dataset, this bias was not detected, likely because the average number of ASVs was approx. three times smaller compared to Scottish samples. For Scotland, the underestimation bias regarding 'adequate' samples, combined with the overestimation bias for 'inadequate' samples, led to a dislocated trend line accounting for a slope of $x = 0.37$. Therefore, the conducted QRS approach was not able to accurately predict IQI values lower than 0.5 based on the investigated dataset.

Nevertheless, these biases had little to no effect on the classification accuracies into ‘adequate’ and ‘inadequate’ samples in this study, but they prevent the model from being used for higher resolution monitoring using additional categories as usual (Aylagas et al., 2017; Cordier et al., 2019b; Frühe et al., 2020; Lanzén et al., 2021).

Spatiotemporal heterogeneity

For the Scottish dataset, 79 ASVs were identified as indicators using the QRS method. This accounted for approx. 30% of the 250 ASVs used for the indicator inference. This number is low compared to the Norwegian dataset (60%), but also compared to other studies that were able to extract up to 89% percent as indicators using the same method (Keeley et al., 2018). This is an indication that it was more difficult to find indicators based on the data used, which resulted in a decreased predictive performance. This is probably due to the fact that in Scotland, a higher variability of environmental factors is expected, which triggers specific responses of the microbial community (Parmar et al., 2015; Yakimov et al., 2007). Mainly, the increased influence of various environmental factors on Scottish samples compared to Norwegian samples is expected to be due to the depth of the seafloor. While the sediments below Norwegian salmon farms were taken from deep fjords with an average depth of approx. 175m, all investigated Scottish sediments are located in shallow waters with an average depth of approx. 30m. It is well known that the influence of environmental parameters on the sediment of shallow waters is much greater than on deeper areas of the seafloor, which potentially applies to the Scottish samples (Nybakken and Bertness, 2004; Zhang et al., 1999).

The bacterial community found in such regions is shaped by a complex interplay of these factors and results in a highly dynamic, heterogeneous composition (Prosser et al., 2007). One of the main differences between shallow and deep-water habitats is salinity (Nybakken and Bertness, 2004). For example, it has been demonstrated that in the water bodies connecting the North Sea and the Baltic Sea, salinity decreases from approx. 35 PSU (Practical Salinity Units) at 200m depth to approx. 30 PSU at shallow surface waters because of freshwater influences from the Baltic Sea (Christensen et al., 2018). This region is a good example of a vertical salinity gradient, consisting of brackish to saline water at the surface, layered on highly dense seawater with a constant salinity (Stigebrandt, 1983). Salinity in shallow coastal waters, especially near estuaries or other freshwater bodies, shows a large variance due to freshwater inflows which harbors drastic influences on the microbial communities (Findlay and Watling, 1998; Vincent et al., 2021; Yang et al., 2016).

Additionally, these freshwater influxes may not only alter salinity but can carry numerous other chemical substances such as nutrients or toxins, which also have an impact on the microbial community (Biddanda et al., 2001; Gao et al., 2016; Hewson and Fuhrman, 2006; Jansson et al., 2006). Furthermore, both light and temperature change with seasonality, leading to enormous changes in the bacterial community over the year (Findlay and Watling, 1998; Fuhrman et al., 2008). The influence of seasons also affects deeper marine soils but is pronounced for shallow sediments. Most of the sediment samples from the Scottish farms were located in the euphotic zone, which means that these areas are under the influence of light. Depending on the availability of light, the composition of the surrounding organism community changes (Nybakken and Bertness, 2004). For example, if there is a lot of light, the growth of autotrophic microbial organisms and phototrophic plants is promoted. The presence of these organisms results in low grazing pressure and high availability of oxygen, which has a significant direct or indirect effect on the bacterial community (Findlay and Watling, 1997). At low light intensity, we expect the promotion of mainly heterotrophic organisms. These organisms affect the bacterial community through grazing (Findlay and Watling, 1998). High temperatures in summer months combined with high light availability and possible nutrient input from the land through wash-off can result in algal blooms, which in turn can once again affect the bacterial community (Hirche, 1987; Tande, 1988).

Other environmental factors that shape the community in a spatiotemporal context include geological, hydrological, and physicochemical factors (Vincent et al., 2021; Zhang et al., 1999). For example, it has been shown that the bacterial community is determined by sediment granulometry, which in turn is shaped by geologic and hydrologic conditions. It has been shown that different microbial communities are found at different sediment structures as muddy or coarse sediments (Zheng et al., 2014). However, the existing granulometry is constantly being altered in some places due to movement caused by ocean currents. If there is a strong current, fine particles are washed away, while coarse particles persist, triggering a shift in microbial abundance (Rusch et al., 2003). This impact on the bacterial community is induced as nutrient content and oxygen availability are altered (Cromey et al., 2002; Findlay and Watling, 1997). Such sediment movements can be observed especially in shallow waters, introducing a further level of heterogeneity for such habitats (Vincent et al., 2021).

Furthermore, it has also been shown that other physicochemical properties such as pH or redox potential can influence the microbial community in both coastal sediments and freshwater (DeAngelis et al., 2010; Liu et al., 2015; Tait et al., 2014).

Indicator distribution among Eco-Groups

It is expected that the seasonal influence of environmental parameters is thought to be greatest where the influence of organic enrichment is least. Therefore, in sediment heavily influenced by salmon farm organic enrichment, other external influences are overshadowed, leading to a more stable microbial community at near-cage samples (Frühe, 2021; Frühe et al., 2021). This potentially explains why most of the indicators found for Scotland indicated poor ecological status (EG IV), as Scottish sediment emerged to be highly heterogeneous in terms of environmental factors at sites less disturbed by organic enrichment (EG I or EG II). For the tested dataset, an exclusively technical bias for the overrepresentation of EG IV indicators induced by sample distribution among EQs can be excluded, as EG IV samples were not overrepresented according to their macrofaunal reference.

This was not the case for indicators based on the Norwegian dataset, as the distribution of samples was reflected in the numbers of indicators found per EG. For Norway, most samples were assigned to EGs II and IV, which was mirrored by the majority of found indicator ASVs. The results for the Norwegian dataset corroborate with a study conducted for New Zealand salmon farms, in which the authors retrieved mostly EG II and IV indicators (Keeley et al., 2018). As EGs I, III, and V were also represented to a similar extent, a balanced number of indicators per EG was achieved. Using those indicators for a prediction of the EQ of the samples by calculating an environmental index, the authors were able to reach an almost perfect agreement for their validation dataset ($R^2 = 0.92$). The even distribution of indicators and therefore the goodness of fit is likely caused by the homogeneity of the three salmon farms under investigation. Despite sampling over three consecutive years, all samples were taken in the month of November, representing the same seasonality of environmental influences. Additionally, all samples had been taken at analogous depths (average of 32.7m depth). Therefore, a high degree of similarity of the microbial community within all tested samples in their respective organic enrichment category is expected which explains the elevated predictive performance and the high number of indicators compared to our study.

Interestingly, the authors demonstrated a great influence of the existing flow regimes for macrofaunal, eukaryotic and foraminiferal data, but at the same time showed a high consistency of the bacterial community disregarding differences in current velocities (Aylagas et al., 2017; Keeley et al., 2018). It can be assumed that increased wave action has a great potential to bias samplings, as smaller particles are washed away and the benthic community and the organic enrichment impact changes with changing sediment structure (Keeley et al., 2013; Lin et al., 2021; Wickramasinghe et al., 2021). Therefore, a potentially false evaluation of the environmental status becomes apparent, as community turnover is triggered by wave action rather than the factor of interest, such as organic enrichment. Therefore, the suitability of bacterial barcodes for monitoring approaches is emphasized, as they can reliably represent the organic enrichment gradient also under increased wave action influence, regardless of the resulting sediment structure, and therefore outperform traditional morphology-based monitoring (Laroche et al., 2017).

Another assumption drawn by previous studies is therefore confirmed, questioning the suitability of traditional macrofauna-based ecological assignments. Lanzén et al. (2020) demonstrated missing correspondence of macrofauna-based ecological evaluation and actual physicochemical parameters. Therefore, in the long term, traditional indices such as AMBI, IQI, or NSI should be replaced by eDNA data, accompanied with physicochemical parameters describing the environmental status. It has been suggested to use a physicochemical pressure index, which includes measures such as total hydrocarbon or copper content (Aylagas et al., 2017; Lanzén et al., 2021). Nevertheless, it is invaluable to document the texture of the soil accompanying sampling, as it has already been demonstrated that coarse sediment can also lead to impediments in eDNA sampling (Pawlowski et al., 2022).

It was also observed that EG I and EG V indicators were severely underrepresented, although the original sample distribution was relatively balanced. This pattern was demonstrated for the Norwegian and the Scottish dataset simultaneously. Those ‘edge classes’ were located at the lowest and highest ends of the IQI scale. This phenomenon seems to be a technical bias that is introduced by prediction as it has often been observed regardless of the method (Cordier et al., 2019b; Frühe et al., 2020; Lanzén et al., 2021). However, for sample classification according to an ‘adequate’ vs. ‘inadequate’ system without interpretation value of edge classes, the influence of this technical relict is negligible for this study.

In summary, indicators for good ecological status seemed to be hard to find for shallow waters under salmon farm impact, as multiple other environmental parameters apart from organic enrichment shape the microbial community. To enable accurate EQ assessments based on QRS, indicators representing all EGs are required, but our results did not cover the whole extent of EGs evenly. To overcome this bias, more samples from Scotland are required, covering different geographical and hydrological conditions as well as different seasons. Environmental parameters which have the potential to shape the bacterial community should be measured while the eDNA samples are taken. Those environmental factors can later help to disentangle the natural variability introduced by geospatial influences rather than organic enrichment derived by the aquaculture installation. This can be done by including those factors as additional features to the SML algorithm, which is expected to increase predictive power.

Shared indicators and taxonomic assignment

Among the Norwegian and the Scottish dataset, 25 QRS indicators were shared, corresponding to 12.4% of the 202 indicators in total. Of those shared indicators, the majority (n = 18) belonged to EG IV, which is not surprising, as most Scottish indicators (76%) were assigned to EG IV. Four indicators have been assigned to non-corresponding EGs among countries, however, the discrepancy only accounted for one EG difference (*Supplementary File 3.10*). As the goal of the study was to compare the EQ inference methods QRS and SML, rather than comparing the indicators among the countries, it will not be discussed further. This is true especially since the comparison was inconclusive based on the available data, as for Scotland, almost exclusively indicators for EG IV have been found. This indicates insufficient sampling for QRS-based indicator inference. Nevertheless, the finding of shared indicators among both countries, even including a under-sampled dataset, corresponds to prior studies recognizing globally distributed indicators (Frühe et al., 2021). Such detection of shared indicators on a large spatiotemporal scale is a prerequisite for a universally applicable monitoring system in the future.

Comparing SML and QRS indicators, it is noticeable that some ASVs were used for SML models, while they were unsuitable for QRS-based EQ inference. It is essential to note that the most important SML indicator (ASV_000011) identified by the RF variable importance measure, was not identified as an indicator by QRS.

For the Norwegian salmon farms Bjørnsvik, Beitveitness Nedre Kvarv, and Aukrasanden, ASV_000011 indicated good environmental health (QRS abundance peaks at 0.65-0.77), but contrary its abundance was highest at polluted sites of Hjelleberget and Storvika farm (QRS abundance peak = 0.1-0.35). Since the abundance peaks of the ASVs were located at different EGs at different locations, this ASV was not identified as an indicator by QRS (*Supplementary File 3.11*). For SML model construction, abundance information of individual ASVs can be processed, as well as combinations from different ASVs. Therefore, ASV_000011 was still identified to be conclusive regarding IQI prediction by SML and was therefore used in the SML algorithm.

EG IV indicator ASVs of the family Helicobacteracea were extracted by SML and QRS simultaneously. This is not surprising as Helicobacteracea have been revealed as indicators of organic enrichment frequently (Aylagas et al., 2017; Frühe et al., 2021; Menchaca et al., 2014). Indicator ASVs belonging to the family of Fusobacteriaceae were identified by SML, while Fusobacteriaceae ASVs were not considered as indicators by QRS. Fusobacteriaceae have been reported in undisturbed sediments (Frühe et al., 2021) as well as in areas highly disturbed by aquacultural impacts by representatives of fish gut microbiota (Dehler et al., 2017; Verhoeven et al., 2018; Yukgehnaish et al., 2020), which could trigger nonconforming QRS peaks, therefore excluding them as indicators. ASVs belonging to the genus *Lutimonas* of the family Flavobacteriaceae, have been reported as indicators by QRS and RF simultaneously. As Falvobacteriaceaea have also been reported in the salmon gut microbiome (Fogarty et al., 2019), it is expected that they indicate a bad ecological status (EG IV). However, some studies also reported Flavobacteriaceae as indicators of a good environmental condition (Frühe et al., 2021). Some ASVs referred to as indicators by SML and QRS belong to Bacteriodales, which have been reported to be most abundant at sediments with a bad environmental status (Dully et al., 2021a). This corroborates with the QRS indicator assessment which indicated EGs III, IV, and V, thus preventing predictions inferring good status, although only a shallow taxonomic assignment to order level was possible. Piscirickettsiaceae ASVs have also been found in various habitats independent of environmental pollution gradients (Frühe et al., 2021; Gribben et al., 2017; Wang et al., 2018). This corroborates well with the reported ASVs indicating EGs II-III which account for a transitional zone. In our study, Myxococcales indicated a good environmental status as they were considered indicators for EGs I-III.

The increased abundance at undisturbed sediments was also reported by Keeley et al. (2018), while in a different study, conducted at another independent salmon farm, Myxococcales indicated a poor to moderate EQ (Dully et al., 2021a). In this study, Alteromonadales ASVs were also reported with especially high abundances at undisturbed sites. This corroborates partly with our findings that Alteromonadales ASVs were indicators for EGs II-IV in Norway and indicators for EG IV in Scotland.

Another group of indicator ASVs inferred simultaneously by SML and QRS were assigned to the family of Desulfobulbaceae. In the literature, Desulfobulbaceae, as a member of sulfate-reducing bacteria (SRB), were often reported at a bad environmental status (Aylagas et al., 2017; Finster et al., 1998; Keeley et al., 2018). But also, they have been reported as ubiquitous along a whole enrichment gradient (Frühe et al., 2021). In our study, three Desulfobulbaceae ASVs were categorized as EG II indicators, where the three Desulfobulbaceae ASVs were identified as indicators for EG IV. For those ASVs, the deepest level of taxonomic assignment possible resulted in Desulfobulbaceae family assignment. However, different ASVs can derive from different Desulfobulbaceae genera or even different Desulfobulbaceae species, which may respond differently to environmental impacts. Therefore, a well-known issue using taxonomic assignments rather than high-resolution sequence variants became apparent (Apothéloz-Perret-Gentil et al., 2021; Cordier, 2020; Cordier et al., 2018; Laroche et al., 2017; Serrana et al., 2022). Therefore, the recommendation of using ASVs rather than taxa is emphasized, which, however, makes comparability between already existing studies more challenging (Lanzén et al., 2020).

Database augmentation

QRS-based EQ inference uses ASV abundance data to identify potential indicators and to construct a sequence-to-EG database. To calculate a molecular environmental index, ASVs of a novel sample are compared to the indicators ASVs of the formerly constructed database. All the above-mentioned studies have shown that the QRS approach is applicable in the respective habitat. However, for the correct assessment of new biogeographical regions at different times, further indicator inference analyses must be performed (Aylagas et al., 2021; Keeley et al., 2018; Lanzén et al., 2020). Both methods, SML and QRS, are expected to improve IQI prediction with the addition of the number of samples (Lanzén et al., 2021).

Thus, to include this procedure into routine monitoring, the sequence-to-EG database must be increasingly augmented with a special focus on underrepresented EG indicators. As soon as indicator ASVs of new samples can be assigned to an EG, they have to be added to the database. At the same time, a query must be made whether this ASV already exists in the database. If this is the case, a new QRS analysis must be performed for this ASV including the new data. Subsequently, it must be decided whether this ASV can still be approved for classification by redefining quality score criteria. If the abundance peaks of an ASV differ between different locations, it is likely that this ASV is no longer suitable as a ubiquitous bioindicator and must be removed from the database (Cordier et al., 2020). Since different bacterial taxa, and thus different ASVs, show high spatial diversity (Aravindraja et al., 2013; Bowman and McCuaig, 2003; Frühe et al., 2021), enough samples to cover the whole span of community dynamics must be portrayed in the databases before a global implementation in standard protocols can be approached. Since the ASV community also changes temporally, samples at different time points are required. There are hints that different time points of the fish production phase succession should be included (Dully et al., 2021c; Keeley et al., 2018).

Also, it has been demonstrated repeatedly that the seasonal dynamics of the bacterial communities require high numbers of samples to describe the whole diversity (Delille, 1993; 1995; Gilbert et al., 2009; Lindh et al., 2015). Therefore, database-deposited indicators should be inferred using samples, covering as much spatial and temporal diversity as possible. For the augmentation of the RF model, on the other hand, the addition of new samples to the RF models is straightforward. Additionally, natural variability can be easily included in the algorithms to improve predictive performance (Cordier et al., 2017; 2020). Novel samples and their associated environmental evaluation can be incorporated by re-learning of the model, eliminating all manual queries that have to be conducted using QRS. Since the initial RF model was constructed without model tuning and independent of quality score parameters, the same EQ evaluations can be achieved by different users. Parameter tuning as the setting of the number of variables tried at each split (*mtry*) or the number of trees build (*ntree*) can improve SML models but was refrained from to keep the influence of the operator as low as possible. Thereby, a shortcut that allows anyone to include new samples in the models without expert knowledge is identified. The influence of the operator to define any parameters is thus obsolete, which is not the case for QRS.

In summary, dataset augmentation via SML is considerably easier than indicator expansion by QRS-based methods. Thus, for building up a huge database, which is needed for correct predictions in a wide spatiotemporal context, SML is recommended.

Conclusions and outlook

In this study, it was demonstrated that both recently used methods for EQ-inference based on eDNA metabarcoding data, SML and QRS, are suitable for the implementation in SOPs for marine coastal monitoring. Despite inferior regression models of QRS compared to SML, sample classification resulted in a high agreement to macrofauna-inferred reference data as reported before (Lanzén et al., 2020). For different monitoring locations and objectives, different EQ-inference methods yielded the best result. This study demonstrates that for spatiotemporal diverse data, SML outperformed QRS. This is because SML is especially suitable for such heterogeneous datasets as it is able to disentangle the monitoring target from the background noise. Therefore, the implementation of SML-based ecological assessments is recommended for SOPs.

To increase predictive performance, targeting samples under diverse geographical, hydrological, and temporal environmental influences, as well as samples around a given threshold of the respective monitoring scheme, is recommended (Cordier et al., 2020; Keeley et al., 2018; Lanzén et al., 2020). Additional measuring of those environmental factors can later help to disentangle the natural variability introduced by geospatial influences from the targeted monitoring goal, such as the organic enrichment introduced by aquaculture installations. Environmental parameters such as temperature, sampling depth, organic enrichment, or pH, can be easily implemented in SML-based approaches, automatically improving predictive power (Cordier et al., 2020). In agreement with previous studies, we recommend eDNA sampling combined with traditional macrofauna sampling and measurements of environmental parameters (Cordier et al., 2020; Lanzén et al., 2020; 2021). With a combination thereof, a robust database can be established, which, after successful completion and testing, can be used for reliable SML-based monitoring in various geographic regions. In the long term, this will be able to replace traditional monitoring, to which biases are increasingly being discovered.

There are further reasons why SML should be preferred over QRS for the development of SOPs for marine coastal monitoring. Firstly, SML is dynamically adaptable to existing and to new regulations.

As the desired output variable is directly used as a feature for model construction, there is no need for index calculations or further transformations of the variable. This results in forward and backward compatibility with existing monitoring frameworks, while a QRS-based system is always limited to the specific variable inputted with a subsequent transformation via calculation of a biotic index (Cordier et al., 2020).

Furthermore, the conducted QRS approach was not able to accurately predict IQI values lower than 0.5 when a dataset is not adequately covered, which can be triggered by the heterogeneity of environmental factors, and therefore a high turnover of the bacterial community. The lack of identifiable indicators prevents the model from being used for higher resolution monitoring using additional categories, which is possible with the constructed SML model. Novel samples and their associated environmental evaluation can be easily incorporated by SML relearning of the model, eliminating all manual queries that have to be conducted when using QRS-based indicator inference and subsequent index calculation. Therefore, for building up a huge database, which is the prerequisite for robust monitoring in a wide spatiotemporal context, the use of SML is recommended.

The next step towards a universally applicable monitoring system is consolidation of already collected data around the globe. Regrettably, inter-study comparability is expected to be low, as the conducted studies showed no consent over used primers and sample processing, which is why SOPs have to be developed (Cordier et al., 2020). In this study, the suitability of high-resolution bacterial ASVs for monitoring approaches is once again emphasized, studies based on taxonomic assignments are thus not recommended. To gain inter-study comparability, attention must be paid in particular to standardized sampling, extraction and PCR primers, as ASVs can only be compared with each other if these characteristics are matched. When samples from a variety of spatiotemporal impacts are sampled under the premise of standardized sample processing, a universally applicable SML tool based on eDNA metabarcoding is expected to be promptly implementable into regulatory monitoring frameworks (Pawlowski et al., 2021).

References

Anderson, M. (2008). Animal-sediment relationships re-visited: Characterising species' distributions along an environmental gradient using canonical analysis and quantile regression splines. *Journal of Experimental Marine Biology and Ecology, 366*, 16-27. doi:10.1016/j.jembe.2008.07.006

Apothéloz-Perret-Gentil, L., Bouchez, A., Cordier, T., Cordonier, A., Guéguen, J., Rimet, F., Vasselon, V., & Pawlowski, J. (2021). Monitoring the ecological status of rivers with diatom eDNA metabarcoding: A comparison of taxonomic markers and analytical approaches for the inference of a molecular diatom index. *Molecular Ecology, 30*(13), 2959-2968. doi:10.1111/mec.15646

Aravindraja, C., Viszwapriya, D., & Karutha Pandian, S. (2013). Ultradeep 16S rRNA Sequencing Analysis of Geographically Similar but Diverse Unexplored Marine Samples Reveal Varied Bacterial Community Composition. *Plos One, 8*(10), e76724. doi:10.1371/journal.pone.0076724

Armstrong, E., & Verhoeven, J. (2020). Machine-learning analyses of bacterial oligonucleotide frequencies to assess the benthic impact of aquaculture. *Aquaculture Environment Interactions, 12*, 131–137. doi:10.3354/aei00353

Aylagas, E., Atalah, J., Sánchez-Jerez, P., Pearman, J. K., Casado, N., Asensi, J., Toledo-Guedes, K., & Carvalho, S. (2021). A step towards the validation of bacteria biotic indices using DNA metabarcoding for benthic monitoring. *Molecular Ecology Resources, 21*(6), 1889-1903. doi:10.1111/1755-0998.13395

Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., Irigoien, X., Danovaro, R., & Rodríguez-Ezpeleta, N. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Marine Pollution Bulletin, 114*(2), 679-688. doi:10.1016/j.marpolbul.2016.10.050

Bates, M. D., & Venables, B. (2011). R Package 'splines': Regression Spline Functions and Classes. *Version 2.0*.

Biddanda, B., Ogdahl, M., & Cotner, J. (2001). Dominance of bacterial metabolism in oligotrophic relative to eutrophic waters. *Limnology and Oceanography, 46*(3), 730-739. doi:10.4319/lo.2001.46.3.0730

Bissett, A., Burke, C., Cook, P. L. M., & Bowman, J. P. (2007). Bacterial community shifts in organically perturbed sediments. *Environmental Microbiology, 9*(1), 46-60. doi:10.1111/j.1462-2920.2006.01110.x

Borja, A., Franco, J., & Pérez, V. (2000). A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments. *Marine Pollution Bulletin, 40*(12), 1100-1114. doi:10.1016/S0025-326X(00)00061-8

Bowman, J., & McCuaig, R. (2003). Biodiversity, Community Structural Shifts, and Biogeography of Prokaryotes within Antarctic Continental Shelf Sediment. *Applied and Environmental Microbiology, 69*, 2463-2483. doi:10.1128/AEM.69.5.2463-2483.2003

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Brown, C. (2012). R Package 'pragma': Provides a pragma / directive / keyword syntax for R. *Version 0.1.3*.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Chariton, A. A., Stephenson, S., Morgan, M. J., Steven, A. D. L., Colloff, M. J., Court, L. N., & Hardy, C. M. (2015). Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental Pollution, 203*, 165-174. doi:10.1016/j.envpol.2015.03.047

Christensen, K. H., Sperrevik, A. K., & Broström, G. (2018). On the variability in the onset of the Norwegian Coastal Current. *Journal of Physical Oceanography, 48*(3), 723-738. doi:10.1175/JPO-D-17-0117.1

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-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology, 30*, 2937-2958. doi:10.1111/mec.15472

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01518

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources, 18*(6), 1381-1391. doi:10.1111/1755-0998.12926

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019b). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology, 27*(5), 387-397. doi:10.1016/j.tim.2018.10.012

Cromey, C., Nickell, T., & Black, K. (2002). DEPOMOD – modeling the deposition and biological effects of waste solids from marine cage farms. *Aquaculture, 214*, 211-239. doi:10.1016/S0044-8486(02)00368-X

DeAngelis, K. M., Silver, W. L., Thompson, A. W., & Firestone, M. K. (2010). Microbial communities acclimate to recurring changes in soil redox potential status. *Environmental Microbiology, 12*(12), 3137-3149. doi:10.1111/j.1462-2920.2010.02286.x

Dehler, C. E., Secombes, C. J., & Martin, S. A. M. (2017). Environmental and physiological factors shape the gut microbiota of Atlantic salmon parr (*Salmo salar L.*). *Aquaculture, 467*, 149-157. doi:10.1016/j.aquaculture.2016.07.017

Delille, D. (1993). Seasonal changes in the abundance and composition of marine heterotrophic bacterial communities in an Antarctic coastal area. *Polar Biology, 13*(7), 463-470. doi:10.1007/BF00233137

Delille, D. (1995). Seasonal changes of subantarctic benthic bacterial communities. *Hydrobiologia, 310*(1), 47-57. doi:10.1007/BF00008182

Dufrene, M., & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs, 67*(3), 345-366. doi:10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2

Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., & Stoeck, T. (2021a). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter-laboratory study. *Ecological Indicators, 121*, e107049. doi:10.1016/j.ecolind.2020.107049

Dully, V., Rech, G., Wilding, T. A., Lanzén, A., MacKichan, K., Berrill, I., & Stoeck, T. (2021b). Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems. *Marine Pollution Bulletin, 173*, e113129. doi:10.1016/j.marpolbul.2021.113129

Dully, V., Wilding, T. A., Mühlhaus, T., & Stoeck, T. (2021c). Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning. *Computational and Structural Biotechnology Journal, 19*, 2256-2268. doi:10.1016/j.csbj.2021.04.005

Findlay, R. H., & Watling, L. (1997). Prediction of benthic impact for salmon net-pens based on the balance of benthic oxygen supply and demand. *Marine Ecology Progress Series, 155*, 147-157. doi:10.3354/meps155147

Findlay, R. H., & Watling, L. (1998). Seasonal Variation in the Structure of a Marine Benthic Microbial Community. *Microbial Ecology, 36*(1), 23-30. doi:10.1007/s002489900089

Finster, K., Liesack, W., & Thamdrup, B. (1998). Elemental sulfur and thiosulfate disproportionation by Desulfocapsa sulfoexigens sp. nov., a new anaerobic bacterium isolated from marine surface sediment. *Applied and Environmental Microbiology, 64*(1), 119-125. doi:10.1128/aem.64.1.119-125.1998

Fogarty, C., Burgess, C. M., Cotter, P. D., Cabrera-Rubio, R., Whyte, P., Smyth, C., & Bolton, D. J. (2019). Diversity and composition of the gut microbiota of Atlantic salmon (Salmo salar) farmed in Irish waters. *Journal of Applied Microbiology, 127*(3), 648-657. doi:10.1111/jam.14291

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment, 189*(7), 316. doi:10.1007/s10661-017-6025-0

Freeman, E. A., Moisen, G. G., Coulston, J. W., & Wilson, B. T. (2015). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Canadian Journal of Forest Research, 46*(3), 323-339. doi:10.1139/cjfr-2014-0562

Frühe, L. (2021). *The potential of benthic microbes as bioindicators in coastal aquaculture impact assessments using eDNA metabarcoding*. Doctoral dissertation, Technische Universität Kaiserslautern.

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

Frühe, L., Dully, V., Forster, D., Keeley, N. B., Laroche, O., Pochon, X., Robinson, S., Wilding, T. A., & Stoeck, T. (2021). Global Trends of Benthic Bacterial Diversity and Community Composition Along Organic Enrichment Gradients of Salmon Farms. *Frontiers in Microbiology, 12*, e637811. doi:10.3389/fmicb.2021.637811

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., & Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences of the United States of America, 105*(22), 7774-7778. doi:10.1073/pnas.0803070105

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). R Package 'irr': Various coefficients of interrater reliability and agreement. *Version 0.84.1*.

Gao, J., Hou, L., Zheng, Y., Liu, M., Yin, G., Li, X., et al. (2016). nirS-Encoding denitrifier community composition, distribution, and abundance along the coastal wetlands of China. *Applied Microbiology and Biotechnology, 100*(19), 8573-8582. doi:10.1007/s00253-016-7659-5

Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Huse, S., & Joint, I. (2009). The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology, 11*(12), 3132-3139. doi:10.1111/j.1462-2920.2009.02017.x

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307. doi:10.1111/2041-210X.12595

Gribben, P. E., Nielsen, S., Seymour, J. R., Bradley, D. J., West, M. N., & Thomas, T. (2017). Microbial communities in marine sediments modify success of an invasive macrophyte. *Scientific Reports, 7*(1), 9845. doi:10.1038/s41598-017-10231-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). Random forests. In *The elements of statistical learning* (2nd ed., pp. 587-604). New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009b). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Helbing, C. C., & Hobbs, J. (2019). Environmental DNA standardization needs for fish and wildlife population assessments and monitoring. Canadian Standards Association. *Retrieved on 10.02.2022 from https://www.csagroup.org/wp-content/uploads/CSA-Group-Research-Environmental-DNA.pdf.*

Herlemann, D. P. R., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., & Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal, 5*(10), 1571-1579. doi:10.1038/ismej.2011.41

Hewson, I., & Fuhrman, J. (2006). Spatial and vertical biogeography of coral reef sediment bacterial and diazotroph communities. *Marine Ecology Progress Series, 306*, 79-86. doi:10.3354/meps306079

Hirche, H. J. (1987). Temperature and plankton. *Marine Biology, 94*(3), 347-356. doi:10.1007/BF00428240

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Jansson, M., Bergström, A. K., Lymer, D., Vrede, K., & Karlsson, J. (2006). Bacterioplankton growth and nutrient use efficiencies under variable organic carbon and inorganic phosphorus ratios. *Microbial Ecology, 52*(2), 358-364. doi:10.1007/s00248-006-9013-4

Kawahara, N., Shigematsu, K., Miyadai, T., & Kondo, R. (2009). Comparison of bacterial communities in fish farm sediments along an organic enrichment gradient. *Aquaculture, 287*(1), 107-113. doi:10.1016/j.aquaculture.2008.10.003

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

Keeley, N. B., Forrest, B. M., Crawford, C., & Macleod, C. K. (2012). Exploiting salmon farm benthic enrichment gradients to evaluate the regional performance of biotic indices and environmental indicators. *Ecological Indicators, 23*, 453-466. doi:10.1016/j.ecolind.2012.04.028

Keeley, N. B., Forrest, B. M., & Macleod, C. K. (2013). Novel observations of benthic enrichment in contrasting flow regimes with implications for marine farm monitoring and management. *Marine Pollution Bulletin, 66*(1-2), 105-116. doi:10.1016/j.marpolbul.2012.10.024

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *Plos One, 9*(1), e86175. doi:10.1371/journal.pone.0086175

Kennedy, R., Arthur, W., & Keegan, B. F. (2011). Long-term trends in benthic habitat quality as determined by Multivariate AMBI and Infaunal Quality Index in relation to natural variability: A case study in Kinsale Harbour, south coast of Ireland. *Marine Pollution Bulletin, 62*(7), 1427-1436. doi:10.1016/j.marpolbul.2011.04.030

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., & Ripley, B. D. (2018). R Package 'quantreg': Estimation and inference methods for models of conditional quantiles. *Version 5.88*.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*, 363-374. doi:10.2307/2529786

Lanzén, A., Dahlgren, T. G., Bagi, A., & Hestetun, J. T. (2021). Benthic eDNA metabarcoding provides accurate assessments of impact from oil extraction, and ecological insights. *Ecological Indicators, 130*, 108064. doi:10.1016/j.ecolind.2021.108064

Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology, 30*, 2969-2987. doi:10.1111/mec.15489

Laroche, O., Wood, S. A., Tremblay, L. A., Lear, G., Ellis, J. I., & Pochon, X. (2017). Metabarcoding monitoring analysis: the pros and cons of using co-extracted environmental DNA and RNA data to assess offshore oil production impacts on benthic communities. *PeerJ, 5*, e3347. doi:10.7717/peerj.3347

Lejzerowicz, F., Esling, P., & Pawlowski, J. (2014). Patchiness of deep-sea benthic Foraminifera across the Southern Ocean: Insights from high-throughput DNA sequencing. *Deep Sea Research Part II: Topical Studies in Oceanography, 108*, 17-26. doi:10.1016/j.dsr2.2014.07.018

Lepš, J., & Šmilauer, P. (2020). *Biostatistics with R: an introductory guide for field biologists*. Cambridge: Cambridge University Press.

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R news, 2(3)*, 18-22.

Lin, W.-J., Chiu, M.-C., Lin, C.-W., & Lin, H.-J. (2021). Effects of Sediment Characteristics on Carbon Dioxide Fluxes Based on Interacting Factors in Unvegetated Tidal Flats. *Frontiers in Marine Science, 8*. doi:10.3389/fmars.2021.670180

Lindh, M. V., Sjöstedt, J., Andersson, A. F., Baltar, F., Hugerth, L. W., Lundin, D., Muthusamy, S., Legrand, C., & Pinhassi, J. (2015). Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environmental Microbiology, 17*(7), 2459-2476. doi:10.1111/1462-2920.12720

Liu, S., Ren, H., Shen, L., Lou, L., Tian, G., Zheng, P., & Hu, B. (2015). pH levels drive bacterial community structure in sediments of the Qiantang River as determined by 454 pyrosequencing. *Frontiers in Microbiology, 6*, e285. doi:10.3389/fmicb.2015.00285

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

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

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal, 6*(3), 610-618. doi:10.1038/ismej.2011.139

Menchaca, I., Rodriguez, J. G., Borja, A., Belzunce, M., Franco, J., Garmendia, J., & Larreta, J. (2014). Determination of polychlorinated biphenyl and polycyclic aromatic hydrocarbon marine regional Sediment Quality Guidelines within the European Water Framework Directive. *Chemistry and Ecology, 30*(8), 693-700. doi:10.1080/02757540.2014.917175

Nybakken, J. W., & Bertness, M. (2004). *Marine biology: an ecological approach* (6th ed.). San Francisco: Pearson Education.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2020). R Package 'vegan': Community Ecology Package. *Version 2.5-7*.

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., & Aerts, H. J. W. L. (2015). Machine learning methods for quantitative radiomic biomarkers. *Scientific Reports, 5*(1), 1-11. doi:10.1038/srep13087

Pawlowski, J., Bonin, A., Boyer, F., Cordier, T., & Taberlet, P. (2021). Environmental DNA for biomonitoring. *Molecular Ecology, 30*(13), 2931-2936. doi:10.1111/mec.16023

Pawlowski, J., Bruce, K., Panksep, K., Aguirre, F. I., Amalfitano, S., Apothéloz-Perret-Gentil, L., et al. (2022). Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. *Science of the Total Environment, 818*, 151783. doi:10.1016/j.scitotenv.2021.151783

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (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

Phillips, G. R., Anwar, A., Brooks, L., Martina, L. J., Miles, A. C., & Prior, A. (2014). Infaunal quality index: Water Framework Directive classification scheme for marine benthic invertebrates. *Retrieved on 01.01.2022 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/314673/Water_Framework_Directive_classification_scheme_for_marine_benthic_invertebrates_-_report.pdf*

Prasad, A., Iverson, L., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems, 9*, 181-199. doi:10.1007/s10021-005-0054-1

Prosser, J. I., Bohannan, B. J., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., et al. (2007). The role of ecological theory in microbial ecology. *Nature Reviews Microbiology, 5*(5), 384-392. doi:10.1038/nrmicro1643

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ, 4*, e2584. doi:10.7717/peerj.2584

Rusch, A., Huettel, M., Reimers, C. E., Taghon, G. L., & Fuller, C. M. (2003). Activity and distribution of bacterial populations in Middle Atlantic Bight shelf sands. *Microbial Ecology, 44*(1), 89-100. doi:10.1111/j.1574-6941.2003.tb01093.x

SEPA (2018). Scottish Environmental Protection Agency. Fish farm survey report-evaluation of a new seabed monitoring approach to investigate the impacts of marine cage fish farms. *Retrived on 20.06.2020 from https://consultation.sepa.org.uk/sectorplan/finfishaquaculture/supporting_documents/Fish%20Farm%20Survey%20Report.pdf.*

Serrana, J. M., Li, B., Sumi, T., Takemon, Y., & Watanabe, K. (2022). Implications of taxonomic and numerical resolution on DNA metabarcoding-based inference of benthic macroinvertebrate responses to river restoration. *Ecological Indicators, 135*, 108508. doi:10.1016/j.ecolind.2021.108508

Stigebrandt, A. (1983). A model for the exchange of water and salt between the Baltic and the Skagerrak. *Journal of Physical Oceanography, 13*, 411-427. doi:10.1175/1520-0485(1983)013<0411:AMFTEO>2.0.CO;2

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Tait, K., Laverock, B., & Widdicombe, S. (2014). Response of an Arctic Sediment Nitrogen Cycling Community to Increased CO2. *Estuaries and Coasts, 37*(3), 724-735. doi:10.1007/s12237-013-9709-x

Tande, K. S. (1988). The effects of temperature on metabolic rates of different life stages of Calanus glacialis in the Barents Sea. *Polar Biology, 8*(6), 457-461. doi:10.1007/BF00264722

UKTAG (2012). United Kingdom Technical Advisory Group. Practitioners guide to the infaunal quality index. Water Framework Directive: Transitional and Coastal Waters. . *Retrieved on 24.02.2022 from https://www.wfduk.org/sites/default/files/Media/Environmental%20standards/Annex%2018%20Transitional%20and%20coastal%20waters%20Invertebrates%20IQI.pdf*

Verhoeven, J. T. P., Salvo, F., Knight, R., Hamoutene, D., & Dufour, S. C. (2018). Temporal Bacterial Surveillance of Salmon Aquaculture Sites Indicates a Long Lasting Benthic Impact With Minimal Recovery. *Frontiers in Microbiology, 9*, e03054. doi:10.3389/fmicb.2018.03054

Vincent, S. G. T., Jennerjahn, T., & Ramasamy, K. (2021). Chapter 3 - Environmental variables and factors regulating microbial structure and functions. In S. G. T. Vincent, T. Jennerjahn, & K. Ramasamy (Eds.), *Microbial Communities in Coastal Sediments* (1st ed., pp. 79-117). Amsterdam, Oxford, Cambridge: Elsevier.

Wang, K., Zou, L., Lu, X., & Mou, X. (2018). Organic carbon source and salinity shape sediment bacterial composition in two China marginal seas and their major tributaries. *Science of the Total Environment, 633*, 1510-1517. doi:10.1016/j.scitotenv.2018.03.295

Wickramasinghe, M. P., Sudarshani, K. A. M., & Wegiriya, H. C. E. (2021). The diversity of marine invertebrate macrofauna in selected rocky intertidal zones of Matara, Sri Lanka. *Asian Journal of Conservation Biology, 10*(1), 15-21. doi:10.53562/ajcb.OZDK5526

Yakimov, M. M., Timmis, K. N., & Golyshin, P. N. (2007). Obligate oil-degrading marine bacteria. *Current opinion in biotechnology, 18*(3), 257-266. doi:10.1016/j.copbio.2007.04.006

Yang, J., Ma, L. a., Jiang, H., Wu, G., & Dong, H. (2016). Salinity shapes microbial diversity and community structure in surface sediments of the Qinghai-Tibetan Lakes. *Scientific Reports, 6*(1), 25078. doi:10.1038/srep25078

Yukgehnaish, K., Kumar, P., Sivachandran, P., Marimuthu, K., Arshad, A., Paray, B. A., & Arockiaraj, J. (2020). Gut microbiota metagenomics in aquaculture: factors influencing gut microbiome and its physiological role in fish. *Reviews in Aquaculture, 12*(3), 1903-1927. doi:10.1111/raq.12416

Zhang, J., Zhang, Z. F., Liu, S. M., Wu, Y., Xiong, H., & Chen, H. T. (1999). Human impacts on the large world rivers: Would the Changjiang (Yangtze River) be an illustration? *Global Biogeochemical Cycles, 13*(4), 1099-1105. doi:10.1029/1999GB900044

Zheng, B., Wang, L., & Liu, L. (2014). Bacterial community structure and its regulating factors in the intertidal sediment along the Liaodong Bay of Bohai Sea, China. *Microbiological Research, 169*(7), 585-592. doi:10.1016/j.micres.2013.09.019

CHAPTER IV
Required sequencing depth for SML-based EQ inference

Summary

Background

Recently, many studies have demonstrated that eDNA metabarcoding-based environmental monitoring has a great potential to be included in legal regulations (Goldberg et al., 2016; Helbing and Hobbs, 2019; Kelly et al., 2014). As eDNA-based monitoring can save time and money and is additionally easier to use, it has the ability to complement or replace traditional macrofauna monitoring (reviewed in Pawlowski et al., 2018). SML methods as RF have emerged as a powerful tool for eDNA data analysis, as it is suitable for such complex datasets containing high numbers of variables (Cordier et al., 2017; Cutler et al., 2007; Evans et al., 2011; Smucker et al., 2020). The RF prediction models based on eDNA were demonstrated to correctly classify various parameters like EQ of sediments, ship ballast water origin at harbors, and other anthropogenic influences (Aylagas et al., 2017; Cordier et al., 2017; Gerhard and Gunsch, 2019). Those predictive models can subsequently enable ecological classification of new samples with unknown conditions, which can thus be used for monitoring approaches.

Datasets recently obtained for SML applications are usually sequenced to full saturation to detect the maximum diversity present. Those studies, therefore, exhibit HTS sequencing, acquiring from approx. 30,000 (Dully et al., 2021a), 40,000 (Lanzén et al., 2020), and 70,000 (Cordier et al., 2017), up to 223,000 HQ sequences per sample (Gerhard and Gunsch, 2019). Because only a certain number of sequences can be obtained by each flow cell compartment, sequencing gets more expensive the more sequence reads are produced. An advantage of using metabarcoding is multiplexing, which involves combining multiple samples into one reaction vessel for sequencing (Illumina, 2017; Taberlet et al., 2018). This is possible as sequences taken from different samples can be tagged with identifier nucleotides and sequenced together (Meyer et al., 2007; Nielsen et al., 2006; Taberlet et al., 2018). Therefore, multiplexing results in simultaneous sequencing of many samples, but consequently also in a decreased number of sequences per sample (Illumina, 2017).

To save sequencing costs and computational time, it is important to assess how many sequences are sufficient to make accurate predictions about unknown samples using corresponding predictive models. To answer this question, the conducted study will give insight into the sequencing depth required for eDNA-based environmental parameter inference when using SML.

Methods

To account for spatiotemporal dynamics of the bacterial community and therefore eDNA, four different datasets focusing on different anthropogenic impacts were analyzed. Three datasets were already published in the literature, so the existing raw read sequences were available. The 'BasCo' dataset contains sequences deriving from a study on urban anthropogenic impacts categorizing the estuarine bacterial communities by a newly developed environmental index (microgAMBI). The samples were classified according to four different quality classes from high to poor quality (Aylagas et al., 2017). Therefore, this dataset was used to predict the microgAMBI classifications 'high', 'good', 'moderate', or 'poor'. Another dataset ('BallWa') was obtained from a study testing RF prediction performance on sample types introduced to harbor water by the discharge of ship ballast water of arriving ships (Gerhard and Gunsch, 2019). For RF predictive models, prediction of the country of origin ('China', 'Singapore', or 'USA') was conducted. The third dataset ('NorSa') which was used for the analysis, contained sequences derived from Norwegian salmon farm sediments (Cordier et al., 2018). According to the traditional EQ inference, the samples were classified into four different EQ categories, namely 'good', 'moderate', 'poor', and 'bad'. Also, the samples could be separated by their salmon farm of origin, allowing for a second analysis of prediction performance. Therefore, the RF algorithm was trained on distinguishing between the aquafarm locations, namely 'Aukrasanden', 'Beitveitnes', 'Bjørnsvik', 'Nedre Kvarv', and 'Storvika'. The fourth dataset was a novel one and was therefore originally processed. It is based on eDNA metabarcoding sequences focusing on V3-V4 bacterial amplicons deriving from sediments beneath a salmon farm in Scotland. Samples were taken at different months representing different stages of the salmon production cycle. The RF models were used to predict the phase of the production cycles ('pre-production phase', 'early production phase' and 'late production phase') as well as the station representing the distance gradient of the samples from the salmon cages (cage edge 'CE', allowable zone of effect 'AZE', and reference 'REF').

In total, the four analyzed datasets were used for the prediction of six separate measures: microgAMBI, country of ballast water origin, EQ, aquafarm location, production cycle phase, and station. For all measures, independent RF analyses were conducted to determine how many sequences are sufficient for correct sample classification. To gain the best prediction accuracy possible, each measure was predicted using the respective full dataset, so every obtained HQ sequence was included for RF model construction. As a measure of model performance, RF internal out-of-bag (OOB) error estimate was used, representing the prediction accuracy when analyzing novel samples (Breiman, 2001; Hastie et al., 2009a). Additionally, the kappa statistic κ was inferred for each predictive model indicating goodness of agreement between reference and prediction excluding potential bias introduced by random agreement (Landis and Koch, 1977). Those models can be referred to as benchmark models, as they represent the best prediction possible by using the whole conglomerate of sequences.

Subsequently, for each of the six measures, predictive models using reduced datasets were constructed. This process included passing different numbers of sequences to the algorithm for model construction, down to a minimum of 50 sequences per sample. All datasets were randomly reduced to 12,500, 10,000, 7,500, 5,000, 2,500, 1,000, 500, 400, 300, 200, 100 and 50 sequences per sample, respectively (*Figure 18*).

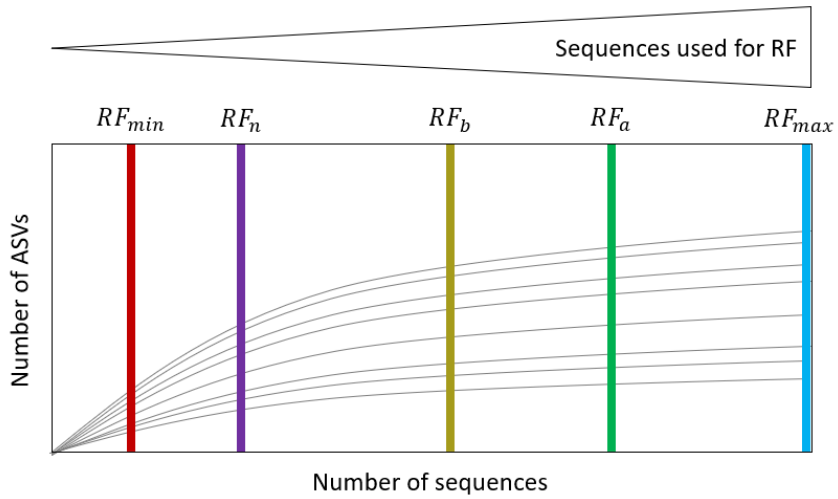


Figure 18) A schematic representation of the RF downsampling process. *A traditional saturation curve is presented by the gray lines indicating sequencing depth. Saturation curves are based on the obtained number of ASV per number of sequences. First, full RF models are constructed using all available sequences (blue line, RF_{max}). The prediction accuracy can be referred to as the benchmark accuracy. For downsampling analysis, a decreasing number of sequences is subsampled randomly. The first downsampling is conducted at the green line (RF_a). For all samples of the respective dataset, a specific number of samples lower than the initial sequence number at RF_{max} is used for the RF model construction. Further downsampling steps (RF_b-RF_n, ochre line and purple line) are included. After that, each sample is reduced regarding the number of sequences until RF_{min} is reached. RF_{min} (red line) represents the stage of downsampling mirroring the benchmark prediction accuracy obtainable by RF_{max}.*

Results and Discussion

It was demonstrated that the various prediction objectives of the datasets were correctly predicted to a different extent. The maximal achievable prediction accuracy, when using all sequences of a dataset, reached 92.6% (κ = 0.88) of correctly classified salmon production phases for the ScoSa dataset indicating almost perfect agreement. The minimal benchmark accuracy was 78.3% (κ = 0.71) in the BasCo dataset for microgAMBI classification, indicating moderate agreement. When comparing benchmark prediction accuracies with the accuracies obtained by using the reduced dataset, a decrease of accuracy was more or less pronounced regarding different measures to be predicted. For the measure 'salmon production phase', as many as 50 sequences were enough to obtain an almost perfect agreement between predicted categories and reference categories (RF_{min}= 50). For the measures 'station' and 'microgAMBI', 5,000 sequences were required (RF_{min}= 5,000) to obtain estimates comparable to the benchmark prediction. For the other measures, 1,000 to 2,500 sequences were sufficient for maintaining benchmark performance.

To exclude bias based on an unequal distribution of categories in the reference data, an analysis of inequality among the categories compared to their respective RF_{min} was conducted. It has been reported before that unequal sample distribution can lead to biased predictions (He and Garcia, 2009). Contrary, RF was also demonstrated to handle those inequalities well (Guo et al., 2016) as was the case for the conducted study. Additionally, no coherence between the number of classes to be predicted and the total number of samples on RF_{min} could be detected. Exemplary, the ScoSa dataset used for the prediction of 'station' and 'salmon production phase' consisted of 76 samples each. Both measures to be predicted consisted of three categories each, while resulting RF_{min} values to reach the benchmark performance differed between 50 and 5,000 sequences.

No influence of the number of used samples in total, the number of categories to be predicted or the evenness of sample distribution on the prediction accuracies were identified. However, the separability of sample groups was a good indicator for the required number of sequences. This corroborates well with the literature, as it has been reported before that a high sample separability allows an improved predictive performance in the field of image recognition (Foody, 2002; Millard and Richardson, 2015). As a measure of separability, ordination analysis like non-metric multidimensional scaling (NMDS) were consulted.

NMDS uses a beta diversity-based sample dissimilarity matrix to visualize (dis)similarities among sample composition (e.g., ASV composition or bacterial community composition) in a simplified, dimension-reduced graph. For a better comprehension of the degree of similarity, confidence intervals per category were added as they can represent the boundaries of sample classes. The less the CIs overlap, the better is the expected separability of the respective categories. If the CIs are mostly overlapping, the separability of categories is reduced. Therefore, more sequences are required to reach the benchmark RF prediction performance. Exemplary, an NMDS diagram for the 'ScoSa' dataset showing sample separation among the different prediction objectives measures is presented in *Figure 19*.

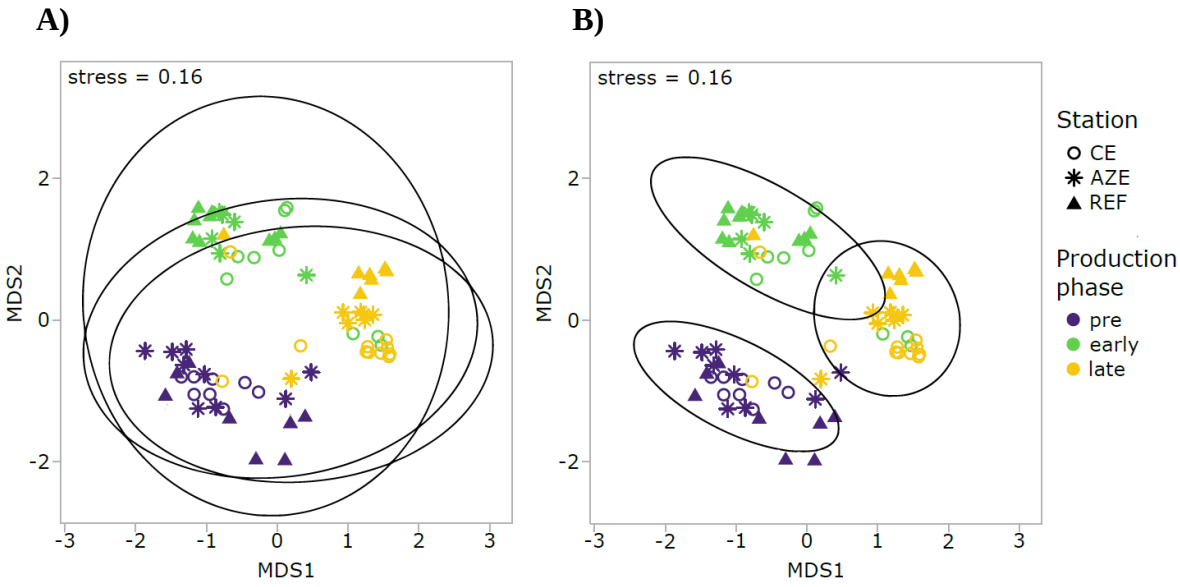


Figure 19) Non-metric multidimensional scaling of the eDNA 'ScoSa' dataset. *The closer the points are to each other, the more similar the bacterial community was among the samples. The sampling station is indicated by different shapes, unfilled points represent cage edge (CE) station which is directly located at the salmon farm cage. The triangle indicates samples taken from a reference site (REF) under pristine conditions. A star represents a sampling station between CE and REF, which were referred to as the allowable zone of effect (AZE). The salmon production phase is indicated by colors. The purple color represents the pre-production phase, the green color represents the early production phase, and the golden color represents the late production phase. The black ellipses represent 95% confidence intervals (CIs) calculated for the station (A) and the salmon production phase (B). The 95% CIs represent the area including a normally distributed sample with a 95% probability. For station (A), confidence intervals are overlapping, while for the salmon production phase (B), a mostly clear separation of sample categories is visible.*

Additionally, analysis of intersections among the ASV inventories can be easily visualized using Venn diagrams. They can give a fast hint on how similar ASV inventories are among the different sample categories. Furthermore, the calculation of the coefficient of variation (CoV) can be conducted.

As the CoV represents the ratio between the standard deviation to the category mean, it can represent the variability of features (here = ASVs) among categories (Lepš and Šmilauer, 2020). The higher the CoV of an ASV, the more variable it is among categories, indicating a better predictive performance. If the combined CoV values of ASVs from a sample grouped in a distinct category are high, a good separation performance is expected. This corroborates well with the findings that different categories showed different CoV values along with different RF_{min} values.

This study focused on various anthropogenic influences as organic enrichment introduced by salmon farming in Norway and Scotland, discharge from urbanization at a coastal area, and impact by ship ballast water discharge. As expected, prediction performance varied between the different datasets and different measures to be categorized. It was demonstrated that in each case, 5,000 sequences randomly extracted per sample were sufficient to reach the benchmark predictive performance, which was based on all sequences of the respective dataset. Occasionally, even as little as 50 sequences per sample are sufficient to reach the benchmark performance. This means, the lower the separability of the samples from a habitat to be studied, the more samples and the more sequences must be available to make a good prediction. Spatiotemporally dynamic habitats, such as marine sediments, must be sampled more extensively, including sufficient replicates (Hestetun et al., 2021a). Also, variation in the bacterial community over the course of the year should be considered, as the composition potentially changes among different environmental factors triggered by seasonal succession or farm production phases (Delille, 1995; Dully et al., 2021c; Findlay and Watling, 1998; Gilbert et al., 2009). However, even in the dataset with the lowest separability of samples, merely 5000 sequences were required to achieve the best possible prediction. Therefore, it was indicated that 5,000 sequences describe the mandatory amount of diversity to foretell various prediction objectives, corresponding with the findings of Lanzén et al. (2017). The authors reported a sharp decrease in sample alpha diversity measures when reducing the number of sequences per sample to less than 5,000. Therefore, it can be assumed that often an excessive number of sequences is obtained, causing avoidable costs and use of computational resources. For eDNA metabarcoding of highly dynamic marine sediments, it is recommended to sample many samples at a shallow sequencing depth rather than deep-sequencing fewer samples, to increase coverage of spatiotemporal variability. This also implies that multiplexing, and thus parallel sequencing, is very well suited to determine the EQ of the respective ecosystem, resulting in an even greater upscaling potential than assumed.

In summary, it was shown that neither the distribution of samples among categories to be predicted, nor the number of those categories, nor the number of samples in total influenced the RF prediction performance. Separability of the samples among categories seemed to play a major role in the inference of the minimum number of sequences needed for RF prediction. The higher the separability of sample categories, the lower was the minimum number of sequences needed for optimal RF predictions. For the implementation of bacterial eDNA-based SOPs for environmental monitoring approaches, it was confirmed that 5,000 sequences per sample are sufficient to make accurate estimations of marine benthic EQ. Therefore, sequencing costs and the usage of computational resources can be reduced. However, as the marine coastal ecosystem is very dynamic in space and time, pilot studies should be conducted for each new habitat type (Loeza-Quintana et al., 2020). Samples should be tested for separability of the categories by using ordination analysis, Venn diagrams, and CoV inference. Using those methods, the required sequencing depth for benchmark predictions can be anticipated, which prevents from using superfluous resources.

References

Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., Irigoien, X., Danovaro, R., & Rodríguez-Ezpeleta, N. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Marine Pollution Bulletin, 114*(2), 679-688. doi:10.1016/j.marpolbul.2016.10.050

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01518

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources, 18*(6), 1381-1391. doi:10.1111/1755-0998.12926

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for classification in Ecology. *Ecology, 88*(11), 2783-2792. doi:doi.org/10.1890/07-0539.1

Delille, D. (1995). Seasonal changes of subantarctic benthic bacterial communities. *Hydrobiologia, 310*(1), 47-57. doi:10.1007/BF00008182

Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., & Stoeck, T. (2021a). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter-laboratory study. *Ecological Indicators, 121*, e107049. doi:10.1016/j.ecolind.2020.107049

Dully, V., Wilding, T. A., Mühlhaus, T., & Stoeck, T. (2021c). Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning. *Computational and Structural Biotechnology Journal, 19*, 2256-2268. doi:10.1016/j.csbj.2021.04.005

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2011). Modeling Species Distribution and Change Using Random Forest. In C. A. Drew, Y. F. Wiersma, & F. Huettmann (Eds.), *Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications* (pp. 139-159). New York: Springer.

Findlay, R. H., & Watling, L. (1998). Seasonal Variation in the Structure of a Marine Benthic Microbial Community. *Microbial Ecology, 36*(1), 23-30. doi:10.1007/s002489900089

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment, 80*(1), 185-201. doi:10.1016/S0034-4257(01)00295-4

Gerhard, W., & Gunsch, C. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environment International, 124*, 312-319. doi:10.1016/j.envint.2018.12.038

Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Huse, S., & Joint, I. (2009). The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology, 11*(12), 3132-3139. doi:10.1111/j.1462-2920.2009.02017.x

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307. doi:10.1111/2041-210X.12595

Guo, F., Wang, G., Su, Z., Liang, H., Wang, W., Lin, F., & Liu, A. (2016). What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire, 25*(5), 505-519. doi:10.1071/WF15121

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). Random forests. In *The elements of statistical learning* (2nd ed., pp. 587-604). New York: Springer.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. doi:10.1109/TKDE.2008.239

Helbing, C. C., & Hobbs, J. (2019). Environmental DNA standardization needs for fish and wildlife population assessments and monitoring. Canadian Standards Association. *Retrieved on 10.02.2022 from https://www.csagroup.org/wp-content/uploads/CSA-Group-Research-Environmental-DNA.pdf.*

Hestetun, J. T., Lanzén, A., & Dahlgren, T. G. (2021a). Grab what you can—an evaluation of spatial replication to decrease heterogeneity in sediment eDNA metabarcoding. *PeerJ, 9*, e11619. doi:10.7717/peerj.11619

Illumina (2017). Metagenomic Sequencing Library Preparation. *Retrieved on 01.01.2022 from https://www.illumina.com/content/dam/illumina-marketing/documents /products/illumina_sequencing_introduction.pdf.*

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *Plos One, 9*(1), e86175. doi:10.1371/journal.pone.0086175

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*, 363-374. doi:10.2307/2529786

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2017). DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *Plos One, 12*(6), e0179443. doi:10.1371/journal.pone.0179443

Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology, 30*, 2969-2987. doi:10.1111/mec.15489

Lepš, J., & Šmilauer, P. (2020). *Biostatistics with R: an introductory guide for field biologists*. Cambridge: Cambridge University Press.

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

Meyer, M., Stenzel, U., Myles, S., Prüfer, K., & Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research, 35*(15), e97. doi:10.1093/nar/gkm566

Millard, K., & Richardson, M. (2015). On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sensing, 7*(7), 8489-8515. doi:10.3390/rs70708489

Nielsen, K. L., Høgh, A. L., & Emmersen, J. (2006). DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Research, 34*(19), e133. doi:10.1093/nar/gkl714

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (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

Smucker, N. J., Pilgrim, E. M., Nietch, C. T., Darling, J. A., & Johnson, B. R. (2020). DNA metabarcoding effectively quantifies diatom responses to nutrients in streams. *Ecological Applications, 30*(8), e02205. doi:10.1002/eap.2205

Taberlet, P., Bonin, A., Zinger, L., & Coissac, É. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press.

Publication:

Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning

Verena Dully[a], Thomas A. Wilding[b], Timo Mühlhaus[c], Thorsten Stoeck[a,*]

[a] *Technische Universität Kaiserslautern, Ecology, D-67663 Kaiserslautern, Germany*

[b] *Scottish Association for Marine Science, Scottish Marine Institute, Oban, Scotland, United Kingdom*

[c] *Technische Universität Kaiserslautern, Computational Systems Biology, D-67663 Kaiserslautern, Germany*

[*] *corresponding author*

From:

Dully, V., Wilding, T.A, Mühlhaus, T., Stoeck, T. (2021). *Computational and Structural Biotechnology Journal, 19,* 2256-2268. doi:10.1016/j.csbj.2021.04.005

Abstract

Environmental DNA metabarcoding is a powerful approach for use in biomonitoring and impact assessments. Amplicon-based eDNA sequence data are characteristically highly divergent in sequencing depth (total reads per sample) as influenced inter alia by the number of samples simultaneously analyzed per sequencing run. The random forest (RF) machine learning algorithm has been successfully employed to accurately classify unknown samples into monitoring categories. To employ RF to eDNA data, and avoid sequencing-depth artifacts, sequence data across samples are normalized using rarefaction, a process that inherently loses information. The aim of this study was to inform future sampling designs in terms of the relationship between sampling depth and RF accuracy. We analyzed three published and one new bacterial amplicon datasets, using a RF, based initially on the maximal rarefied data available (minimum mean of >30,000 reads across all datasets) to give our baseline performance. We then evaluated the RF classification success based on increasingly rarefied datasets. We found that extreme to moderate rarefaction (50–5000 sequences per sample) was sufficient to achieve prediction performance commensurate to the full data, depending on the classification task. We did not find that the number of classification classes, data balance across classes, or the total number of sequences or samples, were associated with predictive accuracy. We identified the ability of the training data to adequately characterize the classes being mapped as the most important criterion and discuss how this finding can inform future sampling design for eDNA based biomonitoring to reduce costs and computation time.

Abbreviations

AMBI, AZTI's marine biotic index; ASV, Amplicon Sequence Variants; AZE, allowable zone of effect, intermediate impact zone; BallWa, ballast water dataset; BasCo, Basque coast dataset; BI, biotic index; bp, base pairs; CE, cage edge; CV, Coefficient of Variance; DADA2, Divisive Amplicon Denoising Algorithm; eDNA, environmental deoxyribonucleic acid; EQ, environmental quality; FM, full model; MDS, multidimensional scaling; microgAMBI, AZTI's marine biotic index based on microbial genes; *mtry*, numbers of variables tried at each split; *n*, number; NEB, New England Biolabs; NorSa, Norway salmon dataset; NW, north west; OOB error, out-of-bag error estimate; PCR, polymerase chain reaction; REF, reference site; RF, random forest algorithm; rRNA, small subunit prokaryotic ribosomal ribonucleic acid; ScoSa, Scottish

salmon farm dataset; SML, supervised machine learning; V3-V4, hypervariable gene regions of the 16s rRNA

1. Introduction

Marine coastal ecosystems offer numerous ecosystem services and therefore are subject to a multitude of stressors from anthropogenic activities, resulting in eutrophication, pollution, overexploitation, and introduction of invasive species [1–4]. These local stressors are complemented by global effects such as increasing temperatures, sea level rise, and ocean acidification [5,6]. These stressors may severely affect marine coastal ecosystems and compromise ecosystem services [7]. Therefore, environmental biomonitoring programs for an efficient management and protection of marine coastal ecosystem are in place, which are laid down in national and international Directives, such as the Marine Strategy Framework Directive [8].

The biological component is a backbone of environmental monitoring. In contrast to chemical monitoring approaches, which provide only an environmental quality snapshot, biological indicators are affected by the total range of environmental species they are exposed to, and thus provide a cumulative measure of environmental health [9]. Traditional methods applied to analyze marine bioindicators (mostly meio- and macrofauna) are based on morphological identification and observational surveys. Such surveys are time consuming, expensive, and characterized by low upscaling potential for high-throughput monitoring to resolve environmental changes on small spatial and temporal scales. In addition, evident limitations of this traditional approach are the identification and quantification of rare species and the ability to distinguish morphologically close or identical species (i.e., cryptic species), or poorly characterized juvenile stages of known species [10].

Concerted efforts of the scientific community in recent years were therefore the development of fast, less expensive, and more robust coastal biomonitoring methods with high potential for automation and upscaling. Environmental DNA (eDNA) metabarcoding of marine communities emerged as a very promising strategy that meets these requirements. It uses short, standardized gene regions obtained from environmental samples as internal taxon tags to provide rapid characterization of whole communities. Recently, a remarkable number of applied environmental metabarcoding studies tested the potential use of metabarcoding data to assess the ecological status of natural marine communities exposed

to various anthropogenic pressures (reviewed in [11]). In specific, bacteria and protists, which dominate most ecosystems in terms of biomass, and structural and functional diversity are likely the best option on which to perform efficient next generation marine biomonitoring [9,12–22].

A major challenge in eDNA-based biomonitoring was the inference of biotic indices (BI), which inform about environmental quality (EQ) of the ecosystem under study.
One solution was the development of specific indices such as the microgAMBI that exploits the taxonomic information obtained from eDNA metabarcodes and the ecological function of identified microbial taxa [12,13]. Because of severely incomplete gene reference databases for microbial taxa, the microgAMBI relies on the ecological functions of higher taxon ranks rather than species, which were obtained from previously published reports. Consequently, a large proportion of the obtained eDNA metabarcode datasets, which could not be assigned to the required taxonomic ranks, but which may be important indicators, cannot be used for the inference of the microgAMBI. As an alternative, other authors correlated obtained amplicon sequence variants (ASVs) of microbial communities to gradients of environmental stressors and assigned an index value to significantly correlating ASVs [16,17]. These indicator ASVs were then used as parameters in modified versions of traditional BIs originally developed for macroinvertebrate bioindicators. The most promising approach, however, to infer EQ from metabarcode datasets is supervised machine learning (SML) [16,23]. The principle and power of this approach is reviewed in detail in Cordier et al. 2018 [9]. In brief, SML is taxonomy independent and does not rely on available knowledge about the ecology of the species hidden behind microbial ASVs. This eliminates difficulties relating to incomplete nucleotide reference databases and a lack of knowledge about the ecology of numerous yet unknown marine microbes. Classification via SML is first used on a training dataset, which consists of two sets of data that are obtained from the same samples. These are the ASVs of the microbial community in this sample and the reference labels (for example the BI obtained from conventional macroinvertebrate monitoring of the same sample). A predictive model is trained to link specific bacterial ASVs to specific reference groups. The accuracy of a model can then be evaluated with a kappa statistic: kappa values ranging from 0.01 to 0.2 indicate ‘‘poor agreement” between two classifications (traditional macrofauna-based vs. eDNA-based EQ classification); and values of > 0.8 indicate ‘‘perfect agreement” [24]. The successfully trained model can then be used for making predictions of reference labels on upcoming genetic metabarcode samples without collecting additional data as reference. The most

successfully applied SML approach for classification using ASVs is Random Forest (RF) for massive and noisy DNA amplicon datasets [9,25–28].

This approach marrying environmental genomics and BI inference is of high relevance for industry and politics (environmental management and decision making) alike. A decisive criterion for the implementation of this approach in routine monitoring practice is the costs associated with each technology.

A part of this costs depends on the required depths of sequencing to make as accurate as possible inference of BI and EQ for an ecosystem under surveillance. In our study we use three eDNA datasets from previous reports and one new original dataset to infer the minimal sequence depths for marine microbial communities to exploit these data with an RF approach to infer the origin of ballast water and to predict EQ of ecosystems under the impact of urban infrastructure and of aquaculture disturbance. The main questions are:(1) What is the lower limit of sequences for accurate RF predictions in marine coastal monitoring using microbial communities? (2) Is this limit the same for different monitoring targets? A major goal of the study is to inform adequate sampling designs for future eDNA metabarcoding-based marine coastal monitoring surveys.

2. Methods

2.1. Datasets

We have analyzed four datasets of bacterial Illumina amplicons of the hypervariable V3-V4 16S rRNA gene region. The first dataset [29] included 6,213,619 sequences, obtained from 51 sediment samples of the Basque coast, subjected to various anthropogenic impacts [12]. The authors inferred a novel biotic index, microgAMBI, from these data to assess EQ for each of the samples. The second dataset [30], published by Gerhard and Gunsch [31] included 22,105,927 sequences, obtained from 68 ballast and harbor water samples, to train a Random Forest algorithm for the prediction of geographic ballast water origin. The third dataset [32] included 15,135,391 sequences, obtained from 129 sediment samples to predict the biotic index AMBI for the assessment of aquaculture-induced benthic disturbance at five Norwegian open cage salmon (*Salmo salar*) farming sites [9]. The fourth dataset [33] is original and was obtained from a time series of a Scottish salmon (*Salmo salar*) farm to predict distance from farm and the salmon production phase in which samples were collected. This dataset included 9,496,674 sequences in 76 samples. Details about sampling and data acquisition for this dataset will be described in the following section.

2.2. Sampling of Scottish salmon farm sediment

Sediment was collected at three stations along a northwest (NW) transect from the northwesterly cage edge (CE) to a reference site (REF) in the direction of the prevailing current flow, located ca. 800 m distant from the CE. An intermediate impact zone (AZE) was located at ca. 100 m distance from the cage edge.

Sampling occurred monthly from March 2018 to March 2019. Due to weather conditions, sampling could not be conducted in November and December 2018. At each site, three biological replicates were taken from a van Veen grab (0.1 m^2 area), each replicate consisting of 10 g of surface sediment (upper few millimeters) collected using plastic spatulas. Immediately following collection, samples were stored in the dark and on ice (max. 6 h) and then stored at -20C°. For the purposes of shipping, samples were then defrosted overnight (4C°) then transferred to equal volumes of LifeGuard nucleic acid preservation solution (Qiagen) until further processing for eDNA metabarcoding.

2.3. DNA extraction, amplification and Illumina sequencing of Scottish salmon farm samples

Following our previously described protocol [16], environmental DNA was obtained from sediment samples using the PowerSoil DNA kit (Qiagen, Hilden, Germany) according to the manufacturer's manual. As DNA metabarcodes, we obtained the ca.450 bp long hypervariable V3-V4 region of the bacterial 16S rRNA gene. The PCR protocol with the Bakt_341F(CCTACGGGNGGCWGCAG) and the Bakt_805R (GACTACHVGGGTATC-TAATCC) primer pair [34] employed an initial activation step of NEB's Phusion High-Fidelity DNA polymerase at 98 C° for 30 s, followed by 27 identical three-step cycles consisting of 98 C° for 10 s, 62 C° for30 s, and 72 C° for 30 s; then a final 5-min extension at 72 C° as previously described [16]. Standard negative controls were run with each PCR assay using the same reaction mixture as described above without adding template DNA to the mixture. From the resulting PCR products, sequencing libraries were constructed using the NEB Next UltraTM DNA Library Prep Kit for Illumina (NEB, USA). The quality of the libraries was assessed with an Agilent Bioanalyzer 2100 system. V3-V4 libraries were sequenced on an Illumina MiSeq platform, generating 2x300-bp paired-end reads. A standard negative control of a DNA template-free library as well as with PhiX Control v3 library spiked in was run with the samples.

2.4. Sequence data processing for all four datasets

Sequences were processed using the Divisive Amplicon Denoising Algorithm DADA2 [35], as described for hypervariable taxonomic marker genes from metabarcoding studies [15] with the model trained on Illumina runs and the following criteria: bacterial V3-V4 sequences were filtered using *filterAndTrim* according to the instructions with *truncLen* = 225 for V3-V4 sequences and *truncLen* = 150 for V4 sequences.

To maximize the quality of the final sequence reads used for downstream analyses, we chose the following *maxEE* values for the individual dataset: BasCo = 1, BallWa = 1, NorSa = 2, ScoSa = 1. Bacterial sequences were merged using 20 base pairs overlap with allowed mismatch of 2. To minimize ecologically uninformative noise, only ASVs with at least 50 reads were maintained for downstream analyses, similar to previous publications [9,36,37]. Samples with less than 15,000 reads were discarded. Furthermore, the South African harbor and ballast water samples from the BallWa dataset consisted of only eight samples. This is a too small sample size to allow location-specific discrimination, and, therefore, we eliminated these eight samples. Thus, fewer samples were used for our analyses than were included in the original datasets. In summary, for our analyses we included 39 samples for the Basque costal dataset (BasCo), 59 samples for the ballast water dataset (BallWa), 95 samples for the Norwegian salmon farm dataset (NorSa) and 76 samples for the new Scottish salmon farm dataset (ScoSa). After processing, the ScoSa dataset was split near-equal (in terms of sample numbers) to represent salmon production phases. These production phases were defined as pre-production phase (n samples = 25, collected between March and May 2018), early salmon production phase (n samples = 24, collected between June and August 2018) and late salmon production phase (n samples = 27, collected between September 2018 and March 2019). For details, we refer to *Supplementary File 1*[*]. Sampling in May 2018 occurred immediately after addition of salmon breed to the cages. Thus, in the preproduction phase, no salmon-related impact on the seafloor is expected. In the early salmon production phase, the average fish biomass was 107 tons in the aquaculture installation under study, whereas in the late production phase, this number had increased toon average of 680 tons.

[*] All supplementary files are additionally available at the appendix of this dissertation

2.5. Supervised Machine learning (SML) predictions

Using RF, we predicted the following measures for the four different datasets: The microgAMBI class ("high", "good", "moderate" or "poor") was predicted for the BasCo dataset, using the microgAMBI identified for each sample in the original publication [12]. Geographic origin (Singapore, USA or China, excluding South Africa) of ship ballast water was predicted for the BallWa dataset, using the ground truth data for each sample from the original publication [31]. The AMBI biotic index class ("good", "moderate", "poor" or "bad") was predicted for the NorSa dataset, which was obtained as reference from an official compliance monitoring survey of these farms using benthic macroinvertebrates [9]. Additionally, the farm of sample origin was predicted. For the new ScoSa dataset, we also predicted two variables for each sample, namely distance from farm and salmon production phase in which samples were collected. Predictive models were trained using the RF algorithm [38] implemented in the randomForest v. 4.6.14 package for classification and regression [39]. First, RF is used on a training dataset, one for each of the four bacterial ASV datasets used in this study (BasCo, BallWa, NorSa and ScoSa). Such training data consist of two sets of data that are obtained from the same sample: (a) the obtained bacterial ASVs as features and (b) the reference labels. These reference labels were microgAMBI class, ballast water origin, AMBI class/Farm and distance/salmon production phase for the BasCo, BallWa, NorSa and ScoSa datasets, respectively. The RF algorithm is then trained to relate specific (combinations of) ASVs to defined reference label values (regression) or categories (classification).

An essential feature of the RF algorithm is its use of out-of-bag (OOB) samples. For each observation, a random forest predictor is constructed by averaging only those decision trees in which this observation did not appear. Therefore, an OOB error estimate (OOB-E) is almost identical to that obtained by N-fold cross validation [40]. Setting the RF parameters for classification approaches, the inventors recommend determining the default *mtry* value to the square root number of features.

In the first step, so-called "full models" (FM) were calculated exploiting all available sequences of each of the four datasets. Prior to RF, we transformed the ASV-to-sample matrix into a relative abundance matrix for each ASV (using the number of reads for the respective ASV divided by the number of all reads in a sample) to compensate for any differences in the sequencing depth between samples [12,31]. With this matrix RF

models were calculated using features and reference labels as mentioned above. For each model we ran 6000 trees.

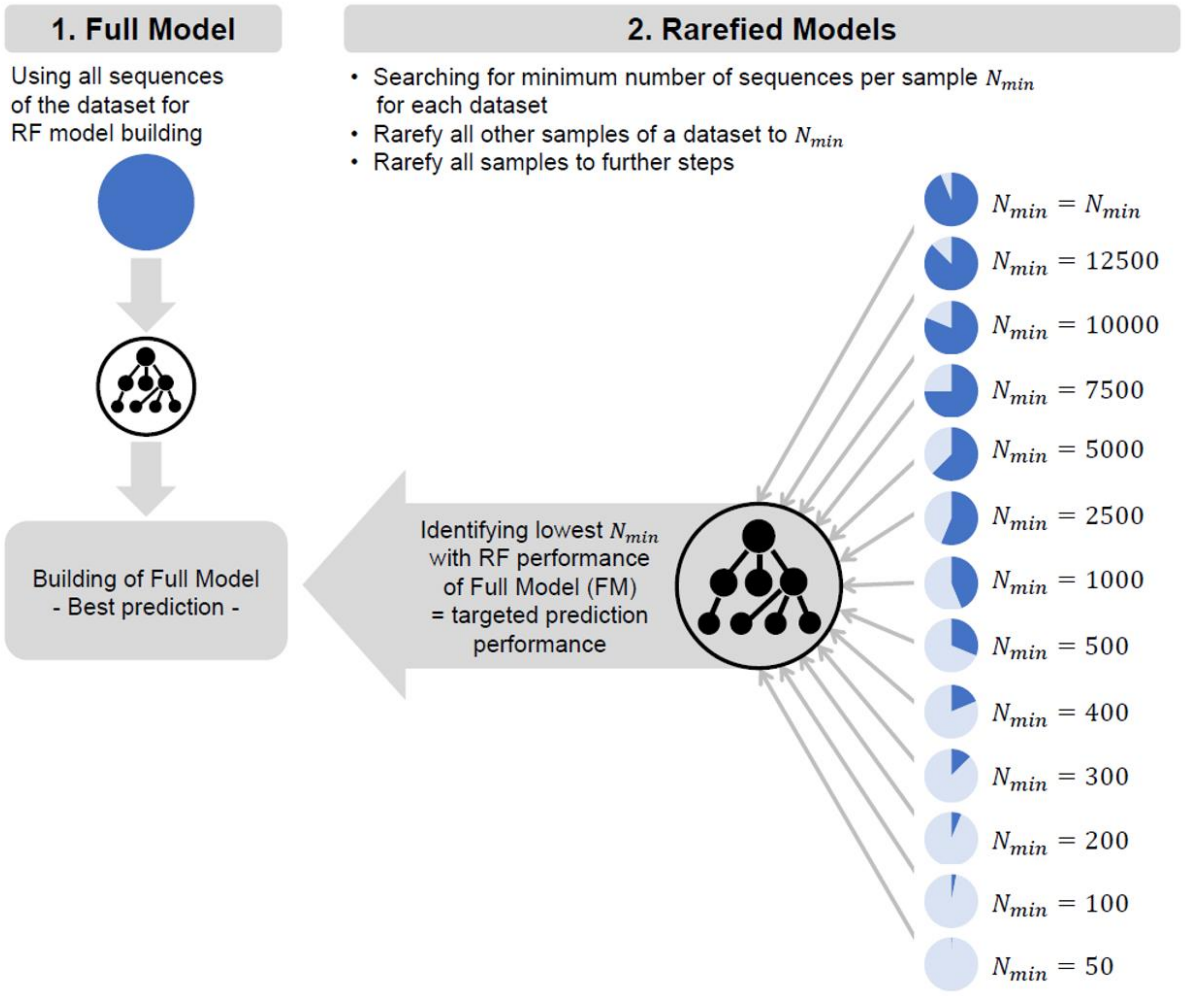


Fig. 1. *Workflow. Using the full ASVs-to-sample matrix from each dataset analyzed in this study, we build RF models for each of these datasets. The best model of each dataset (=full model) was then used a reference (benchmark) to assess to which degree sequences can be removed from each dataset (=downsampling) without losing prediction performance compared to the full model. The full model is thus the targeted RF prediction performance.*

Randomly chosen datasets did not show any OOB-E improvements when increasing the number of decision trees further. For each dataset we repeated the calculation of the FM was repeated several times: for the first FM, the *mtry* value was set to default as recommended. Depending on this default value, six further models were calculated, each with the default *mtry* value plus/minus one, plus/minus two and plus/minus three. Each of these analyses were repeated two times using different base trees to secure the prediction capacity of the full model. The R package *caret* [41] was used to infer kappa statistics for

each RF model. In the second step we then constructed RF models with the rarefied (downsampled) datasets to ask the following question: what minimum number of sequences within a sample is required to obtain an RF prediction performance which is of the same or similar quality as the prediction obtained from the full dataset (FM analysis). From here on, we refer to the prediction accuracy obtained from the full dataset (FM) as "targeted RF prediction performance". From each of the four full datasets, we created 13 rarefied datasets (52 datasets in total). With exception of the BallWa dataset, the number of sequences per sample in the first rarefied dataset was equal to the number of sequences obtained for the sample with the lowest sequence coverage. This was 15,177 for the ScoSa dataset, 16,501 for the BasCo dataset, and 15,048 for the NorSa dataset. Because the lowest number of sequences per sample within the BallWa dataset was nearly twice as high as for the other three datasets, we have set the number of this first downsampled BallWa model to 15,000 sequences to enable more solid comparisons with the other three datasets. Downsampling for the following 12 models in each of the four datasets employed 12,500, 10,000, 7500, 5000, 2500, 1000, 500,400, 300, 200, 100 and 50 sequences. For each of the 52 downsampled datasets, RF analyses were conducted as described above for the full model. A schematic graphic of this study design is shown in *Fig. 1*. Additionally, exemplary information about model construction and downsampling can be found in *Supplementary File 2*[*].

3. Results

3.1. Sequence data overview and rarefaction

The number of merged high-quality bacterial amplicons and obtained ASVs for the individual datasets were as follows (amplicons/ASVs): ScoSa 2,653,449/28,151, BasCo 2,310,195/63,943, BallWa 5,024,841/49,843, and NorSa 2,472,237/89,883. The number of ASVs with at least 50 sequence reads, which were used for downstream RF analyses were 3039 (ScoSa), 8406 (BasCo), 6318 (BallWa) and 6012 (NorSa). At their maximum sampling depth, nearly all samples were approaching sample saturation (*Fig. 2a–d*). Compared to the Chao1 estimator (=100% ASV coverage), the full sequence datasets reached a coverage of 99% (ScoSa), 98% (BasCo), 85% (BallWa) and 67% (NorSa) (*Fig. 3a–d*). The decrease of coverage (saturation) with downsampling for subsequent RF

[*] All supplementary files are additionally available at the appendix of this dissertation

analyses was notably different for the four datasets. As an example, when downsampling each dataset to 5000 reads, the coverage was 59% for the ScoSa dataset, 49% for the BasCo dataset, 34% for theBallWa dataset and 31% for the NorSa dataset. At the lowest sequence number (n sequences = 50), all samples were severely undersampled with saturation ranging between 4.2% (ScoSa,*Figs. 2a* and *3a*) and 1.8% (NorSa, *Figs. 2d* and *3d*) compared to the full community ASV richness as estimated by Chao1.

3.2. Random forest predictions of full and downsampled datasets

ScoSa (*Fig. 4*): For the ScoSa dataset we predicted the distance of samples from the salmon farm and the salmon production phase based on the V3-V4 metabarcodes of the benthic bacterial communities. To predict the distance from the salmon farm (*Fig. 4a*, prediction categories: cage edge, allowable zone of effect, reference), the RF model, which was trained on the full dataset, achieved a mean prediction accuracy of 89.2% (mean out-of-bag error: 10.8%) at a kappa > 0.8. When downsampling, kappa remained above this threshold for "almost perfect agreement" down to5000 sequences per sample. At this sampling size, the mean prediction accuracy was still 82.2%. Thus, a minimum of 5000 sequences within each sample was required to achieve the targeted RF prediction performance (FM-based prediction accuracy as reference). At sampling sizes ranging between 2500 and 300 sequences per sample, RF could still predict the distance from the salmon farm with a minimal precision of 71.3% (model_300) at a kappa ranging between 0.6 and 0.8 (moderate agreement). Poor agreement was obtained when less than 300 sequences per sample were used for the RF model.

When predicting the salmon production phase in which a specific sample was taken (*Fig.4b*, prediction categories: preproduction phase, early production phase, late production phase), prediction accuracy was 92.6% for the full dataset at a kappa of 0.88. When downsampling the full ScoSa dataset, 50 sequences per sample emerged as sufficient to maintain a prediction accuracy as high as 89.5% at an almost perfect agreement (kappa > 0.8 (0.82)) between the predicted and the actual salmon production phase of a sample. Thus, as few as 50 sequences within each sample were sufficient to achieve the targeted RF prediction performance.

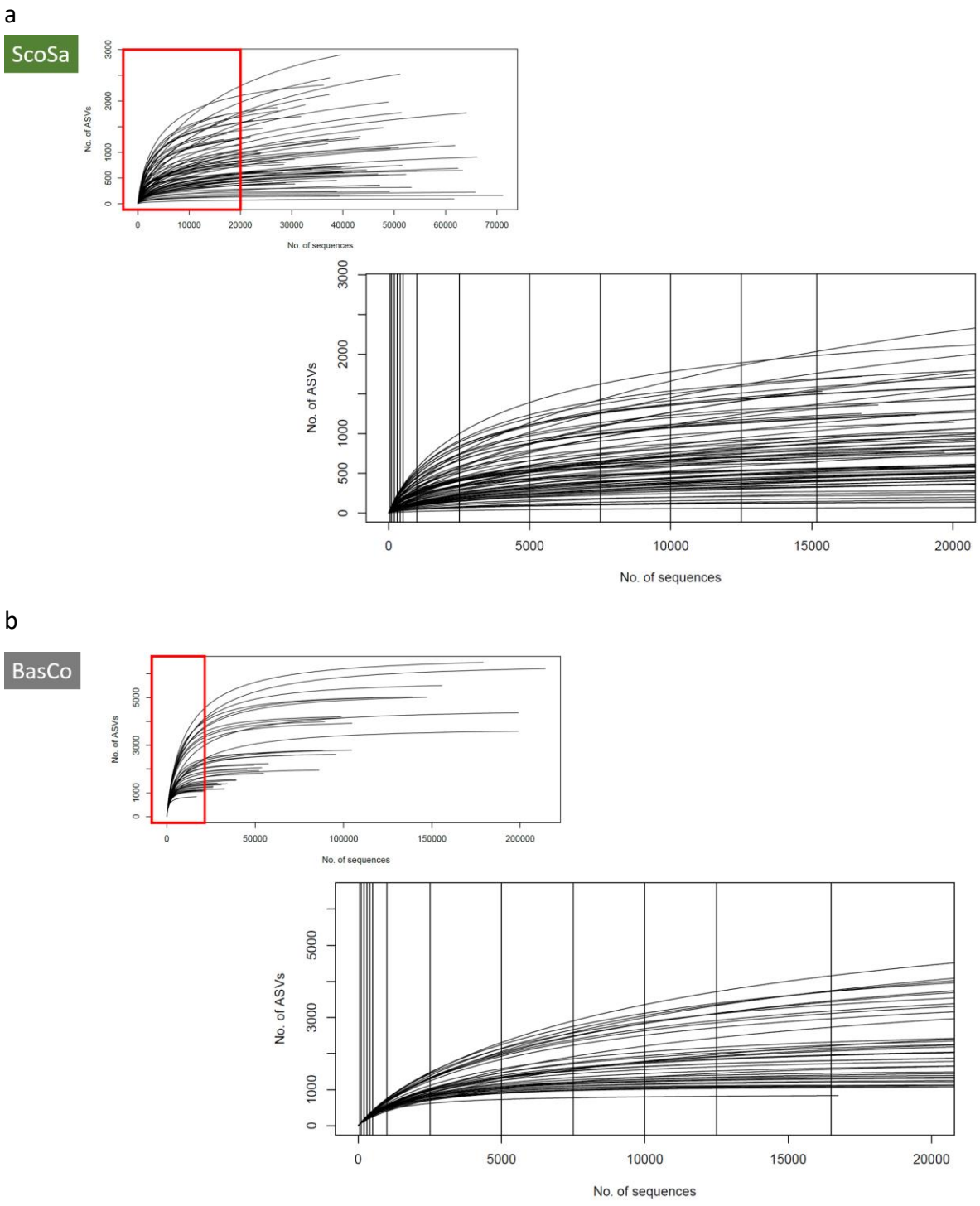
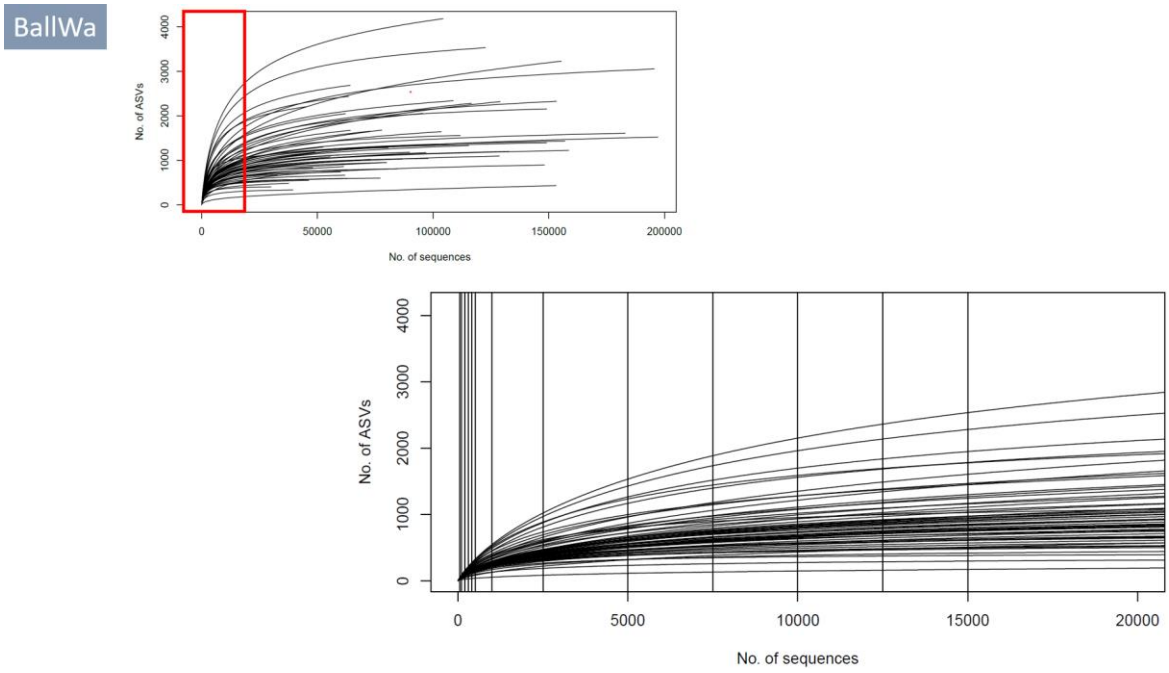


Fig. 2. *Rarefaction curves (sampling saturation profiles) of the datasets used in this study. (a) ScoSa (Scottish salmon farm sediment samples) dataset; (b) BasCo (marine sediment samples from the Basque coast); (c) BallWa (ballast water samples); (d) NorSa (Norwegian salmon farm sediment samples). The upper smaller graphics shows the full data of each dataset, based on which the full model used to define the ''targeted RF prediction performance" was build (see Fig. 1). The red square is the excerpt of the full dataset that is displayed in the lower graphics. In the lower graphics, sampling sizes used for downsampling (see Fig. 1) are marked with vertical lines. This visualizes the level of sample saturation (ASV coverage) at each downsampling size for each of the four datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)*

c



d

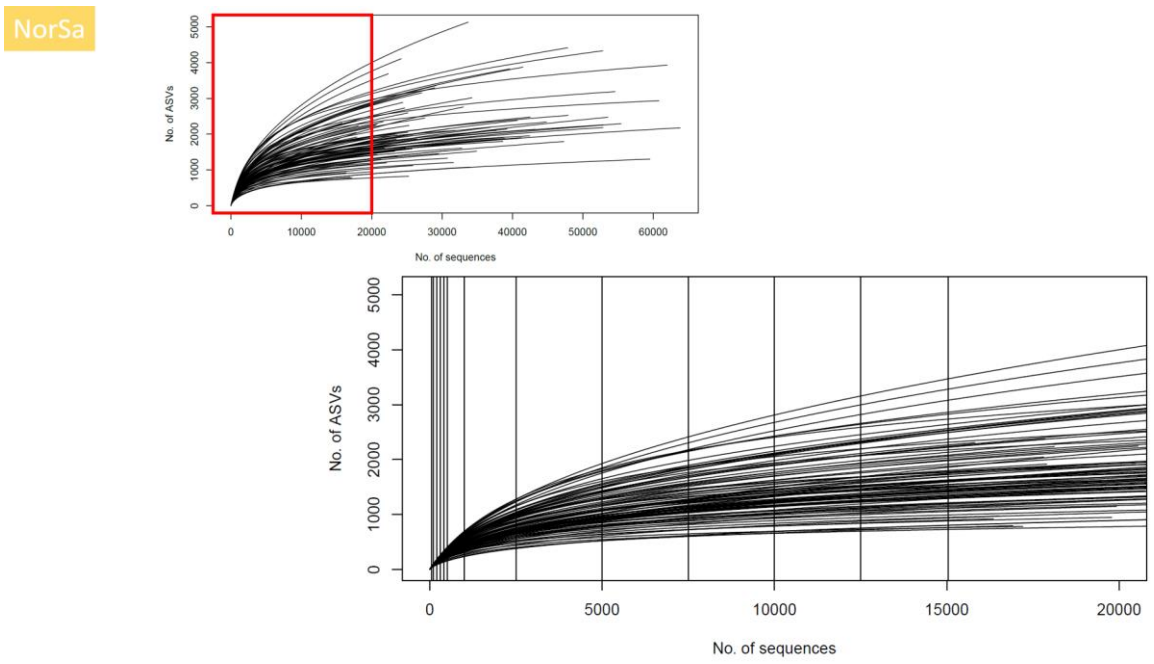


Fig. 2 *(continued)*

BasCo (*Fig. 5*): For the BasCo dataset we predicted the biotic index microgAMBI, based on which the ecological quality (EQ) of a sample was inferred (prediction categories: high, good, moderate, or poor ecological quality). Even when the full dataset was used to train the RF model, the mean precision of prediction was only 78.3% at moderate agreement (kappa = 0.71) between reference EQ values and predicted EQ values. To

achieve the targeted RF prediction performance of the full model (based on 214,250 sequences per sample) a minimum of 5000 sequences per sample was sufficient. Kappa values indicated moderate agreement down to 500 sequences per sample. Compared to the full dataset, the precision decreased, however, for 12.5% (OOB error at model_500: 31.5%).

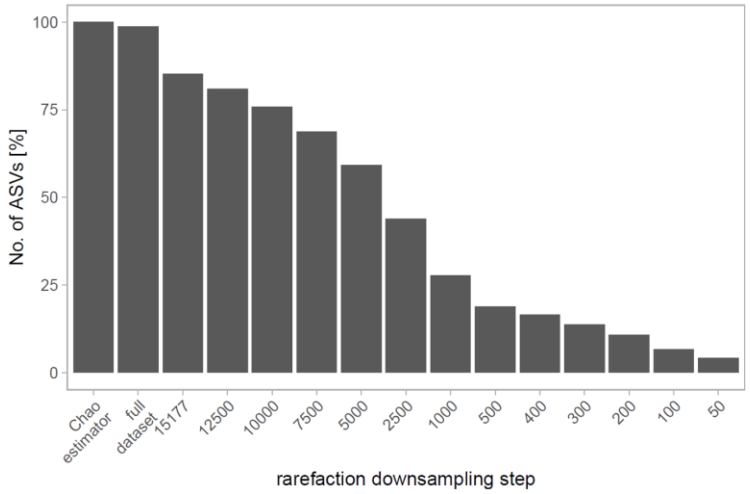
BallWa (*Fig. 6*): For the BallWa dataset we predicted the origin of ballast water samples using the bacterial V3-V4 16S rDNA markers detected in each of the 59 individual samples (categories: China, Singapore, USA). Despite this dataset had the highest number of sequences per sample of all four datasets tested, kappa statistics returned only a moderate agreement between the actual origin of a ballast water sample and the predicted origin even for the full dataset. Mean precision for origin prediction for the full dataset was 84.6% (kappa = 0.74). In the downsampled datasets, at least 2500 sequences per sample were required to match the performance category of the full dataset (targeted RF prediction performance). At this number of sequences within each sample, the mean prediction accuracy remained in the same order of magnitude compared to the full dataset (82.2% for model_2500) at moderate agreement (kappa = 0.70). Down to 300 sequences per sample, kappa remained in this category with a mean prediction accuracy of 75.3%. Only when less than 300 sequences per sample were used, agreement between observed and predicted sample origin was poor with OOB errors exceeding 25%.

NorSa (*Fig. 7*): For the NorSa dataset, we used the bacterial V3- V4 16S rDNA metabarcodes obtained from each of the 95 individual samples to predict two variables. First, the geographic location of the salmon farming site (*Fig. 7a*, predicted categories: Aukrasanden, Beitveitness, Bjornsvik, Nedre Kvarv and Storvika). Second, the biotic index AMBI (and resulting EQ category) for each sample, which was originally obtained from macroinvertebrates during a routine compliance monitoring of these salmon farms (predicted categories: good, moderate, poor, or bad ecological quality). Despite this dataset had the lowest sample saturation (coverage) (*Fig. 3d*), prediction accuracy for EQ category was similarly high for the full dataset (92.6%) and for the dataset that included only 1000 sequences per sample (90.8%) (*Fig. 7b*). Kappa statistics revealed a high agreement between actual and predicted EQ category for each of the 95 samples for the full dataset and for the model_1000 datasets (0.88 and 0.84, respectively). At 500 and 400 sequences per sample, the mean precision of prediction dropped to 89.0% and 88.1% respectively, yet kappa values were still >0.8. Only when number of sequences decreased below 400 sequences per sample, kappa values decreased below this threshold.

Table 1 summarizes these results and provides a comparative overview of the average number of sequences in the full data of each dataset and the minimum sampling size at which RF predictions for the individual variables were still in the same (kappa and accuracy) category compared to the full dataset. This overview shows that in the "worst case scenario" (ScoSa – distance from cage prediction, *Fig. 4a*), as few as 5000 sequences per sample were required to achieve RF prediction accuracies as good as for the corresponding full dataset (with 37,642 sequences per sample on average).

a

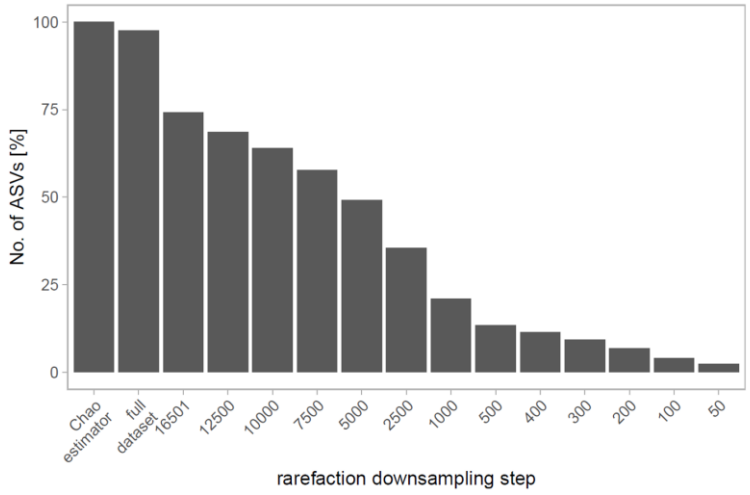
ScoSa



Step	Percent
Chao	100
full	98.76
15177	85.24
12500	81.01
10000	75.89
7500	68.82
5000	59.19
2500	43.92
1000	27.73
500	18.84
400	16.46
300	13.79
200	10.73
100	6.67
50	4.21

b

BasCo

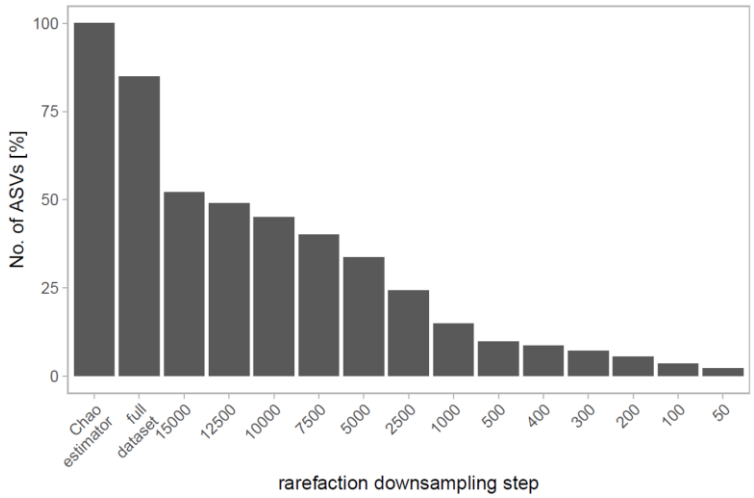


Step	Percent
Chao	100
full	97.67
16501	74.22
12500	68.6
10000	63.94
7500	57.7
5000	49.16
2500	35.47
1000	21
500	13.34
400	11.47
300	9.33
200	6.87
100	4.03
50	2.28

Fig. 3*. Relative proportion of remaining ASVs at individual downsampling steps for the ScoSa dataset (a), the BasCo dataset (b), the BallWa dataset (c) and the NorSa dataset (d). The Chao1 estimator was used to infer the maximum number of ASVs from each dataset, which is the number of ASVs that could be detected in theory, when all ASVs were sampled in each dataset. This number was set as 100%. ''Full" refers to the actually sampled ASVs in all datasets and indicates the discrepancy between the actual ASVs in a dataset and the expected number of ASVs if they were sampled to completion (=Chao1 value).*

c

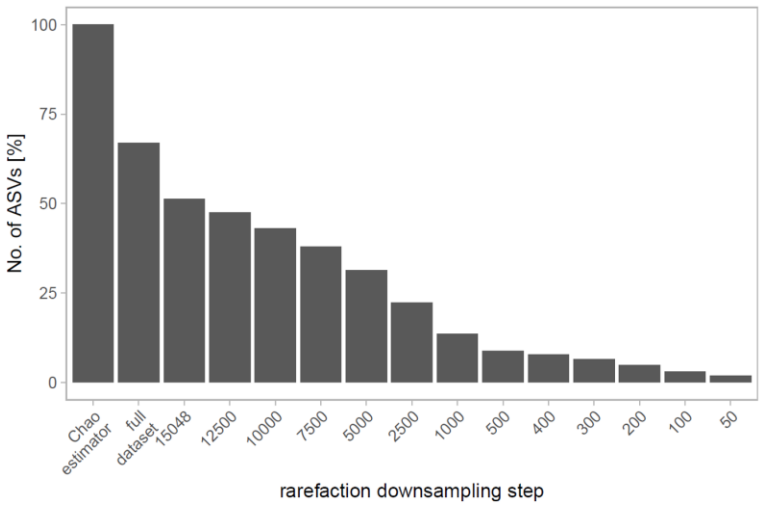
BallWa



Step	Percent
Chao	100
full	84.95
15000	52.16
12500	48.97
10000	45.02
7500	40.14
5000	33.68
2500	24.25
1000	14.89
500	9.83
400	8.6
300	7.09
200	5.52
100	3.46
50	2.11

d

NorSa



Step	Percent
Chao	100
full	66.87
15048	51.36
12500	47.44
10000	43.06
7500	37.99
5000	31.4
2500	22.34
1000	13.48
500	8.84
400	7.7
300	6.4
200	4.88
100	3
50	1.83

Fig. 3 *(continued)*

Table 1. *Summary of RF prediction results. The table shows the lower boundary (sequence numbers) at which an RF prediction performance was achieved in downsampled datasets that matched the prediction performance of the respective full dataset (=targeted RF prediction performance).*

Data set	n samples*	Averaged n reads per sample in full dataset	RF reference label for prediction	n prediction classes	Min n sequences required for targeted RF prediction performance**
ScoSa	76	37,642	Station (distance)	3	5000
			Salmon production phase	3	50
BasCo	39	76,259	microgAMBI	4	5000
BallWa	59	91,598	Country of origin	3	2500
NorSa	95	31,057	Aquafarm location	5	1000
			AMBI	4	1000

**Numbers refer to sample numbers which were used in this study. These are samples which had at least 15,000 sequences, which was the minimum number that we chose as highest threshold for downsampled datasets.*

***Full datasets as reference.*

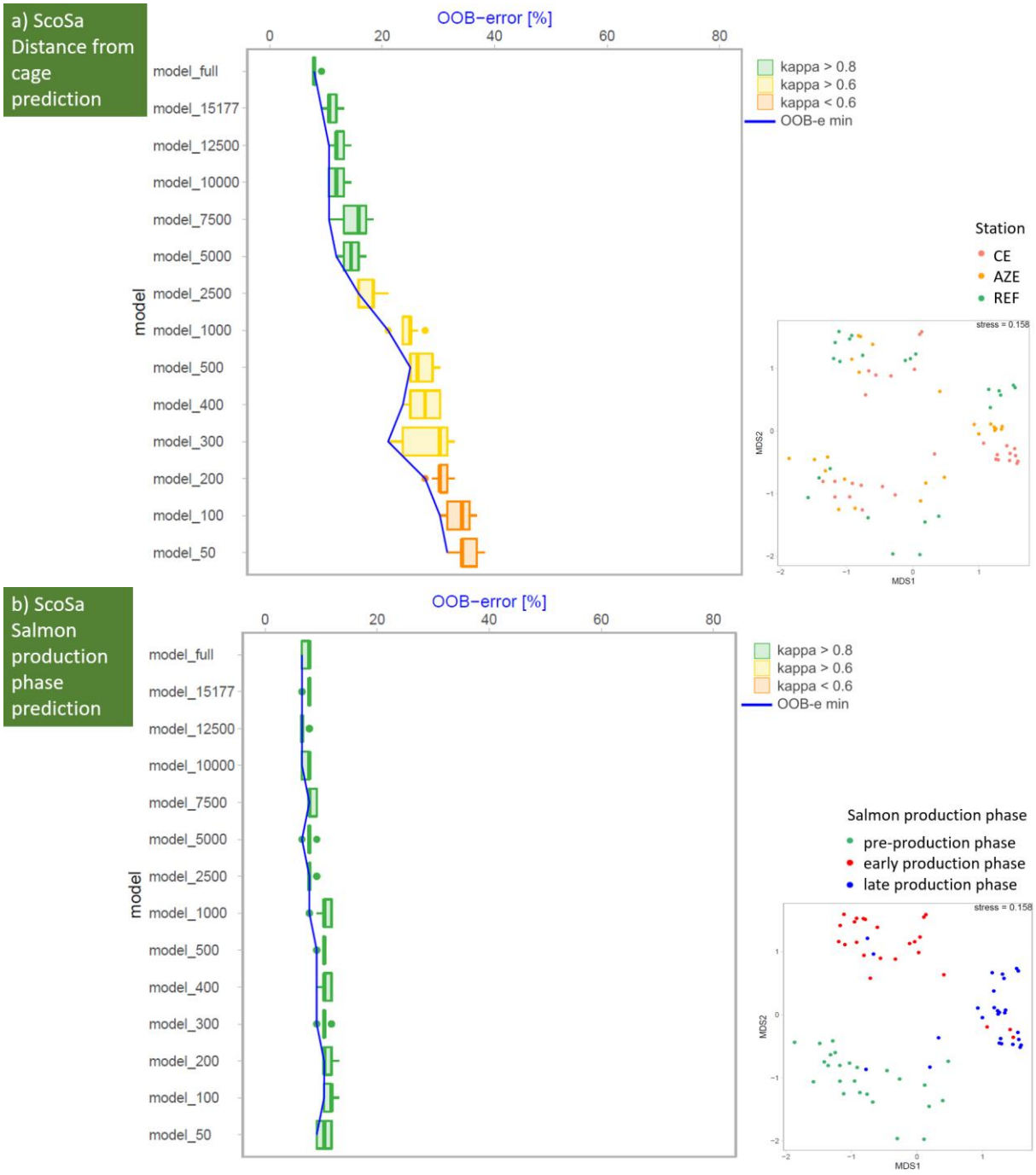


Fig. 4. *Change of OOB (out of bag error) and kappa value with decreasing dataset size (number of sequences) used for RF prediction for the ScoSa dataset to predict distance from the salmon cage site (a), and the salmon production phase (b) based on benthic bacterial community composition. The boxplot of each downsampled dataset consists of RF prediction results obtained from 21 models. Each boxplot shows median, quartiles (25%–75%), min and max values as well as outliers. Kappa values of > 0.8 (marked in green) indicate ''perfect agreement" between observed and predicted classifications (distance from salmon cages for the ScoSa distance dataset in Fig. 4a and salmon production phase in Fig. 4b). Kappa values marked in red (<0.6) indicate poor agreement. In case of distance prediction (Fig. 4a), perfect agreements can be achieved when the full dataset is downsampled to 5000 sequences. In case of salmon production phase predictions as few as 50 sequences still allow for a perfect prediction accuracy compared to the full model (Fig. 4b). Nonmetric multidimensional scaling (NMDS) plots in the lower right corner show the clustering of all individual samples of the full ASV dataset, colored by the specific prediction classes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.*

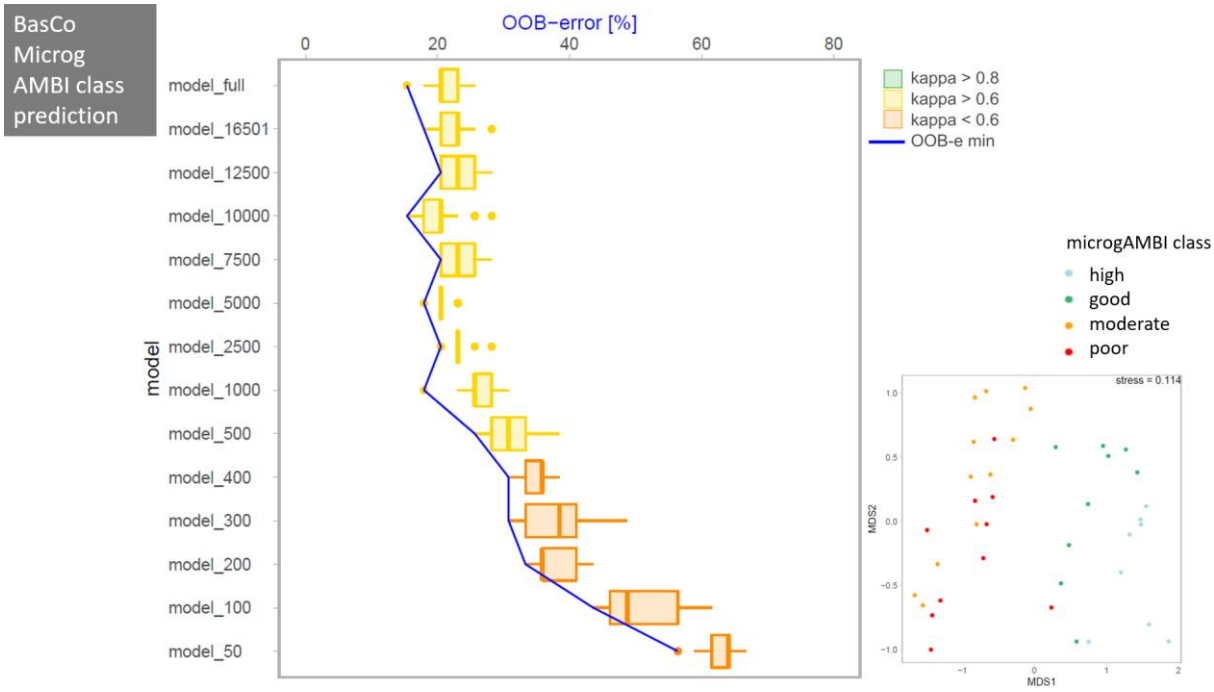


Fig. 5. *Change of OOB (out of bag error) and kappa value with decreasing dataset size (number of sequences) used for RF prediction for the BasCo to predict microgAMBI index based on benthic bacterial community composition. For further details see legend of Fig. 4.*

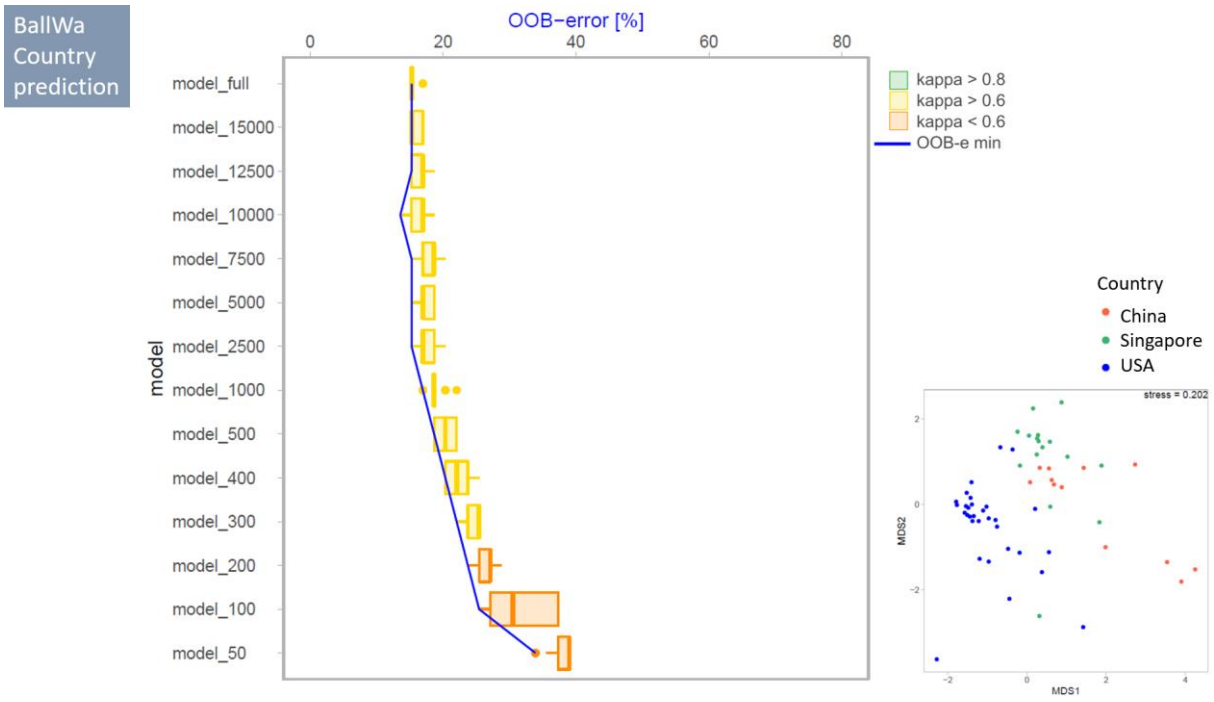


Fig. 6. *Change of OOB (out of bag error) and kappa value with decreasing dataset size (number of sequences) used for RF prediction for the BallWa dataset to predict country of origin for ballast water samples based on bacterial community composition. For further details see legend of Fig. 4.*

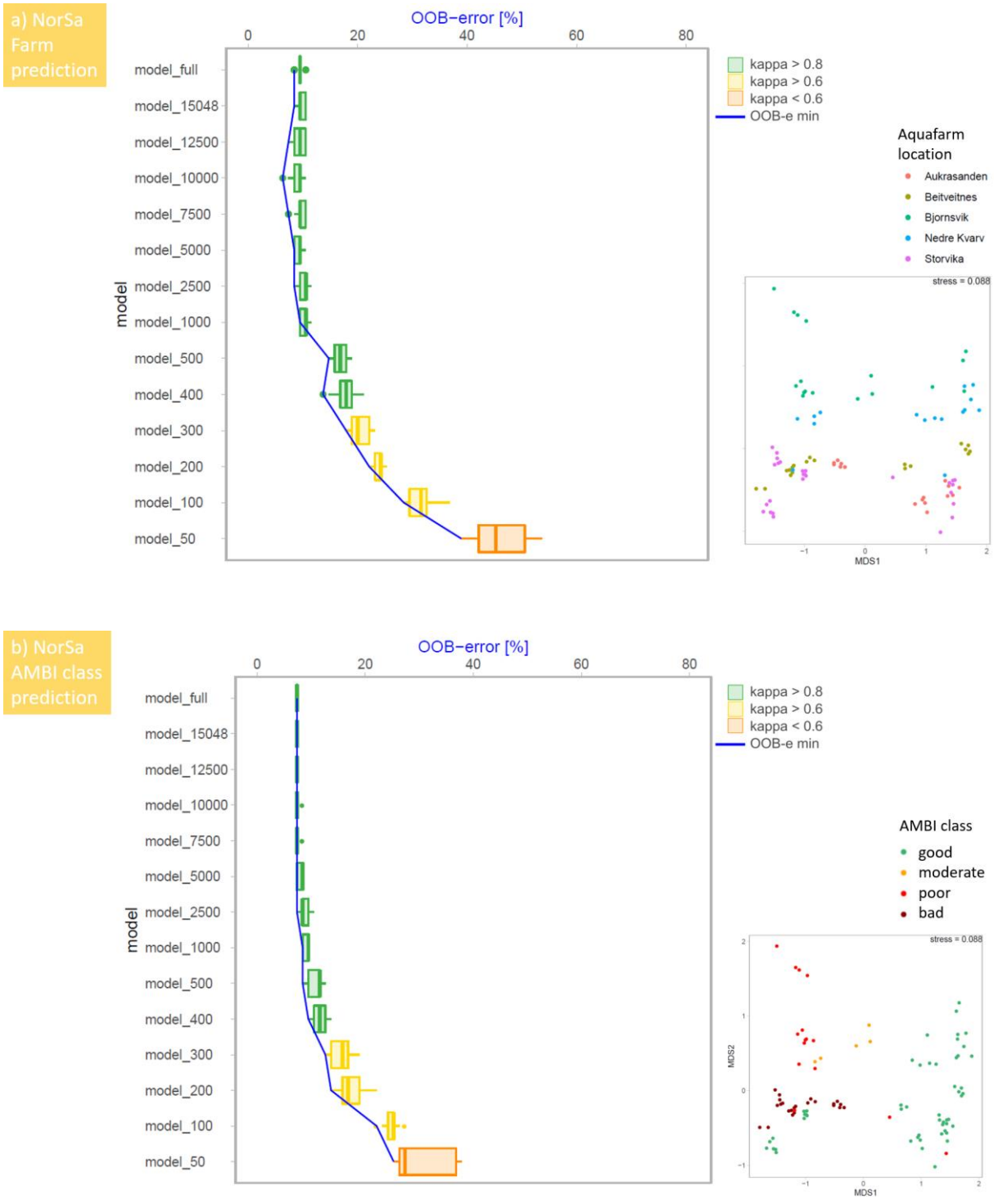


Fig. 7. *Change of OOB (out of bag error) and kappa value with decreasing dataset size (number of sequences) used for RF prediction for the NorSa dataset to predict the aquafarm location (a) and the AMBI ecological quality index based on benthic bacterial community composition. For further details see legend of Fig. 4.*

4. Discussion

Advances in various high-throughput sequencing technologies have opened up new opportunities to exploit massive sequence datasets from environmental samples for developing prognostic and predictive markers for biomonitoring. RF learning algorithms are increasingly used in microbial ecology for classification problems [31,42]. While several studies have compared the power of RF classifications with other classifiers or different model parameters within RF, no environmental sequencing study has evaluated the relationship between training data volume and RF accuracy. Therefore, in the specific problem addressed in our study, we are asking, how many amplicon sequences of marine bacterial communities are required per sample to achieve a desired level of RF prediction performance. A common sense logic could be ‘‘the more data, the better”. However, the larger the (amplicon) datasets, the more expensive are the sequencing costs and the longer is the computation time needed to train the model. In addition, too much data (=too many features) may have a tendency to overfit (lower variation in individual trees resulting in a ‘‘less random” forest). This is prevented using smaller datasets (increase of variation in the individual trees within the forest, decorrelation) [40,43]. However, removing too many features can impair model performance. Thus, finding the ideal number of sequences per sample is an important step towards an optimal sampling design for environmental sequencing studies which use RF learning algorithms for sample classification. Our results from four test datasets (V3-V4 16S rDNA gene amplicon surveys of marine bacterial communities from different environments) showed that there is no general answer to this question. Depending on the dataset and the question asked, we identified subsets as little as 50 sequences as powerful as the full dataset of 15,000 sequences (ScoSa salmon production phase prediction) and in other cases the subsets required a minimum of 5000 sequences (ScoSa distance predictions, BasCo microgAMBI predictions) to achieve as accurate predictions as the full dataset.

The discrepancies could be explained by the complexity of the problem addressed in this study. Statistical heuristic to determine a suitable number of sequences per sample for classification problems are a function of the number of classes in the classification, the distribution of classes across the complete dataset (=balance), the total number of samples, and the ability of the training data to adequately characterize the classes being mapped [44]. Also, the tuning parameters when building a random forest could play a role in this context [45].

Here, we can largely exclude some of these factors as explanations for the observed discrepancies of 50 versus 5000 sequences per sample as minimal data size. The number of classes is relatively even among the test datasets of this study (min 3, max 5). In addition, the maximum discrepancy in the minimum dataset was observed within the same number of classes (n = 3, ScoSa salmon production phase prediction and ScoSa distance prediction). And, finally, the dataset with the highest number of classes (NorSa aquafarm location prediction) required fewer sequences per sample (n = 1000) for a full dataset-like classification than a dataset with the minimum number of classes (BallWa, min n sequences = 2500). Likewise, the total number of samples in a dataset is not an obvious decisive factor that determines the minimum number of sequences required for the targeted level of RF performance. The highest number of samples (n = 95) was available for the NorSa dataset. However, the minimum number of sequences required for the targeted level of RF performance was notably higher compared to a dataset with only 76 samples (ScoSa salmon production phase prediction). Furthermore, in the ScoSa dataset, the same number of samples (n = 76) required only 50 sequences minimum to predict salmon production phase (3 classes), while 5000 min sequences were needed to predict distance to the salmon farming site (also 3 classes).

We also consider it unlikely that the tuning parameters when building a random forest could play a role for the obtained results. Relevant tuning parameters are the choice of the base tree, the number of trees in the forest, size of the leaf nodes and the rate of data subsampling [45]. We have built 6000 trees, which outnumbers the number of trees used in comparable analyses. For example, for the original analysis of the NorSa dataset, 300 trees were used [9]. 5000 trees were used for the original BallWa dataset analysis [31], and Smith et al. [28] used 1000 trees for RF analyses to classify unpolluted sites from those contaminated with uranium, nitrate, or oil using V4 16S rRNA gene amplicons of bacterial communities. Also, we have constructed ~300 models with varying tuning parameters for each dataset. This strategy should minimize the effect of tuning parameters in the identification of the smallest number of sequences needed for the targeted classification performance.

In case of training data class imbalance, classifications may favor the classes that represent the largest proportion in the training samples (majority classes) [46]. Thus, classes that are underrepresented in the training data may thus be difficult to classify correctly. It therefore seems reasonable to assume that the smaller the number of sequences become within under-represented classes, the more erroneous the classification

performance. To test whether this helps to explain the different minimum sequence numbers that achieve targeted classification in the datasets analyzed in this study, we analyzed the (in)equality of class frequency distributions using the Gini coefficient (not to be confused with Gini impurity) and compared with the minimum number of sequences in a sample that is required for the targeted prediction performance. The results are presented in *Supplementary File 3*[*]. They clearly show that a data class imbalance is irrelevant as factor to determine the required minimum of sequences within a sample for targeted prediction performance. Previous analyses showed that among several commonly used classifiers for ‘‘omics” data RF is the optimal choice when feature distributions are skewed and when class distributions are unbalanced [47].

We consider the ability of the training data to adequately characterize the classes being mapped as the best proxy to assess the number of sequences in a sample that is required for the targeted classification performance. Especially from RF classifications using image analyses, it is well known that classifications are more accurate, when classes are mutually exclusive and have hard, welldefined boundaries [44,48]. A possibility to visualize how well the boundaries between classes are established based on the features observed for the samples within each of the class is ordination plots such as multidimensional scaling (MDS). The boundaries of classes (clusters in ordination plots) in such plots can be calculated and visualized using confidence intervals [49]. The smaller the overlap of confidence intervals of individual class-specific clusters, the higher is the probability that targeted RF classification accuracy can be achieved with a low number of sequences in a sample. *Supplementary File 4*[*] shows an example for the ScoSa dataset. While salmon production phase clusters are well separated with only little overlap of cluster-specific confidence intervals, distance clusters are less clearly separated. This corroborates well with our finding that in the former case only 50 sequences per sample are sufficient for the targeted RF classification accuracy, while in the latter case at least 5000 sequences are required. This observation confirms our assumption that datasets with well-defined boundaries of classes require fewer sequences within each sample in the training dataset to achieve the targeted classification performance. Subtle differences among the classes, such as environmental gradients, season or geographic origin will require more sequences within a sample. But even with very large sequence datasets, RF predictions may not be satisfying, if these sequences do not succeed to better define the boundaries of the classes. This was

[*] All supplementary files are additionally available at the appendix of this dissertation

the case for the BasCo dataset analyzed in this study. Even when the full dataset (n = 76,259 sequences, sampled to near saturation) was used to train the RF model, the mean precision of prediction was only 78.3% at moderate agreement (kappa = 0.71) between reference EQ values and predicted EQ values.

Well-defined, hard boundaries between classes with little to no gradual transitions or edge overlap occur when the features within a class (here: bacterial V3-V4 16S rRNA gene amplicons) are as specific as possible for each individual class [44]. One way to visualize the class specificity of features is Venn diagrams. Not surprisingly, we found that the specificity of ASVs was notably higher for the ScoSa salmon production phase dataset (n min sequences required = 50) compared to the ScoSa distance dataset (n min sequences required = 5000) (Supplementary File 5*). In the latter, 35% of all sequences were common to all three distance classes, whereas only 11% of all sequences were common to all three salmon production phases.

The coefficient of variation (CV) of each feature in a dataset can be interrogated as a measure to assess the ability of features (here: ASVs) to discriminate prediction classes. The CV measures the standard deviation of an individual feature across the individual prediction classes relative to the CV group mean (‘‘group” describes the prediction target) [50]. The general expectation is as follows: the higher the CV of an individual feature (ASV), the more specific is its occurrence in individual prediction classes (=uneven distribution of an ASV, which leads to a higher CV). In conclusion, the more features with a higher CV are included in a dataset, the higher is the likelihood to obtain a more accurate RF prediction with only a subset of the features included in a dataset. This is because the subset with fewer features still includes sufficient information in each feature for reliable RF predictions. To test this logic, we have exemplary calculated the CV for the ScoSa salmon production phase dataset and for the ScoSa station dataset. Indeed, the ScoSa salmon production phase dataset, which has still reliable RF prediction accuracies with as few as 50 features, has a notably higher CV density distribution compared to the ScoSa station dataset, which requires at least 5000 features to achieve the targeted RF prediction performance. For CV kernel density plots and more detailed information and analyses we refer to *Supplementary File 6**.

* All supplementary files are additionally available at the appendix of this dissertation

5. Conclusions

In conclusion of our study, even for the ‘‘worst case scenario” when classes had no hard boundaries but substantial gradual transition and edge overlap (*Supplementary File 4*[*]) we identified 5000 sequences as a threshold for the number of sequences within a sample, beyond which no substantial improvements are achieved in RF classification performance. This could be a rule of thumb guiding future studies using taxonomic metabarcodes of marine microbial communities for RF classification in ecological studies. Our examples included classifications of environmental quality and stressor impact, as well as spatial and temporal scaling, all of which are central topics in microbial ecology. Considering that environmental DNA metabarcoding studies of marine microbial communities usually acquire substantially higher number of sequences [9,17,31,36,37,51] without prior adaptations of sequencing depths to the research questions addressed, our study may guide future sampling designs in RF classification based on microbial amplicon sequences to save financial and computational resources, while avoiding possible bias of overfitting and reducing noise due to too large datasets. Also, our study has identified parameters that are helpful to assess whether fewer or more sequences are needed as features to distinguish prediction classes. Both, feature specificities as well as multidimensional scaling plots allow for assessments of the minimum sequencing depth required for an RF performance that does not improve substantially with notably larger sequencing efforts. We therefore recommend a small-scale pilot study before designing large-scale experiments to assess the general tendency of features (sequences within each sample) to distinguish prediction classes.

CRediT authorship contribution statement

Verena Dully: Writing - review & editing. Thomas A. Wilding: Writing - review & editing. Timo Mühlhaus: Writing - review & editing. Thorsten Stoeck: Supervision, Conceptualization, Funding acquisition, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by grants from the Deutsche Forschungsgemeinschaft (DFG, grant STO414/15-1 and STO414/15-2) and from the ASSEMBLE plus program, both awarded to TS. The authors are very grateful for the support of Scottish Sea Farms Limited, which granted access to the Scottish salmon farm. We also thank Gail Twigg (SAMS) and the crew of the R/V Seol Mara and for technical support during sampling. Furthermore, we thank two anonymous reviewers for their valuable comments on our manuscript.

Appendix A.

Supplementary data Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.04.005*.

* All supplementary files are additionally available at the appendix of this dissertation

References

[1] Miller, M. L., & Auyong, J. (1991). Coastal zone tourism: A potent force affecting environment and society. *Marine Policy, 15*(2), 75-99. doi:10.1016/0308-597X(91)90008-Y

[2] Olenin, S., Elliott, M., Bysveen, I., Culverhouse, P. F., Daunys, D., Dubelaar, G. B., et al. (2011). Recommendations on methods for the detection and control of biological pollution in marine coastal waters. *Marine Pollution Bulletin, 62*(12), 2598-2604. doi:10.1016/j.marpolbul.2011.08.011

[3] Rosenberg, R. (1985). Eutrophication - The future marine coastal nuisance? *Marine Pollution Bulletin, 16*(6), 227-231. doi:10.1016/0025-326X(85)90505-3

[4] Shahidul Islam, M., & Tanaka, M. (2004). Impacts of pollution on coastal and marine ecosystems including coastal and marine fisheries and approach for management: a review and synthesis. *Marine Pollution Bulletin, 48*(7), 624-649. doi:10.1016/j.marpolbul.2003.12.004

[5] IPCC (2007). Intergovernmental Panel on Climate Change. Climate Change 2007: Fourth Assessment Report. The Physical Science Basis, Summary for Policymakers. . *Retrieved on 25.01.2021 from https://previa.uclm.es/area/amf/antoine/energias/-Ipcc_anotado.pdf*

[6] Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., et al. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment, 70*, 186-203. doi:10.1016/j.atmosenv.2012.11.060

[7] Hoegh-Guldberg, O., & Bruno, J. F. (2010). The impact of climate change on the world's marine ecosystems. *Science, 328*(5985), 1523-1528. doi:10.1126/science.1189930

[8] MSFD (2008). Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the Field of marine environmental policy. *Retrieved on 25.01.2021 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008L0056.*

[9] Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources, 18*(6), 1381-1391. doi:10.1111/1755-0998.12926

[10] Danovaro, R., Carugati, L., Berzano, M., Cahill, A. E., Carvalho, S., Chenuil, A., et al. (2016). Implementing and Innovating Marine Monitoring Approaches for Assessing Marine Environmental Status. *Frontiers in Marine Science, 3*, 213. doi:10.3389/fmars.2016.00213

[11] Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (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

[12] Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., Irigoien, X., Danovaro, R., & Rodríguez-Ezpeleta, N. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Marine Pollution Bulletin, 114*(2), 679-688. doi:10.1016/j.marpolbul.2016.10.050

[13] Borja, A. (2018). Testing the efficiency of a bacterial community-based index (microgAMBI) to assess distinct impact sources in six locations around the world. *Ecological Indicators, 85*, 594-602. doi:10.1016/j.ecolind.2017.11.018

[14] Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology, 27*(5), 387-397. doi:10.1016/j.tim.2018.10.012

[15] Forster, D., Filker, S., Kochems, R., Breiner, H.-W., Cordier, T., Pawlowski, J., & Stoeck, T. (2019). A Comparison of Different Ciliate Metabarcode Genes as Bioindicators for Environmental Impact Assessments of Salmon Aquaculture. *Journal of Eukaryotic Microbiology, 66*(2), 294-308. doi:10.1111/jeu.12670

[16] Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

[17] Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

[18] Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources, 14*(6), 1129-1140. doi:10.1111/1755-0998.12261

[19] Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., & Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: time for change. *European Journal of Protistology, 55*, 12-25. doi:10.1016/j.ejop.2016.02.003

[20] Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

[21] Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., & Pawlowski, J. (2018b). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecological Indicators, 85*, 153-164. doi:10.1016/j.ecolind.2017.10.041

[22] Verhoeven, J. T. P., Salvo, F., Knight, R., Hamoutene, D., & Dufour, S. C. (2018). Temporal Bacterial Surveillance of Salmon Aquaculture Sites Indicates a Long Lasting Benthic Impact With Minimal Recovery. *Frontiers in Microbiology, 9*, e03054. doi:10.3389/fmicb.2018.03054

[23] Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology, 30*, 2937-2958. doi:10.1111/mec.15472

[24] Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*, 363-374. doi:10.2307/2529786

[25] Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics, 99*(6), 323-329. doi:10.1016/j.ygeno.2012.04.003

[26] Deist, T. M., Dankers, F., Valdes, G., Wijsman, R., Hsu, I. C., Oberije, C., et al. (2018). Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Medical Physics, 45*(7), 3449-3459. doi:10.1002/mp.12967

[27] Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. *Geoderma, 170*, 70-79. doi:10.1016/j.geoderma.2011.10.010

[28] Smith, M., B., Rocha, A., M., Smillie, C., S., Olesen, S., W., Paradis, C., Wu, L., et al. (2015). Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. *mBio, 6*(3), e00326-00315. doi:10.1128/mBio.00326-15

[29] Dataset. AZTI. Sediment samples for bacterial diversity analysis. NCBI SRA, Accession Number: PRJNA322754, 2016. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA322754

[30] Dataset. Duke University. Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. NCBI SRA, Accession Number: PRJNA628526, 2020. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA628526

[31] Dataset. Gerhard WA, Gunsch CK. Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environ Int* 2019; 124:312

[32] Dataset. University of Kaiserslautern. Bacterial eDNA metabarcodes for Environmental monitoring. NCBI SRA, Accession Number: PRJNA417767, 2017. https://www. ncbi.nlm.nih.gov/bioproject/?term=PRJNA417767

[33] Dataset. University of Kaiserslautern. V3V4 Data Salmon Farm Scotland. NCBI SRA, Accession Number: PRJNA667346, 2020. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA667346.

[34] Herlemann, D. P. R., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., & Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal, 5*(10), 1571-1579. doi:10.1038/ismej.2011.41

[35] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

[36] Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., & Stoeck, T. (2021). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter-laboratory study. *Ecological Indicators*, *121*, e107049. doi:10.1016/j.ecolind.2020.107049

[37] Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology, 30*, 2969-2987. doi:10.1111/mec.15489

[38] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

[39] Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R news, 2(3)*, 18-22.

[40] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. In *The elements of statistical learning* (2nd ed., pp. 587-604). New York: Springer.

[41] Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., & Kenkel, B. (2020). R package 'caret' Classification and Regression Training. *https://github.com/topepo/caret/*

[42] Roguet, A., Eren, A. M., Newton, R. J., & McLellan, S. L. (2018). Fecal source identification using random forest. *Microbiome, 6*(1), 185. doi:10.1186/s40168-018-0568-3

[43] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

[44] Millard, K., & Richardson, M. (2015). On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sensing, 7*(7), 8489-8515. doi:10.3390/rs70708489

[45] Tang, C., Garreau, D., & Luxburg, U. (2018). When do random forests fail? Conference on Neural Information Processing Systems. *Retrieved on 26.01.2021 from https://www.researchgate.net/publication/328229072_When_do_random_-forest_fail*

[46] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. doi:10.1109/TKDE.2008.239

[47] Guo, F., Wang, G., Su, Z., Liang, H., Wang, W., Lin, F., & Liu, A. (2016). What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire, 25*(5), 505-519. doi:10.1071/WF15121

[48] Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment, 80*(1), 185-201. doi:10.1016/S0034-4257(01)00295-4

[49] Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2019). R Package 'vegan': Community Ecology Package. *Version 2.5-6*.

[50] Lepš, J., & Šmilauer, P. (2020). *Biostatistics with R: an introductory guide for field biologists*. Cambridge: Cambridge University Press.

[51] Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01

SYNTHESIS AND OUTLOOK

Retrospective

Environmental genomics, such as eDNA metabarcoding, harbors a tremendous potential for the implementation into regulatory frameworks for compliance monitoring. To protect the coastal environment and ensure ecosystem functioning, compliance monitoring is mandatory for operators of aquaculture installations and other anthropogenically induced pollution sources. This thesis mainly focused on aquaculture farms enriching the surrounding environment with organic carbon, which has a major impact on the natural community. For those marine coastal habitats, environmental monitoring based on bacterial eDNA has recently emerged. The sensitivity of the bacterial community to environmental influences, such as organic enrichment, can be reflected in an eDNA dataset (Aylagas et al., 2021; Dowle et al., 2015; Frühe et al., 2020; Stoeck et al., 2018a). Therefore, it is recommended to target eDNA deriving from bacterial sequences, as those are powerful indicators of environmental health, outperforming traditional methods (Laroche et al., 2018). In contrast to macrofauna organisms traditionally used for monitoring, benthic bacteria mirror the ecosystem status in a timely manner as they have short generations times. Sampling and processing of eDNA is faster and easier than using traditional macrofauna indicators, therefore allowing timely interventions by farm operators (Pawlowski et al., 2018). Misidentifications of macrofauna organisms due to variation in life stages and cryptic diversity can be bypassed by eDNA metabarcoding. Additionally, microbial organisms that were previously undetectable via traditional methods can now be treated as potential bioindicators. The eDNA method can be suggested to regulators as a cheap alternative, as HTS costs, mandatory for this method, are decreasing more and more. Exemplary, it is estimated that the average cost to identify an operational species can be reduced from 32.54$ to 3.32$ using metabarcoding, mostly by circumventing costs of labor (Le et al., 2021).

This thesis emphasizes the remarkable potential of eDNA metabarcoding-based methods for determining the environmental status of various habitats in marine coastal environments. To implement the novel eDNA-based method into legislative determinations for routine monitoring, standardized processes examined for their robustness and reproducibility are required (Bourlat et al., 2013; Kelly et al., 2014; Pawlowski et al., 2018). The individual steps of the method must be standardized and recorded in protocols for SOPs (Goldberg et al., 2016; Helbing and Hobbs, 2019; Loeza-Quintana et al., 2020).

Outcome of the thesis

The calls for standardization of sample preservation during transport, PCR, sequencing, and bioinformatic evaluation of eDNA data have been answered by this work (Loeza-Quintana et al., 2020; Pawlowski et al., 2022; Schloss, 2018). This study thus provides a better understanding of crucial steps involved in evaluating a marine coastal environment, focusing on the detection of organic enrichment induced by aquafarm installations.

Reproducibility of the metabarcoding-based ecological interpretation is a precondition for the implementation of such eDNA-based methods into SOPs (Bowers et al., 2021; Hestetun et al., 2021a). As eDNA degradation and other undesirable alterations can bias downstream analysis and therefore the evaluation of the environment under surveillance, it is essential to preserve sediment samples during transport. In **Chapter I**, it was demonstrated that the environmental assessment of marine coastal ecosystems is robust among the tested strategies of sample preservation during transport. It was demonstrated that both commonly used preservation methods, either adding of preservation solution to samples, or the freezing of samples in a timely manner without addition of preservation solution, allow for the same conclusions regarding environmental quality based on bacterial eDNA. For both preservation strategies, the ecological disturbance gradient induced by salmon farm organic enrichment was equally well mirrored. The DNA extraction efficiencies, the alpha and beta diversity measures, and the bacterial community profiles were demonstrated to be highly congruent among preservation strategies. As treatment effects were marginal compared to sampling location effects, the feasibility of the eDNA-based method is ensured. Thus, it is recommended to implement both preservation methods into SOPs for bacterial eDNA-based monitoring of marine coastal sediments. This allows the operator to choose the method easiest to employ according to the logistical circumstances. The use of nucleic acid preservation solutions is recommended if freezing cannot be accomplished within a few hours after sampling or if the cost of refrigerated transport exceeds the cost of the nucleic acid preservation solution.

Furthermore, conclusions based on HQ data can only be drawn when robust methodologies are applied, representing the prerequisite for the method to be accepted for standardized protocols (Nicholson et al., 2020; Zinger et al., 2019). A crucial point in the creation of such SOPs is the reproducibility and robustness of the molecular method to be used regardless of the sample processing laboratory.

This means that when the same samples are analyzed following the same instructions, different laboratories must achieve the same ecological conclusions. **Chapter II** of this thesis confirmed the reproducibility of the eDNA-based method since independent sample processing among different laboratories lead to the same ecological interpretation regarding EQ. When samples were handed to two different laboratories, the identified indicators for EQ assessment, the alpha and beta diversity measures, and the taxonomic profiles were highly congruent. Additionally, SML was able to accurately predict the EQ for one laboratory based on data from the other laboratory, indicating a high congruency among the bacterial community turnover along the enrichment gradient. Therefore, the conducted eDNA metabarcoding method and the used data analysis protocol are demonstrated to be sufficiently robust for marine monitoring of aquacultural impacts and can be implemented in SOPs.

Throughout the first two chapters, it was highlighted that the bacterial community composition in marine coastal sediments exhibits a considerable small-scale natural variance. This variability has been reported several times and can be explained by microhabitat patchiness, induced by changes of various environmental impacts on the seafloor on a small area (Lejzerowicz et al., 2014; Parsons et al., 1977). To exclude these variations, it is suggested for future studies to take the samples in several replicates as it has already been proposed in the literature (Hestetun et al., 2021a; Keeley et al., 2018; Lanzén et al., 2020; Lejzerowicz et al., 2014). Nevertheless, it was emphasized in this work that microhabitat patchiness does not decisively affect the evaluation of an ecosystem along a pollution gradient, as it does not overshadow the effects of organic enrichment induced by the salmon farm (Dowle et al., 2015; Hornick and Buschmann, 2018; Keeley et al., 2018; Stoeck et al., 2018a).

Moreover, it was demonstrated that the taxonomic assignment of the used bacterial dataset was insufficient for monitoring purposes. The detection of potentially strong indicators on lower taxonomic levels such as species level was prohibited. Such biases regarding taxonomic assignments are introduced by incomplete databases, which is why ASVs should be preferred (Chariton et al., 2010; Lanzén et al., 2016; Lejzerowicz et al., 2015). This recommendation is emphasized in Chapter III, as indicator ASVs were identified, for which the taxonomically assigned equivalents were inconclusive regarding their ecological response to a distinct EQ. For the construction of a universally applicable monitoring system based on eDNA metabarcoding, it is thus recommended to use high-resolution ASVs rather than assigned taxa for indicator inference.

To make the eDNA method operational as a future monitoring tool, a transformation of ASV data into EQ assessments of the environment must be conducted. In **Chapter III**, it was demonstrated that both methods currently used for this purpose, SML and QRS analysis, are suitable for the assessments of EQ in marine coastal sediment under salmon farm influence. However, it was indicated that one of the used datasets was insufficient to capture the whole spatiotemporal variability of the bacterial community. Since these challenges did arise especially when using the QRS approach, the use of SML for such compromised datasets is recommended. In agreement with other studies, a good prediction of the already investigated regions in their respective seasons is expected, while it emerged that a large number of additional samples is required in order for the tool to be universally applicable (Aylagas et al., 2021; Frühe et al., 2021; Keeley et al., 2018; Lanzén et al., 2020). It has long been known that the bacterial community composition is incredibly diverse between different localities and time points. To develop a universally applicable monitoring tool, sampling along different stages of salmon farm production, across different regions, and throughout different seasons is thus required (Bowman and McCuaig, 2003; Delille, 1995; Frühe et al., 2021; Lindh et al., 2015; Prodinger et al., 2021). To obtain valuable bacterial eDNA community data of those environments, standardization of sample processing is a prerequisite to combine the acquired genetic information in a universally applicable system (Aylagas et al., 2021; Jeunen et al., 2019). The overall goal is to build a database of bacterial indicator ASVs, which exclusively react to salmon farm influences, regardless of geospatial or seasonal variations. With that, eDNA metabarcoding-based monitoring can replace traditional macrofauna-based monitoring for impacts of aquaculture installations on marine coastal environments.

For the extension of these necessary databases, SML has proven to be the best tool since its augmentation is very simple, and additionally environmental parameters can be incorporated into the algorithm (Cordier et al., 2020). This is expected to increase the predictive power by allowing to detangle the bacterial community patterns reflecting the desired organic enrichment from background noise induced by other environmental factors. Therefore, it is strongly recommended to measure environmental factors such as salinity, pH, temperature, metal contamination and organic carbon content simultaneously with eDNA samplings. As more samples are considered necessary for mapping the spatiotemporal heterogeneity of marine coastal sediments, an increase in project costs is expected, but can be counteracted by pooling of replicates before sequencing and multiplexing of samples (Aylagas et al., 2021; Esling et al., 2015; Hestetun et al., 2021a).

To estimate how many samples can be sequenced simultaneously without information loss, it is mandatory to approximate the required sequencing depth for the habitat under surveillance. So far, it was not known how much sequencing depth is required for eDNA-based inference of the environmental health status when using SML. This question is addressed in **Chapter IV**, where it was demonstrated that SML-based characterizations of the samples under investigation can be accomplished with relatively low sequencing depth, ranging from 50-5,000 sequences per sample, according to the habitat and factor under inspection. It was discovered that the separability of sample groups corresponding to abiotic or biotic conditions plays a large role in anticipating the accuracy of the prediction rather than the number of used sequences. Therefore, this thesis demonstrated that a shallow sequencing depth does not prevent accurate environmental assessments, which is why multiplexing harbors an even greater potential for monitoring purposes than expected. As sequencing depth can be reduced, more samples are allowed to be sequenced for the same amount of money. Additionally, the forward- and backward comparability of SML approaches described by Cordier et al. (2020) was emphasized, as RF was able to be applied for various prediction objectives describing several geospatial and temporal characteristics.

Recommendations for eDNA metabarcoding-based monitoring approaches

In summary, this thesis gives the following recommendations for the development of SOPs towards a universally applicable eDNA metabarcoding-based monitoring, combining the recent findings with the current literature: In order to obtain the greatest benefits from compliance monitoring, it is emphasized to sample eDNA in replicates to account for microhabitat patchiness of the seafloor. As reviewed in Hestetun et al. (2021a), such replicates can be pooled before sequencing to keep sequencing costs down. Accompanying to eDNA sampling, analysis of the benthic macrofauna is still recommended during database development as it can depict ecological health (Lanzén et al., 2021). Additional measuring of spatiotemporal environmental factors is strongly recommended to build a robust database and to later disentangle potential environmentally induced influences from the desired EQ induced changes (Cordier et al., 2020; Lanzén et al., 2020). Depending on the logistic circumstances, the samples can be preserved during transport either by freezing or by the addition of a nucleic acid preservation solution.

Extraction of eDNA should be carried out with a standardized kit that has already been tested and confirmed, such as the Qiagen's DNeasy PowerSoil kit (Pearman et al., 2020). If required, different laboratories can be commissioned for the subsequent sample processing. To ensure subsequent comparability between ASVs, the same amplification protocol and the same amplification primers should be used for the PCR (Golebiewski and Tretyn, 2020). A well-studied bacterial primer set is recommended for this purpose, accompanied by the use of a proofreading, high-fidelity DNA polymerase (Monroe et al., 2013). The subsequent sequencing should be carried out using equivalent methods, though Illumina sequencing is the most widely used method and therefore favorable. In order to anticipate the required sequencing depth, a pilot study of selected samples is recommended. These samples should cover the biggest possible variability of the habitat to be investigated. With the help of ordination analysis and variance assessments, the required number of sequences for the large-scale study can then be anticipated. In order to ensure latter inter-study comparability of the samples, selecting the sequencing depth in order to retrieve approx. 5,000 HQ-sequences after bioinformatic quality filtering steps is recommended. Sequence data processing should be conducted using a well-described tool, for example the DADA2 algorithm for ASV inference (Callahan et al., 2016).

Subsequently, SML regression models for the inference of the EQ can be constructed using the ASV-to-sample matrix. With each new sample added, the SML algorithm will become more robust against community characteristics induced by spatiotemporal heterogeneity rather than the desired factor under surveillance, e.g., organic enrichment. This can be optimized by a targeted sampling as well as by the addition of measured environmental parameters to the model (Cordier et al., 2020). When the database is sufficiently augmented, a universally applicable monitoring tool is produced which is not further reliant on laborious macrofauna analysis. When a novel sediment sample should be assessed regarding its EQ, the eDNA has to be processed as usual until the ASV-to-sample matrix is constructed. This matrix will then be inputted to the SML model, which automatically infers the EQ while detangling the response of the priorly inferred indicator ASVs for organic enrichment from potential background noise.

Future monitoring scenarios

The more dependable data already incorporated into the model, the more reliable the EQ inference will be. However, in order to build up this database, one first needs a large inter-study compatibility. To achieve this, molecular details must also be standardized as far as possible. Unfortunately, consensus of the research community is still missing regarding further sample processing steps as the choice of primers, DNA polymerase for PCR, and sequencing strategy, thus preventing the acquisition of comparable sequence datasets (Berry et al., 2011; Clarke et al., 2014; Elbrecht and Leese, 2015).

Recently, ASVs are most commonly used for eDNA studies, but one could also attempt to corroborate the ecosystem assessment on either native oligonucleotides or metagenomics (Armstrong and Verhoeven, 2020; Pawlowski et al., 2021). Instead of the inference of ASVs via amplicon sequencing, raw nucleotide segments or derivatives thereof could be directly incorporated into SML models. This alternative requires shotgun-sequencing of either eDNA for metagenomics describing potential gene functions, or eRNA for metatranscriptomics to assess gene expression (Taberlet et al., 2018). The advantage of the shotgun-sequencing method is that it requires no prior amplification of certain gene regions before sequencing. Hence, the use of amplification primers and a PCR polymerase is no longer necessary resulting in reduced PCR bias and potentially better inter-study comparability. Because no amplification of the template takes place, relative abundance information can be recovered in an uncompromised manner, and organisms which escaped metabarcoding approaches due to potential primer mismatches can be detected (Logares et al., 2014). The whole genetic information present in the samples is sequenced, including bacterial eDNA as well as eDNA of eukaryotes, which allows for the detection of potential bioindicators across taxonomic groups (Metzker, 2010). Since complete genomes are sequenced in random short fragments, extensive sequencing is required to obtain a good picture of the environment when using shotgun methods. Subsequently, further analyses such as classification and assembly of gained sequence fragments has to be conducted using bioinformatic tools. This process of genome recovery is presenting novel procedures to be established and standardized. Regarding shotgun sequencing, especially eRNA based methods such as meta-transcriptomics, are a promising alternative to traditional monitoring. It is assumed that beyond spatiotemporal taxonomic turnover, a conserved functional diversity of the microbial community can be recovered when using eRNA (Birrer et al., 2019; Cordier, 2020; Laroche et al., 2018).

Thus, eRNA has a high potential to indicate the year-round environmental influence of organic enrichment induced by salmon farms, even beyond national boundaries. Using this methodology, the number of samples required for the construction of a universally applicable tool for ecological quality inference could be dramatically reduced. Nevertheless, due to the known fragility of RNA, implementation in routine compliance monitoring is difficult, as sample handling becomes more complicated. To resolve this issue, eRNA-targeting microarrays can be integrated into so-called PhyloChips or GeoChips (Brodie et al., 2007; He et al., 2007). It is conceivable that EQ indicating genes, previously established by metatranscriptome analyses of marine sediments, can be detected and quantified using a microarray strategy. When added to the chip, the sample to be assessed reacts with a specific gene template in the form of cDNA on the chip, causing an e.g., photometrically measurable reaction (Singh et al., 2021). This reaction is recorded, and the environmental status can then be calculated and outputted in an easy to perceive manner, allowing quick estimates on EQ. Similar approaches for pollution monitoring using such microarrays seem promising (Cai et al., 2021; Lu and Su, 2021; Yin et al., 2022), and have recently been suggested for environmental monitoring of marine habitats (Lekang et al., 2020). However, the inter-usability of desired eRNA sequences and actual cDNA microarray detection has not yet been fully unraveled (Rachinger et al., 2021). Moreover, even though the technology for the chips has already been established, it is still necessary to find indicator genes and confirm their spatiotemporal applicability for marine coastal sediments under influence of aquaculture-induced organic enrichment. Alternatively, it has been suggested that eRNA sampling can help to improve eDNA results bioinformatically. By considering only shared sequences that are found in the eDNA and eRNA simultaneously, distinguishing between active and dormant organisms is feasible. Thus, extracellular non-target eDNA can be bioinformatically removed. As a result, technical artefacts can be detected and removed by this procedure, improving the overall sensitivity of the eDNA method (Laroche et al., 2018; Laroche et al., 2017).

Conclusion

To develop a universally applicable monitoring tool, augmentation of the eDNA datasets including samples showing spatiotemporal variability is required. Sequencing of a large number of such samples rather than individual samples to full saturation is recommended, which will result in the saving of valuable resources. In combination with the recording of prevailing environmental parameters, fast attainment of samples required for accurate EQ inference for biomonitoring purposes is enabled. Additionally, sampling of eRNA seems promising, because functional redundancy may be the key to a globally applicable EQ indication tool. Until this framework is technically mature for the implementation in SOPs, the currently available eDNA metabarcoding-based method can reliably characterize samples from already studied areas in the respective seasons. Throughout the thesis, eDNA metabarcoding was demonstrated to be sufficiently robust and reproducible regarding EQ assessments. Sample preservation approaches and EQ inference strategies were analyzed regarding their influence on the resulting environmental assessment and can now be proposed for implementation in SOPs. By aiding towards standardization, this work has laid the foundation for the inclusion of novel habitats, as comparable results can be obtained. In combination with SML, eDNA metabarcoding is suitable for assessments of organic enrichment induced by aquafarm installations. As soon as sufficient data has been fed into an SML-based monitoring system, it is universally applicable and can replace traditional macrofauna-based monitoring for the assessment of marine coastal environments.

In summary, this work has made a major contribution to the development of standardized eDNA-metabarcoding-based monitoring protocols. Based on made recommendations, SOPs can be established, which will enable all future studies to be effectively compared, taking the step from a validated method to a universally applicable monitoring tool.

References

Armstrong, E., & Verhoeven, J. (2020). Machine-learning analyses of bacterial oligonucleotide frequencies to assess the benthic impact of aquaculture. *Aquaculture Environment Interactions, 12,* 131–137. doi:10.3354/aei00353

Aylagas, E., Atalah, J., Sánchez-Jerez, P., Pearman, J. K., Casado, N., Asensi, J., Toledo-Guedes, K., & Carvalho, S. (2021). A step towards the validation of bacteria biotic indices using DNA metabarcoding for benthic monitoring. *Molecular Ecology Resources, 21*(6), 1889-1903. doi:10.1111/1755-0998.13395

Berry, D., Mahfoudh, K. B., Wagner, M., & Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology, 77*(21), 7846-7849. doi:10.1128/AEM.05220-11

Birrer, S. C., Dafforn, K. A., Sun, M. Y., Williams, R. B. H., Potts, J., Scanes, P., et al. (2019). Using meta-omics of contaminated sediments to monitor changes in pathways relevant to climate regulation. *Environmental Microbiology, 21*(1), 389-401. doi:10.1111/1462-2920.14470

Bourlat, S. J., Borja, A., Gilbert, J., Taylor, M. I., Davies, N., Weisberg, S. B., et al. (2013). Genomics in marine monitoring: New opportunities for assessing marine health status. *Marine Pollution Bulletin, 74*(1), 19-31. doi:10.1016/j.marpolbul.2013.05.042

Bowers, H. A., Pochon, X., von Ammon, U., Gemmell, N., Stanton, J.-A. L., Jeunen, G.-J., Sherman, C. D. H., & Zaiko, A. (2021). Towards the Optimization of eDNA/eRNA Sampling Technologies for Marine Biosecurity Surveillance. *Water, 13*(8), 1113. doi:10.3390/w13081113

Bowman, J., & McCuaig, R. (2003). Biodiversity, Community Structural Shifts, and Biogeography of Prokaryotes within Antarctic Continental Shelf Sediment. *Applied and Environmental Microbiology, 69,* 2463-2483. doi:10.1128/AEM.69.5.2463-2483.2003

Brodie, E. L., DeSantis, T. Z., Parker, J. P. M., Zubietta, I. X., Piceno, Y. M., & Andersen, G. L. (2007). Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences, 104*(1), 299-304. doi:10.1073/pnas.0608255104

Cai, W., Li, Y., Hu, J., & Cheng, H. (2021). Exploring the Microbial Ecological Functions in Response to Vertical Gradients in a Polluted Urban River. *CLEAN – Soil, Air, Water, 49*(9), 2100004. doi:10.1002/clen.202100004

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Chariton, A. A., Court, L. N., Hartley, D. M., Colloff, M. J., & Hardy, C. M. (2010). Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment, 8*(5), 233-238. doi:10.1890/090115

Clarke, L., Soubrier, J., Weyrich, L., & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources, 14,* 1160-1170. doi:10.1111/1755-0998.12265

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-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology, 30*, 2937-2958. doi:10.1111/mec.15472

Delille, D. (1995). Seasonal changes of subantarctic benthic bacterial communities. *Hydrobiologia, 310*(1), 47-57. doi:10.1007/BF00008182

Dowle, E., Pochon, X., Keeley, N., & Wood, S. A. (2015). Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *Fems Microbiology Ecology, 91*(8), fiv089. doi:10.1093/femsec/fiv089

Elbrecht, V., & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass-Sequence Relationships with an Innovative Metabarcoding Protocol. *Plos One, 10*(7), e0130324. doi:10.1371/journal.pone.0130324

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research, 43*(5), 2513-2524. doi:10.1093/nar/gkv107

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

Frühe, L., Dully, V., Forster, D., Keeley, N. B., Laroche, O., Pochon, X., Robinson, S., Wilding, T. A., & Stoeck, T. (2021). Global Trends of Benthic Bacterial Diversity and Community Composition Along Organic Enrichment Gradients of Salmon Farms. *Frontiers in Microbiology, 12*, e637811. doi:10.3389/fmicb.2021.637811

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307. doi:10.1111/2041-210X.12595

Golebiewski, M., & Tretyn, A. (2020). Generating amplicon reads for microbial community assessment with next-generation sequencing. *Journal of Applied Microbiology, 128*(2), 330-354. doi:10.1111/jam.14380

He, Z., Gentry, T. J., Schadt, C. W., Wu, L., Liebich, J., Chong, S. C., et al. (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal, 1*(1), 67-77. doi:10.1038/ismej.2007.2

Helbing, C. C., & Hobbs, J. (2019). Environmental DNA standardization needs for fish and wildlife population assessments and monitoring. Canadian Standards Association. *Retrieved on 10.02.2022 from https://www.csagroup.org/wp-content/uploads/CSA-Group-Research-Environmental-DNA.pdf.*

Hestetun, J. T., Lanzén, A., & Dahlgren, T. G. (2021a). Grab what you can—an evaluation of spatial replication to decrease heterogeneity in sediment eDNA metabarcoding. *PeerJ, 9*, e11619. doi:10.7717/peerj.11619

Hornick, K. M., & Buschmann, A. H. (2018). Insights into the diversity and metabolic function of bacterial communities in sediments from Chilean salmon aquaculture sites. *Annals of Microbiology, 68*(2), 63-77. doi:10.1007/s13213-017-1317-8

Jeunen, G.-J., Knapp, M., Spencer, H. G., Taylor, H. R., Lamare, M. D., Stat, M., Bunce, M., & Gemmell, N. J. (2019). Species-level biodiversity assessment using marine environmental DNA metabarcoding requires protocol optimization and standardization. *Ecology and Evolution, 9*(3), 1323-1335. doi:10.1002/ece3.4843

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *Plos One, 9*(1), e86175. doi:10.1371/journal.pone.0086175

Lanzén, A., Dahlgren, T. G., Bagi, A., & Hestetun, J. T. (2021). Benthic eDNA metabarcoding provides accurate assessments of impact from oil extraction, and ecological insights. *Ecological Indicators, 130*, 108064. doi:10.1016/j.ecolind.2021.108064

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology, 25*(17), 4392-4406. doi:10.1111/mec.13761

Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology, 30*, 2969-2987. doi:10.1111/mec.15489

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lear, G., & Pochon, X. (2018). A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Marine Pollution Bulletin, 127*, 97-107. doi:10.1016/j.marpolbul.2017.11.042

Laroche, O., Wood, S. A., Tremblay, L. A., Lear, G., Ellis, J. I., & Pochon, X. (2017). Metabarcoding monitoring analysis: the pros and cons of using co-extracted environmental DNA and RNA data to assess offshore oil production impacts on benthic communities. *PeerJ, 5*, e3347. doi:10.7717/peerj.3347

Le, J. T., Levin, L. A., Lejzerowicz, F., Cordier, T., Gooday, A. J., & Pawlowski, J. (2021). Scientific and budgetary trade-offs between morphological and molecular methods for deep-sea biodiversity assessment. *Integrated Environmental Assessment and Management*. doi:10.1002/ieam.4466

Lejzerowicz, F., Esling, P., & Pawlowski, J. (2014). Patchiness of deep-sea benthic Foraminifera across the Southern Ocean: Insights from high-throughput DNA sequencing. *Deep Sea Research Part II: Topical Studies in Oceanography, 108*, 17-26. doi:10.1016/j.dsr2.2014.07.018

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D., & Pawlowski, J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports, 5*, e13932 doi:10.1038/srep13932.

Lekang, K., Lanzén, A., Jonassen, I., Thompson, E., & Troedsson, C. (2020). Evaluation of a eukaryote phylogenetic microarray for environmental monitoring of marine sediments. *Marine Pollution Bulletin, 154*, 111102. doi:10.1016/j.marpolbul.2020.111102

Lindh, M. V., Sjöstedt, J., Andersson, A. F., Baltar, F., Hugerth, L. W., Lundin, D., Muthusamy, S., Legrand, C., & Pinhassi, J. (2015). Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environmental Microbiology, 17*(7), 2459-2476. doi:10.1111/1462-2920.12720

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology, 16*(9), 2659-2671. doi:10.1111/1462-2920.12250

Lu, Z., & Su, H. (2021). Employing gene chip technology for monitoring and assessing soil heavy metal pollution. *Environmental Monitoring and Assessment, 194*(1), 2. doi:10.1007/s10661-021-09650-6

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics, 11*(1), 31-46. doi:10.1038/nrg2626

Monroe, C., Grier, C., & Kemp, B. M. (2013). Evaluating the efficacy of various thermostable polymerases against co-extracted PCR inhibitors in ancient DNA samples. *Forensic Science International, 228*(1), 142-153. doi:10.1016/j.forsciint.2013.02.029

Nicholson, A., McIsaac, D., MacDonald, C., Gec, P., Mason, B. E., Rein, W., et al. (2020). An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines. *Environmental DNA, 2*(3), 343-349. doi:10.1002/edn3.81

Parsons, T. R., Takahashi, M., & Hargrave, B. (1977). *Biological oceanographic processes* (2nd ed.). Oxford: Pergamon Press.

Pawlowski, J., Bonin, A., Boyer, F., Cordier, T., & Taberlet, P. (2021). Environmental DNA for biomonitoring. *Molecular Ecology, 30*(13), 2931-2936. doi:10.1111/mec.16023

Pawlowski, J., Bruce, K., Panksep, K., Aguirre, F. I., Amalfitano, S., Apothéloz-Perret-Gentil, L., et al. (2022). Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. *Science of the Total Environment, 818*, 151783. doi:10.1016/j.scitotenv.2021.151783

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (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, J. K., Keeley, N. B., Wood, S. A., Laroche, O., Zaiko, A., Thomson-Laing, G., Biessy, L., Atalah, J., & Pochon, X. (2020). Comparing sediment DNA extraction methods for assessing organic enrichment associated with marine aquaculture. *PeerJ, 8*, e10231. doi:10.7717/peerj.10231

Prodinger, F., Endo, H., Takano, Y., Li, Y., Tominaga, K., Isozaki, T., et al. (2021). Year-round dynamics of amplicon sequence variant communities differ among eukaryotes, Imitervirales and prokaryotes in a coastal ecosystem. *FEMS microbiology ecology, 97*(12). doi:10.1101/2021.02.02.429489

Rachinger, N., Fischer, S., Böhme, I., Linck-Paulus, L., Kuphal, S., Kappelmann-Fenzl, M., & Bosserhoff, A. K. (2021). Loss of Gene Information: Discrepancies between RNA Sequencing, cDNA Microarray, and qRT-PCR. *International Journal of Molecular Sciences, 22*(17), e9349. doi:10.3390/ijms22179349

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio, 9*(3), e00525-00518. doi:10.1128/mBio.00525-18

Singh, J., Gupta, M., Singh, K. K., Kumar, A., Yadav, D., Wenjing, W., & Singh, P. K. (2021). Chapter 18 - Advancement in bioinformatics and microarray-based technologies for genome sequence analysis and its application in bioremediation of soil and water pollutants. In A. Kumar, V. K. Singh, P. Singh, & V. K. Mishra (Eds.), *Microbe Mediated Remediation of Environmental Contaminants* (pp. 209-225). Sawston: Woodhead Publishing.

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Taberlet, P., Bonin, A., Zinger, L., & Coissac, É. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press.

Yin, X., Wang, W., Wang, A., He, M., Lin, C., Ouyang, W., & Liu, X. (2022). Microbial community structure and metabolic potential in the coastal sediments around the Yellow River Estuary. *Science of the Total Environment, 816*, e151582. doi:10.1016/j.scitotenv.2021.151582

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., et al. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology, 28*(8), 1857-1862. doi:10.1111/mec.15060

BIBLIOGRAPHY

Ahmed, K. R., Akter, S., Marandi, A., & Schüth, C. (2021). A simple and robust wetland classification approach by using optical indices, unsupervised and supervised machine learning algorithms. *Remote Sensing Applications: Society and Environment, 23*, e100569. doi:10.1016/j.rsase.2021.100569

Alon, S., Vigneault, F., Eminaga, S., Christodoulou, D. C., Seidman, J. G., Church, G. M., & Eisenberg, E. (2011). Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Research, 21*(9), 1506-1511. doi:10.1101/gr.121715.111

Alve, E., Korsun, S., Schönfeld, J., Dijkstra, N., Golikova, E., Hess, S., Husum, K., & Panieri, G. (2016). Foram-AMBI: A sensitivity index based on benthic foraminiferal faunas from North-East Atlantic and Arctic fjords, continental shelves and slopes. *Marine Micropaleontology, 122*, 1-12. doi:10.1016/j.marmicro.2015.11.001

Amir, A., McDonald, D., Navas-Molina Jose, A., Kopylova, E., Morton James, T., Zech Xu, Z., et al. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems, 2*(2), e00191-00116. doi:10.1128/mSystems.00191-16

Anderson, M. (2008). Animal-sediment relationships re-visited: Characterising species' distributions along an environmental gradient using canonical analysis and quantile regression splines. *Journal of Experimental Marine Biology and Ecology, 366*, 16-27. doi:10.1016/j.jembe.2008.07.006

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32-46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Apothéloz-Perret-Gentil, L., Bouchez, A., Cordier, T., Cordonier, A., Guéguen, J., Rimet, F., Vasselon, V., & Pawlowski, J. (2021). Monitoring the ecological status of rivers with diatom eDNA metabarcoding: A comparison of taxonomic markers and analytical approaches for the inference of a molecular diatom index. *Molecular Ecology, 30*(13), 2959-2968. doi:10.1111/mec.15646

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources, 17*(6), 1231-1242. doi:10.1111/1755-0998.12668

Aravindraja, C., Viszwapriya, D., & Karutha Pandian, S. (2013). Ultradeep 16S rRNA Sequencing Analysis of Geographically Similar but Diverse Unexplored Marine Samples Reveal Varied Bacterial Community Composition. *Plos One, 8*(10), e76724. doi:10.1371/journal.pone.0076724

Armstrong, E., & Verhoeven, J. (2020). Machine-learning analyses of bacterial oligonucleotide frequencies to assess the benthic impact of aquaculture. *Aquaculture Environment Interactions, 12*, 131–137. doi:10.3354/aei00353

Armstrong, E. G., Mersereau, J., Salvo, F., Hamoutene, D., & Dufour, S. C. (2020). Temporal change in the spatial distribution of visual organic enrichment indicators at aquaculture sites in Newfoundland, Canada. *Aquaculture International, 28*(2), 569-586. doi:10.1007/s10499-019-00478-z

Aylagas, E., Atalah, J., Sánchez-Jerez, P., Pearman, J. K., Casado, N., Asensi, J., Toledo-Guedes, K., & Carvalho, S. (2021). A step towards the validation of bacteria biotic indices using DNA metabarcoding for benthic monitoring. *Molecular Ecology Resources, 21*(6), 1889-1903. doi:10.1111/1755-0998.13395

Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Frontiers in Marine Science, 3*(96). doi:10.3389/fmars.2016.00096

Aylagas, E., Borja, A., & Rodriguez-Ezpeleta, N. (2014). Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). *Plos One, 9*(3), e90529. doi:10.1371/journal.pone.0090529

Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., Irigoien, X., Danovaro, R., & Rodríguez-Ezpeleta, N. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Marine Pollution Bulletin, 114*(2), 679-688. doi:10.1016/j.marpolbul.2016.10.050

Aylagas, E., & Rodriguez-Ezpeleta, N. (2016). Analysis of Illumina MiSeq Metabarcoding Data: Application to Benthic Indices for Environmental Monitoring. *Methods in molecular biology, 1452*, 237-249. doi:10.1007/978-1-4939-3774-5_16

AZTI (2022). AMBI by AZTI, AZTI's Marine Biotic Index. *Retrieved on 15.03.2022 from https://ambi.azti.es/.*

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411-421. doi:10.1097/01.psy.0000127692.23278.a9

Bagley, M., Pilgrim, E., Knapp, M., Yoder, C., Santo Domingo, J., & Banerji, A. (2019). High-throughput environmental DNA analysis informs a biological assessment of an urban stream. *Ecological Indicators, 104*, 378-389. doi:10.1016/j.ecolind.2019.04.088

Baird, D., & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology, 21*, 2039-2044. doi:10.1111/j.1365-294X.2012.05519.x

Bannister, R. J., Valdemarsen, T., Hansen, P. K., Holmer, M., & Ervik, A. (2014). Changes in benthic sediment conditions under an Atlantic salmon farm at a deep, well-flushed coastal site. *Aquaculture Environment Interactions, 5*(1), 29-47. doi:10.3354/aei00092

Barbier, E. B., Hacker, S. D., Kennedy, C., Koch, E. W., Stier, A. C., & Silliman, B. R. (2011). The value of estuarine and coastal ecosystem services. *Ecological Monographs, 81*(2), 169-193. doi:10.1890/10-1510.1

Bates, M. D., & Venables, B. (2011). R Package 'splines': Regression Spline Functions and Classes. *Version 2.0.*

Berry, D., Mahfoudh, K. B., Wagner, M., & Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology, 77*(21), 7846-7849. doi:10.1128/AEM.05220-11

Biddanda, B., Ogdahl, M., & Cotner, J. (2001). Dominance of bacterial metabolism in oligotrophic relative to eutrophic waters. *Limnology and Oceanography, 46*(3), 730-739. doi:10.4319/lo.2001.46.3.0730

Birrer, S. C., Dafforn, K. A., Sun, M. Y., Williams, R. B. H., Potts, J., Scanes, P., et al. (2019). Using meta-omics of contaminated sediments to monitor changes in pathways relevant to climate regulation. *Environmental Microbiology, 21*(1), 389-401. doi:10.1111/1462-2920.14470

Bissett, A., Burke, C., Cook, P. L. M., & Bowman, J. P. (2007). Bacterial community shifts in organically perturbed sediments. *Environmental Microbiology, 9*(1), 46-60. doi:10.1111/j.1462-2920.2006.01110.x

Boers, S. A., Jansen, R., & Hays, J. P. (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *European Journal of Clinical Microbiology & Infectious Diseases, 38*(6), 1059-1070. doi:10.1007/s10096-019-03520-3

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods, 10*(1), 57. doi:10.1038/nmeth.2276

Bonada, N., Prat, N., Resh, V. H., & Statzner, B. (2006). Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology, 51*, 495-523. doi:10.1146/annurev.ento.51.110104.151124

Borja, A. (2018). Testing the efficiency of a bacterial community-based index (microgAMBI) to assess distinct impact sources in six locations around the world. *Ecological Indicators, 85*, 594-602. doi:10.1016/j.ecolind.2017.11.018

Borja, A., Franco, J., & Pérez, V. (2000). A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments. *Marine Pollution Bulletin, 40*(12), 1100-1114. doi:10.1016/S0025-326X(00)00061-8

Bourlat, S. J., Borja, A., Gilbert, J., Taylor, M. I., Davies, N., Weisberg, S. B., et al. (2013). Genomics in marine monitoring: New opportunities for assessing marine health status. *Marine Pollution Bulletin, 74*(1), 19-31. doi:10.1016/j.marpolbul.2013.05.042

Bowers, H. A., Pochon, X., von Ammon, U., Gemmell, N., Stanton, J.-A. L., Jeunen, G.-J., Sherman, C. D. H., & Zaiko, A. (2021). Towards the Optimization of eDNA/eRNA Sampling Technologies for Marine Biosecurity Surveillance. *Water, 13*(8), 1113. doi:10.3390/w13081113

Bowman, J., & McCuaig, R. (2003). Biodiversity, Community Structural Shifts, and Biogeography of Prokaryotes within Antarctic Continental Shelf Sediment. *Applied and Environmental Microbiology, 69*, 2463-2483. doi:10.1128/AEM.69.5.2463-2483.2003

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Brodie, E. L., DeSantis, T. Z., Parker, J. P. M., Zubietta, I. X., Piceno, Y. M., & Andersen, G. L. (2007). Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences, 104*(1), 299-304. doi:10.1073/pnas.0608255104

Brown, C. (2012). R Package 'pragma': Provides a pragma / directive / keyword syntax for R. *Version 0.1.3*.

Brown, J. R., Gowen, R. J., & McLusky, D. S. (1987). The effect of salmon farming on the benthos of a Scottish sea loch. *Journal of Experimental Marine Biology and Ecology, 109*(1), 39-51. doi:10.1016/0022-0981(87)90184-5

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1842*(10), 1932-1941. doi:10.1016/j.bbadis.2014.06.015

Burridge, L., Weis, J. S., Cabello, F., Pizarro, J., & Bostick, K. (2010). Chemical use in salmon aquaculture: A review of current practices and possible environmental effects. *Aquaculture, 306*(1-4), 7-23. doi:10.1016/j.aquaculture.2010.05.020

Cai, W., Li, Y., Hu, J., & Cheng, H. (2021). Exploring the Microbial Ecological Functions in Response to Vertical Gradients in a Polluted Urban River. *CLEAN – Soil, Air, Water, 49*(9), 2100004. doi:10.1002/clen.202100004

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal, 11*(12), 2639-2643. doi:10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335-336. doi:10.1038/nmeth.f.303

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., et al. (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol, 12*(1), 158. doi:10.1186/1471-2180-12-158

Carignan, V., & Villard, M.-A. (2002). Selecting Indicator Species to Monitor Ecological Integrity: A Review. *Environmental Monitoring and Assessment, 78*(1), 45-61. doi:10.1023/A:1016136723584

Carroll, M. L., Cochrane, S., Fieler, R., Velvin, R., & White, P. (2003). Organic enrichment of sediments from salmon farming in norway: environmental factors, management practices, and monitoring techniques. *Aquaculture, 226*, 165-180. doi:10.1016/S0044-8486(03)00475-7

Chambers, J. M., & Hastie, T. (1992). *Statistical Models in S*. New York: Chapman and Hall/CRC.

Chandler, D. P., Fredrickson, J. K., & Brockman, F. J. (1997). Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. *Molecular Ecology, 6*(5), 475-482. doi:10.1046/j.1365-294X.1997.00205.x

Chariton, A. A., Court, L. N., Hartley, D. M., Colloff, M. J., & Hardy, C. M. (2010). Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment, 8*(5), 233-238. doi:10.1890/090115

Chariton, A. A., Stephenson, S., Morgan, M. J., Steven, A. D. L., Colloff, M. J., Court, L. N., & Hardy, C. M. (2015). Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental Pollution, 203*, 165-174. doi:10.1016/j.envpol.2015.03.047

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics, 99*(6), 323-329. doi:10.1016/j.ygeno.2012.04.003

Chen, Z., Hui, P. C., Hui, M., Yeoh, Y. K., Wong, P. Y., Chan, M. C. W., et al. (2019). Impact of Preservation Method and 16S rRNA Hypervariable Region on Gut Microbiota Profiling. *mSystems, 4*(1), e00271-00218. doi:10.1128/mSystems.00271-18

Chiang, F., Mazdiyasni, O., & AghaKouchak, A. (2021). Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. *Nature Communications, 12*, e2754. doi:10.1038/s41467-021-22314-w

Christensen, K. H., Sperrevik, A. K., & Broström, G. (2018). On the variability in the onset of the Norwegian Coastal Current. *Journal of Physical Oceanography, 48*(3), 723-738. doi:10.1175/JPO-D-17-0117.1

Clarke, L., Soubrier, J., Weyrich, L., & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources, 14*, 1160-1170. doi:10.1111/1755-0998.12265

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-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology, 30*, 2937-2958. doi:10.1111/mec.15472

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology, 51*(16), 9118-9126. doi:10.1021/acs.est.7b01518

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources, 18*(6), 1381-1391. doi:10.1111/1755-0998.12926

Cordier, T., Frontalini, F., Cermakova, K., Apothéloz-Perret-Gentil, L., Treglia, M., Scantamburlo, E., Bonamin, V., & Pawlowski, J. (2019a). Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Marine Environmental Research, 146*, 24-34. doi:10.1016/j.marenvres.2018.12.009

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019b). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology, 27*(5), 387-397. doi:10.1016/j.tim.2018.10.012

Cromey, C., Nickell, T., & Black, K. (2002). DEPOMOD – modeling the deposition and biological effects of waste solids from marine cage farms. *Aquaculture, 214*, 211-239. doi:10.1016/S0044-8486(02)00368-X

Cuthbertson, L., Rogers, G. B., Walker, A. W., Oliver, A., Hoffman, L. R., Carroll, M. P., Parkhill, J., Bruce, K. D., & van der Gast, C. J. (2015). Implications of multiple freeze-thawing on respiratory samples for culture-independent analyses. *Journal of Cystic Fibrosis, 14*(4), 464-467. doi:10.1016/j.jcf.2014.10.004

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for classification in Ecology. *Ecology, 88*(11), 2783-2792. doi:doi.org/10.1890/07-0539.1

Dafforn, K., Baird, D., Chariton, A., Sun, M., Brown, M., Simpson, S., Kelaher, B., & Johnston, E. (2014). Chapter One - Faster, Higher and Stronger? The Pros and Cons of Molecular Faunal Data for Assessing Ecosystem Condition. In G. Woodward, A. J. Dumbrell, D. J. Baird, & M. Hajibabaei (Eds.), *Advances in Ecological Research: Big Data in Ecology* (pp. 1-40). Essex: Academic Press, Elsevier Ltd.

Daily, G. C. (2013). Nature's Services: Societal Dependence on Natural Ecosystems (1997). In L. Robin, S. Sörlin, & P. Warde (Eds.), *The Future of Nature: Documents of Global Change* (pp. 454-464). London: Yale University Press.

Danovaro, R., Carugati, L., Berzano, M., Cahill, A. E., Carvalho, S., Chenuil, A., et al. (2016). Implementing and Innovating Marine Monitoring Approaches for Assessing Marine Environmental Status. *Frontiers in Marine Science, 3*, 213. doi:10.3389/fmars.2016.00213

Darling, J., & Mahon, A. (2011). From molecules to management: Adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research, 111*, 978-988. doi:10.1016/j.envres.2011.02.001

DeAngelis, K. M., Silver, W. L., Thompson, A. W., & Firestone, M. K. (2010). Microbial communities acclimate to recurring changes in soil redox potential status. *Environmental Microbiology, 12*(12), 3137-3149. doi:10.1111/j.1462-2920.2010.02286.x

Dehler, C. E., Secombes, C. J., & Martin, S. A. M. (2017). Environmental and physiological factors shape the gut microbiota of Atlantic salmon parr (*Salmo salar L.*). *Aquaculture, 467*, 149-157. doi:10.1016/j.aquaculture.2016.07.017

Deist, T. M., Dankers, F., Valdes, G., Wijsman, R., Hsu, I. C., Oberije, C., et al. (2018). Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Medical Physics, 45*(7), 3449-3459. doi:10.1002/mp.12967

Delille, D. (1993). Seasonal changes in the abundance and composition of marine heterotrophic bacterial communities in an Antarctic coastal area. *Polar Biology, 13*(7), 463-470. doi:10.1007/BF00233137

Delille, D. (1995). Seasonal changes of subantarctic benthic bacterial communities. *Hydrobiologia, 310*(1), 47-57. doi:10.1007/BF00008182

Dominianni, C., Wu, J., Hayes, R. B., & Ahn, J. (2014). Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiolology, 14*(1), 103. doi:10.1186/1471-2180-14-103

Dowle, E., Pochon, X., Keeley, N., & Wood, S. A. (2015). Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. *Fems Microbiology Ecology, 91*(8), fiv089. doi:10.1093/femsec/fiv089

Drius, M., Bongiorni, L., Depellegrin, D., Menegon, S., Pugnetti, A., & Stifter, S. (2019). Tackling challenges for Mediterranean sustainable coastal tourism: An ecosystem service perspective. *Science of the Total Environment, 652*, 1302-1317. doi:10.1016/j.scitotenv.2018.10.121

Dufrene, M., & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs, 67*(3), 345-366. doi:10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2

Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., & Stoeck, T. (2021a). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter-laboratory study. *Ecological Indicators, 121*, e107049. doi:10.1016/j.ecolind.2020.107049

Dully, V., Rech, G., Wilding, T. A., Lanzén, A., MacKichan, K., Berrill, I., & Stoeck, T. (2021b). Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems. *Marine Pollution Bulletin, 173*, e113129. doi:10.1016/j.marpolbul.2021.113129

Dully, V., Wilding, T. A., Mühlhaus, T., & Stoeck, T. (2021c). Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning. *Computational and Structural Biotechnology Journal, 19*, 2256-2268. doi:10.1016/j.csbj.2021.04.005

Dyksma, S., Pjevac, P., Ovanesov, K., & Mussmann, M. (2018). Evidence for H2 consumption by uncultured Desulfobacterales in coastal sediments. *Environmental Microbiology, 20*(2), 450-461. doi:10.1111/1462-2920.13880

Edgar, R. (2016). Preprint: UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. doi:10.1101/081257

EEA (2018). European Environment Agency. Indicator Assessment - Aquaculture production in Europe. *Retrieved on 15.02.2022 from www.eea.europa.eu/data-and-maps/indicators/aquaculture-production-4/assessment.*

Elbrecht, V., & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass-Sequence Relationships with an Innovative Metabarcoding Protocol. *Plos One, 10*(7), e0130324. doi:10.1371/journal.pone.0130324

Elmogy, A. M., Tariq, U., Mohammed, A., & Ibrahim, A. (2021). Fake Reviews Detection using Supervised Machine Learning. *International Journal of Advanced Computer Science and Applications, 12*(1), 301-606. doi:10.14569/IJACSA.2021.0120169

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research, 43*(5), 2513-2524. doi:10.1093/nar/gkv107

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2011). Modeling Species Distribution and Change Using Random Forest. In C. A. Drew, Y. F. Wiersma, & F. Huettmann (Eds.), *Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications* (pp. 139-159). New York: Springer.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research, 8*(3), 175-185. doi:10.1101/gr.8.3.175

FAO (1996). Food and Agriculture Organization of the United Nations. The contributions of science to integrated coastal management, Joint Group of Experts on the Scientific Aspects of Marine Environmental Protection (GESAMP). *Retrieved on 21.01.2022 from http://www.gesamp.org/site/assets/files/1239/the-contributions-of-science-to-integrated-coastal-management-en.pdf.*

FAO (2018). Food and Agriculture Organization of the United Nations. The State of World Fisheries and Aquaculture 2018. *Retreived on 05.01.2022 from https://www.fao.org/documents/card/en/c/I9540EN.*

FAO (2020). Food and Agriculture Organization of the United Nations. The State of World Fisheries and Aquaculture 2020. *Retrieved on 03.03.2022 from https://www.fao.org/publications/card/en/c/CA9229EN.*

Faraway, J. J. (2004). *Linear models with R.* New York: Chapman and Hall/CRC.

Filippidou, S., Junier, T., Wunderlin, T., Lo, C.-C., Li, P.-E., Chain, P. S., & Junier, P. (2015). Under-detection of endospore-forming Firmicutes in metagenomic data. *Computational and Structural Biotechnology Journal, 13*, 299-306. doi:10.1016/j.csbj.2015.04.002

Findlay, R. H., & Watling, L. (1997). Prediction of benthic impact for salmon net-pens based on the balance of benthic oxygen supply and demand. *Marine Ecology Progress Series, 155*, 147-157. doi:10.3354/meps155147

Findlay, R. H., & Watling, L. (1998). Seasonal Variation in the Structure of a Marine Benthic Microbial Community. *Microbial Ecology, 36*(1), 23-30. doi:10.1007/s002489900089

Finlay, B. J. (2002). Global dispersal of free-living microbial eukaryote species. *Science, 296*(5570), 1061-1063. doi:10.1126/science.1070710

Finster, K., Liesack, W., & Thamdrup, B. (1998). Elemental sulfur and thiosulfate disproportionation by Desulfocapsa sulfoexigens sp. nov., a new anaerobic bacterium isolated from marine surface sediment. *Applied and Environmental Microbiology, 64*(1), 119-125. doi:10.1128/aem.64.1.119-125.1998

Fogarty, C., Burgess, C. M., Cotter, P. D., Cabrera-Rubio, R., Whyte, P., Smyth, C., & Bolton, D. J. (2019). Diversity and composition of the gut microbiota of Atlantic salmon (Salmo salar) farmed in Irish waters. *Journal of Applied Microbiology, 127*(3), 648-657. doi:10.1111/jam.14291

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment, 80*(1), 185-201. doi:10.1016/S0034-4257(01)00295-4

Forrest, B. M., Keeley, N., Gillespie, P., Hopkins, G., Knight, B., & Govier, D. (2007). Review of the Ecological Effects of Marine Finfish Aquaculture: Final Report. The Ministry of Fisheries, New Zealand. *Retrieved on 18.01.2022 from https://www.yumpu.com/en/document/read/20615870/review-of-the-ecological-effects-of-marine-finfish-aquaculture-final-*.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., Lopez, P., Stoeck, T., & Bapteste, E. (2015). Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *Bmc Biology, 13*(1), 16. doi:10.1186/s12915-015-0125-5

Forster, D., Filker, S., Kochems, R., Breiner, H.-W., Cordier, T., Pawlowski, J., & Stoeck, T. (2019a). A Comparison of Different Ciliate Metabarcode Genes as Bioindicators for Environmental Impact Assessments of Salmon Aquaculture. *Journal of Eukaryotic Microbiology, 66*(2), 294-308. doi:10.1111/jeu.12670

Forster, D., Lentendu, G., Filker, S., Dubois, E., Wilding, T. A., & Stoeck, T. (2019b). Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environmental Microbiology, 21*(11), 4109-4124. doi:10.1111/1462-2920.14764

Fortunato, C. S., Eiler, A., Herfort, L., Needoba, J. A., Peterson, T. D., & Crump, B. C. (2013). Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. *The ISME Journal, 7*(10), 1899-1911. doi:10.1038/ismej.2013.79

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment, 189*(7), 316. doi:10.1007/s10661-017-6025-0

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., & Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Applied and Environmental Microbiology, 74*(8), 2461-2470. doi:10.1128/aem.02272-07

Freeman, E. A., Moisen, G. G., Coulston, J. W., & Wilson, B. T. (2015). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Canadian Journal of Forest Research, 46*(3), 323-339. doi:10.1139/cjfr-2014-0562

Frontalini, F., & Coccioni, R. (2011). Benthic foraminifera as bioindicators of pollution: a review of Italian research over the last three decades. *Revue de micropaléontologie, 54*(2), 115-127. doi:10.1016/j.revmic.2011.03.001

Frühe, L. (2021). *The potential of benthic microbes as bioindicators in coastal aquaculture impact assessments using eDNA metabarcoding*. Doctoral dissertation, Technische Universität Kaiserslautern.

Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology, 30*, 2988–3006. doi:10.1111/mec.15434

Frühe, L., Dully, V., Forster, D., Keeley, N. B., Laroche, O., Pochon, X., Robinson, S., Wilding, T. A., & Stoeck, T. (2021). Global Trends of Benthic Bacterial Diversity and Community Composition Along Organic Enrichment Gradients of Salmon Farms. *Frontiers in Microbiology, 12*, e637811. doi:10.3389/fmicb.2021.637811

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., & Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences of the United States of America, 105*(22), 7774-7778. doi:10.1073/pnas.0803070105

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). R Package 'irr': Various coefficients of interrater reliability and agreement. *Version 0.84.1*.

Gao, J., Hou, L., Zheng, Y., Liu, M., Yin, G., Li, X., et al. (2016). nirS-Encoding denitrifier community composition, distribution, and abundance along the coastal wetlands of China. *Applied Microbiology and Biotechnology, 100*(19), 8573-8582. doi:10.1007/s00253-016-7659-5

Gerhard, W., & Gunsch, C. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environment International, 124*, 312-319. doi:10.1016/j.envint.2018.12.038

Gihring, T. M., Green, S. J., & Schadt, C. W. (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environmental Microbiology, 14*(2), 285-290. doi:10.1111/j.1462-2920.2011.02550.x

Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Huse, S., & Joint, I. (2009). The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology, 11*(12), 3132-3139. doi:10.1111/j.1462-2920.2009.02017.x

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300. doi:10.1016/j.patrec.2005.08.011

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., et al. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology, 34*(9), 942-949. doi:10.1038/nbt.3601

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307. doi:10.1111/2041-210X.12595

Golebiewski, M., & Tretyn, A. (2020). Generating amplicon reads for microbial community assessment with next-generation sequencing. *Journal of Applied Microbiology, 128*(2), 330-354. doi:10.1111/jam.14380

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351. doi:10.1038/nrg.2016.49

Gowen, R. J., & Bradbury, N. B. (1987). The Ecological Impact of Salmonid Farming in Coastal Waters - a Review. *Oceanography and Marine Biology, 25*, 563-575. doi:10.1016/0198-0254(88)92716-1

Grall, J., & Glémarec, M. (1997). Using biotic indices to estimate macrobenthic community perturbations in the Bay of Brest. *Estuarine, Coastal and Shelf Science, 44*, 43-53. doi:10.1016/S0272-7714(97)80006-6

Gray, J. S. (1997). Marine biodiversity: patterns, threats and conservation needs. *Biodiversity & Conservation, 6*(1), 153-175. doi:10.1023/A:1018335901847

Gray, M. A., Pratte, Z. A., & Kellogg, C. A. (2013). Comparison of DNA preservation methods for environmental bacterial community samples. *FEMS Microbiology Ecology, 83*(2), 468-477. doi:10.1111/1574-6941.12008

Gribben, P. E., Nielsen, S., Seymour, J. R., Bradley, D. J., West, M. N., & Thomas, T. (2017). Microbial communities in marine sediments modify success of an invasive macrophyte. *Scientific Reports, 7*(1), 9845. doi:10.1038/s41598-017-10231-2

Guo, F., Wang, G., Su, Z., Liang, H., Wang, W., Lin, F., & Liu, A. (2016). What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire, 25*(5), 505-519. doi:10.1071/WF15121

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research, 21*(3), 494-504. doi:10.1101/gr.112730.110

Hajishengallis, G., Darveau, R. P., & Curtis, M. A. (2012). The keystone-pathogen hypothesis. *Nature Reviews Microbiology, 10*(10), 717-725. doi:10.1038/nrmicro2873

Harley, C. D., Randall Hughes, A., Hultgren, K. M., Miner, B. G., Sorte, C. J., Thornber, C. S., Rodriguez, L. F., Tomanek, L., & Williams, S. L. (2006). The impacts of climate change in coastal marine systems. *Ecology Letters, 9*(2), 228-241. doi:10.1111/j.1461-0248.2005.00871.x

Hassan, R., Scholes, R., & Ash, N. (2005). Millennium Ecosystem Assessment Series. Ecosystems and Human Well-Being: Current State and Trends: Findings of the Condition and Trends Working Group. *Retrieved on 15.02.2022 from https://www.millenniumassessment.org/documents/document.766.aspx.pdf.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). Random forests. In *The elements of statistical learning* (2nd ed., pp. 587-604). New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009b). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. doi:10.1109/TKDE.2008.239

He, Z., Gentry, T. J., Schadt, C. W., Wu, L., Liebich, J., Chong, S. C., et al. (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal, 1*(1), 67-77. doi:10.1038/ismej.2007.2

Heal, G. M., Barbier, E. B., Boyle, K. J., Covich, A. P., Gloss, S. P., Carlton H. Hershner, J., et al. (2005). *Valuing Ecosystem Services: Toward Better Environmental Decision-Making*. Washington DC: The National Academies Press.

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society - Biological Sciences, 270*(1512), 313-321. doi:10.1098/rspb.2002.2218

Helbing, C. C., & Hobbs, J. (2019). Environmental DNA standardization needs for fish and wildlife population assessments and monitoring. Canadian Standards Association. *Retrieved on 10.02.2022 from https://www.csagroup.org/wp-content/uploads/CSA-Group-Research-Environmental-DNA.pdf.*

Herlemann, D. P. R., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., & Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal, 5*(10), 1571-1579. doi:10.1038/ismej.2011.41

Hestetun, J. T., Lanzén, A., & Dahlgren, T. G. (2021a). Grab what you can—an evaluation of spatial replication to decrease heterogeneity in sediment eDNA metabarcoding. *PeerJ, 9*, e11619. doi:10.7717/peerj.11619

Hestetun, J. T., Lanzén, A., Skaar, K. S., & Dahlgren, T. G. (2021b). The impact of DNA extract homogenization and replication on marine sediment metabarcoding diversity and heterogeneity. *Environmental DNA, 3*, 997–1006. doi:10.1002/edn3.223

Hewson, I., & Fuhrman, J. (2006). Spatial and vertical biogeography of coral reef sediment bacterial and diazotroph communities. *Marine Ecology Progress Series, 306*, 79-86. doi:10.3354/meps306079

Hiergeist, A., Reischl, U., & Gessner, A. (2016). Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *International Journal of Medical Microbiology, 306*(5), 334-342. doi:10.1016/j.ijmm.2016.03.005

Hino, A., Maruyama, H., & Kikuchi, T. (2016). A novel method to assess the biodiversity of parasites using 18S rDNA Illumina sequencing; parasitome analysis method. *Parasitology International, 65*(5), 572-575. doi:10.1016/j.parint.2016.01.009

Hintermeier-Erhard, G., & Zech, W. (1997). *Wörterbuch der Bodenkunde - Systematik, Genese, Eigenschaften, Ökologie und Verbreitung von Böden*. Stuttgart: Springer Spektrum.

Hirche, H. J. (1987). Temperature and plankton. *Marine Biology, 94*(3), 347-356. doi:10.1007/BF00428240

Hoegh-Guldberg, O., & Bruno, J. F. (2010). The impact of climate change on the world's marine ecosystems. *Science, 328*(5985), 1523-1528. doi:10.1126/science.1189930

Holmer, M., Wildish, D., & Hargrave, B. T. (2005). Organic Enrichment from Marine Finfish Aquaculture and Effects on Sediment Biogeochemical Processes. In B. T. Hargrave (Ed.), *Environmental Effects of marine Finfish Aquaculture* (pp. 182-206). New York: Springer.

Hornick, K. M., & Buschmann, A. H. (2018). Insights into the diversity and metabolic function of bacterial communities in sediments from Chilean salmon aquaculture sites. *Annals of Microbiology, 68*(2), 63-77. doi:10.1007/s13213-017-1317-8

Hughes, S. (1975). *National environmental policy act of 1969*. Washington, D.C.: Congressional Research Service, Library of Congress.

Hvas, M., Folkedal, O., Solstorm, D., Vågseth, T., Fosse, J. O., Gansel, L. C., & Oppedal, F. (2017). Assessing swimming capacity and schooling behaviour in farmed Atlantic salmon Salmo salar with experimental push-cages. *Aquaculture, 473*, 423-429. doi:10.1016/j.aquaculture.2017.03.013

Illumina (2013). Metagenomic Sequencing Library Preparation. *Retrieved on 20.06.2020 from https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf*

Illumina (2017). Metagenomic Sequencing Library Preparation. *Retrieved on 01.01.2022 from https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf*

Inagaki, F., Suzuki, M., Takai, K., Oida, H., Sakamoto, T., Aoki, K., Nealson, K. H., & Horikoshi, K. (2003). Microbial Communities Associated with Geological Horizons in Coastal Subseafloor Sediments from the Sea of Okhotsk. *Applied and Environmental Microbiology, 69*(12), 7224-7235. doi:10.1128/AEM.69.12.7224-7235.2003

IPCC (2007). Intergovernmental Panel on Climate Change. Climate Change 2007: Fourth Assessment Report. The Physical Science Basis, Summary for Policymakers. *Retrieved on 25.01.2021 from https://previa.uclm.es/area/amf/antoine/energias/Ipcc_anotado.pdf*

Iturbe-Espinoza, P., Brandt, B. W., Braster, M., Bonte, M., Brown, D. M., & van Spanning, R. J. M. (2021). Effects of DNA preservation solution and DNA extraction methods on microbial community profiling of soil. *Folia Microbiologica, 66*, 597–606. doi:10.1007/s12223-021-00866-0

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Jansson, M., Bergström, A. K., Lymer, D., Vrede, K., & Karlsson, J. (2006). Bacterioplankton growth and nutrient use efficiencies under variable organic carbon and inorganic phosphorus ratios. *Microbial Ecology, 52*(2), 358-364. doi:10.1007/s00248-006-9013-4

Jeunen, G.-J., Knapp, M., Spencer, H. G., Taylor, H. R., Lamare, M. D., Stat, M., Bunce, M., & Gemmell, N. J. (2019). Species-level biodiversity assessment using marine environmental DNA metabarcoding requires protocol optimization and standardization. *Ecology and Evolution, 9*(3), 1323-1335. doi:10.1002/ece3.4843

Kalle, E., Kubista, M., & Rensing, C. (2014). Multi-template polymerase chain reaction. *Biomolecular Detection and Quantification, 2*, 11-29. doi:10.1016/j.bdq.2014.11.002

Karr, J. R. (1991). Biological Integrity: A Long-Neglected Aspect of Water Resource Management. *Ecological Applications, 1*(1), 66-84. doi:10.2307/1941848

Kawahara, N., Shigematsu, K., Miura, S., Miyadai, T., & Kondo, R. (2008). Distribution of sulfate-reducing bacteria in fish farm sediments on the coast of southern Fukui Prefecture, Japan. *Plankton and Benthos Research, 3*(1), 42-45. doi:10.3800/pbr.3.42

Kawahara, N., Shigematsu, K., Miyadai, T., & Kondo, R. (2009). Comparison of bacterial communities in fish farm sediments along an organic enrichment gradient. *Aquaculture, 287*(1), 107-113. doi:10.1016/j.aquaculture.2008.10.003

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators, 85*, 1044-1057. doi:10.1016/j.ecolind.2017.11.014

Keeley, N. B., Forrest, B. M., Crawford, C., & Macleod, C. K. (2012). Exploiting salmon farm benthic enrichment gradients to evaluate the regional performance of biotic indices and environmental indicators. *Ecological Indicators, 23*, 453-466. doi:10.1016/j.ecolind.2012.04.028

Keeley, N. B., Forrest, B. M., & Macleod, C. K. (2013). Novel observations of benthic enrichment in contrasting flow regimes with implications for marine farm monitoring and management. *Marine Pollution Bulletin, 66*(1-2), 105-116. doi:10.1016/j.marpolbul.2012.10.024

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *Plos One, 9*(1), e86175. doi:10.1371/journal.pone.0086175

Kendall, M. G. (1948). *Rank correlation methods*. Oxford: Griffin.

Kennedy, K., Hall, M. W., Lynch, M. D., Moreno-Hagelsieb, G., & Neufeld, J. D. (2014). Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Applied and Environmental Microbiology, 80*(18), 5717-5722. doi:10.1128/aem.01451-14

Kennedy, R., Arthur, W., & Keegan, B. F. (2011). Long-term trends in benthic habitat quality as determined by Multivariate AMBI and Infaunal Quality Index in relation to natural variability: A case study in Kinsale Harbour, south coast of Ireland. *Marine Pollution Bulletin, 62*(7), 1427-1436. doi:10.1016/j.marpolbul.2011.04.030

Klemetsen, T., Willassen, N., & Karlsen, C. (2019). Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis. *MicrobiologyOpen, 8*. doi:10.1002/mbo3.898

Kobayashi, M., Msangi, S., Batka, M., Vannuccini, S., Dey, M. M., & Anderson, J. L. (2015). Fish to 2030: The Role and Opportunity for Aquaculture. *Aquaculture economics & management, 19*(3), 282-300. doi:10.1080/13657305.2015.994240

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., & Ripley, B. D. (2018). R Package 'quantreg': Estimation and inference methods for models of conditional quantiles. *Version 5.88*.

Kondo, R., Shigematsu, K., & Butani, J. (2008). Rapid enumeration of sulphate-reducing bacteria from aquatic environments using real-time PCR. *Plankton and Benthos Research, 3*(3), 180-183. doi:10.3800/pbr.3.180

Kondo, R., Shigematsu, K., Kawahara, N., Okamura, T., Yoon, Y. H., Sakami, T., Yokoyama, H., & Koizumi, Y. (2012). Abundance of sulphate-reducing bacteria in fish farm sediments along the coast of Japan and South Korea. *Fisheries Science, 78*(1), 123-131. doi:10.1007/s12562-011-0439-3

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology, 79*(17), 5112-5120. doi:10.1128/aem.01043-13

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., & Kenkel, B. (2020). R package 'caret' Classification and Regression Training. *https://github.com/topepo/caret/*

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*, 363-374. doi:10.2307/2529786

Lanzén, A., Dahlgren, T. G., Bagi, A., & Hestetun, J. T. (2021). Benthic eDNA metabarcoding provides accurate assessments of impact from oil extraction, and ecological insights. *Ecological Indicators, 130*, 108064. doi:10.1016/j.ecolind.2021.108064

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology, 25*(17), 4392-4406. doi:10.1111/mec.13761

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2017). DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *Plos One, 12*(6), e0179443. doi:10.1371/journal.pone.0179443

Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial mandala for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology, 30*, 2969-2987. doi:10.1111/mec.15489

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lear, G., & Pochon, X. (2018). A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Marine Pollution Bulletin, 127*, 97-107. doi:10.1016/j.marpolbul.2017.11.042

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lejzerowicz, F., Pawlowski, J., Lear, G., Atalah, J., & Pochon, X. (2016). First evaluation of foraminiferal metabarcoding for monitoring environmental impact from an offshore oil drilling site. *Marine Environmental Research, 120*, 225-235. doi:10.1016/j.marenvres.2016.08.009

Laroche, O., Wood, S. A., Tremblay, L. A., Lear, G., Ellis, J. I., & Pochon, X. (2017). Metabarcoding monitoring analysis: the pros and cons of using co-extracted environmental DNA and RNA data to assess offshore oil production impacts on benthic communities. *PeerJ, 5*, e3347. doi:10.7717/peerj.3347

Le, J. T., Levin, L. A., Lejzerowicz, F., Cordier, T., Gooday, A. J., & Pawlowski, J. (2021). Scientific and budgetary trade-offs between morphological and molecular methods for deep-sea biodiversity assessment. *Integrated Environmental Assessment and Management*. doi:10.1002/ieam.4466

Lear, G., Dopheide, A., Ancion, P., & Lewis, G. D. (2011). A comparison of bacterial, ciliate and macroinvertebrate indicators of stream ecological health. *Aquatic Ecology, 45*(4), 517-527. doi:10.1007/s10452-011-9372-x

Lejzerowicz, F., Esling, P., & Pawlowski, J. (2014). Patchiness of deep-sea benthic Foraminifera across the Southern Ocean: Insights from high-throughput DNA sequencing. *Deep Sea Research Part II: Topical Studies in Oceanography, 108*, 17-26. doi:10.1016/j.dsr2.2014.07.018

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D., & Pawlowski, J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports, 5*, e13932 doi:10.1038/srep13932.

Lekang, K., Lanzén, A., Jonassen, I., Thompson, E., & Troedsson, C. (2020). Evaluation of a eukaryote phylogenetic microarray for environmental monitoring of marine sediments. *Marine Pollution Bulletin, 154*, 111102. doi:10.1016/j.marpolbul.2020.111102

Lepš, J., & Šmilauer, P. (2020). *Biostatistics with R: an introductory guide for field biologists*. Cambridge: Cambridge University Press.

Li, L., Kato, C., & Horikoshi, K. (1999). Microbial Diversity in Sediments Collected from the Deepest Cold-Seep Area, the Japan Trench. *Marine Biotechnology, 1*(4), 391-400. doi:10.1007/PL00011793

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R news, 2(3)*, 18-22.

Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. *Geoderma, 170*, 70-79. doi:10.1016/j.geoderma.2011.10.010

Lin, W.-J., Chiu, M.-C., Lin, C.-W., & Lin, H.-J. (2021). Effects of Sediment Characteristics on Carbon Dioxide Fluxes Based on Interacting Factors in Unvegetated Tidal Flats. *Frontiers in Marine Science, 8*. doi:10.3389/fmars.2021.670180

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature, 362*(6422), 709-715. doi:10.1038/362709a0

Lindh, M. V., Sjöstedt, J., Andersson, A. F., Baltar, F., Hugerth, L. W., Lundin, D., Muthusamy, S., Legrand, C., & Pinhassi, J. (2015). Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environmental Microbiology, 17*(7), 2459-2476. doi:10.1111/1462-2920.12720

Liu, S., Ren, H., Shen, L., Lou, L., Tian, G., Zheng, P., & Hu, B. (2015). pH levels drive bacterial community structure in sediments of the Qiantang River as determined by 454 pyrosequencing. *Frontiers in Microbiology, 6*, e285. doi:10.3389/fmicb.2015.00285

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260. doi:10.1002/edn3.112

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology, 16*(9), 2659-2671. doi:10.1111/1462-2920.12250

Lotze, H. K., Lenihan, H. S., Bourque, B. J., Bradbury, R. H., Cooke, R. G., Kay, M. C., et al. (2006). Depletion, degradation, and recovery potential of estuaries and coastal seas. *Science, 312*(5781), 1806-1809. doi:10.1126/science.1128035

Lu, Z., & Su, H. (2021). Employing gene chip technology for monitoring and assessing soil heavy metal pollution. *Environmental Monitoring and Assessment, 194*(1), 2. doi:10.1007/s10661-021-09650-6

Lundin, D., Severin, I., Logue, J. B., Östman, Ö., Andersson, A. F., & Lindström, E. S. (2012). Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity? *Environmental Microbiology Reports, 4*(3), 367-372. doi:10.1111/j.1758-2229.2012.00345.x

Macher, T.-H., Beermann, A. J., & Leese, F. (2021). TaxonTableTools: A comprehensive, platform-independent graphical user interface software to explore and visualise DNA metabarcoding data. *Molecular Ecology Resources, 21*(5), 1705-1714. doi:10.1111/1755-0998.13358

Madoni, P. (1994). A sludge biotic index (SBI) for the evaluation of the biological performance of activated sludge plants based on the microfauna analysis. *Water Research, 28*, 67-75. doi:10.1016/0043-1354(94)90120-1

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution, 25*(10), 574-582. doi:10.1016/j.tree.2010.06.016

Markert, B., Wappelhorst, O., Weckert, V., Herpin, U., Siewers, U., Friese, K., & Breulmann, G. (1999). The use of bioindicators for monitoring the heavy-metal status of the environment. *Journal of Radioanalytical and Nuclear Chemistry, 240*(2), 425-429. doi:10.1007/BF02349387

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

Martinez-Crego, B., Alcoverro, T., & Romero, J. (2010). Biotic indices for assessing the status of coastal waters: a review of strengths and weaknesses. *Journal of Environmental Monitoring, 12*(5), 1013-1028. doi:10.1039/b920937a

McCarthy, A., Chiang, E., Schmidt, M. L., & Denef, V. J. (2015). RNA Preservation Agents and Nucleic Acid Extraction Method Bias Perceived Bacterial Community Composition. *Plos One, 10*(3), e0121659. doi:10.1371/journal.pone.0121659

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal, 6*(3), 610-618. doi:10.1038/ismej.2011.139

Mcleod, E., Chmura, G. L., Bouillon, S., Salm, R., Björk, M., Duarte, C. M., Lovelock, C. E., Schlesinger, W. H., & Silliman, B. R. (2011). A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO2. *Frontiers in Ecology and the Environment, 9*(10), 552-560. doi:10.1890/110004

Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., et al. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment, 70*, 186-203. doi:10.1016/j.atmosenv.2012.11.060

Menchaca, I., Rodriguez, J. G., Borja, A., Belzunce, M., Franco, J., Garmendia, J., & Larreta, J. (2014). Determination of polychlorinated biphenyl and polycyclic aromatic hydrocarbon marine regional Sediment Quality Guidelines within the European Water Framework Directive. *Chemistry and Ecology, 30*(8), 693-700. doi:10.1080/02757540.2014.917175

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics, 11*(1), 31-46. doi:10.1038/nrg2626

Meyer, M., Stenzel, U., Myles, S., Prüfer, K., & Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research, 35*(15), e97. doi:10.1093/nar/gkm566

Millard, K., & Richardson, M. (2015). On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sensing, 7*(7), 8489-8515. doi:10.3390/rs70708489

Miller, M. E., Belote, R. T., Bowker, M. A., & Garman, S. L. (2011). Alternative states of a semiarid grassland ecosystem: implications for ecosystem services. *Ecosphere, 2*(5), 55. doi:10.1890/ES11-00027.1

Miller, M. L., & Auyong, J. (1991). Coastal zone tourism: A potent force affecting environment and society. *Marine Policy, 15*(2), 75-99. doi:10.1016/0308-597X(91)90008-Y

Mizrahi-Man, O., Davenport, E. R., & Gilad, Y. (2013). Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *Plos One, 8*(1), e53608. doi:10.1371/journal.pone.0053608

MoBio (2011). LifeGuard Soil Preservation Solution - Instruction Manual. *Retrieved on 01.01.2022 from https://www.qiagen.com/us/resources/download.aspx?id=982fd584-9776-4dff-9324-587291cfe0fb&lang=en*

Monroe, C., Grier, C., & Kemp, B. M. (2013). Evaluating the efficacy of various thermo-stable polymerases against co-extracted PCR inhibitors in ancient DNA samples. *Forensic Science International, 228*(1), 142-153. doi:10.1016/j.forsciint.2013.02.029

Morey, K. C., Bartley, T. J., & Hanner, R. H. (2020). Validating environmental DNA metabarcoding for marine fishes in diverse ecosystems using a public aquarium. *Environmental DNA, 2*(3), 330-342. doi:10.1002/edn3.76

Morgan, J. L., Darling, A. E., & Eisen, J. A. (2010). Metagenomic Sequencing of an In Vitro-Simulated Microbial Community. *Plos One, 5*(4), e10209. doi:10.1371/journal.pone.0010209

MSFD (2008). Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the Field of marine environmental policy. *Retrieved on 25.01.2021 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008L0056.*

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN computer science, 2*(1), 1-13. doi:10.1007/s42979-020-00394-7

Muxika, I., Borja, A., & Bald, J. (2007). Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. *Marine Pollution Bulletin, 55*(1-6), 16-29. doi:10.1016/j.marpolbul.2006.05.025

Muxika, I., Borja, A., & Bonne, W. (2005). The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. *Ecological Indicators, 5*(1), 19-31. doi:10.1016/j.ecolind.2004.08.004

Myers, S., & Smith, M. (2018). Impact of anthropogenic CO2 emissions on global human nutrition. *Nature Climate Change, 8*. doi:10.1038/s41558-018-0253-3

Nicholson, A., McIsaac, D., MacDonald, C., Gec, P., Mason, B. E., Rein, W., et al. (2020). An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines. *Environmental DNA, 2*(3), 343-349. doi:10.1002/edn3.81

Nielsen, K. L., Høgh, A. L., & Emmersen, J. (2006). DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Research, 34*(19), e133. doi:10.1093/nar/gkl714

Nielsen, K. M., Johnsen, P. J., Bensasson, D., & Daffonchio, D. (2007). Release and persistence of extracellular DNA in the environment. *Environmental Biosafety Research, 6*(1-2), 37-53. doi:10.1051/ebr:2007031

Nogales, B., Lanfranconi, M. P., Piña-Villalonga, J. M., & Bosch, R. (2011). Anthropogenic perturbations in marine microbial communities. *FEMS Microbiology Reviews, 35*(2), 275-298. doi:10.1111/j.1574-6976.2010.00248.x

Nybakken, J. W., & Bertness, M. (2004). *Marine biology: an ecological approach* (6th ed.). San Francisco: Pearson Education.

Odelberg, S. J., Weiss, R. B., Hata, A., & White, R. (1995). Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. *Nucleic Acids Research, 23*(11), 2049-2057. doi:10.1093/nar/23.11.2049

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2019). R Package 'vegan': Community Ecology Package. *Version 2.5-6.*

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2020). R Package 'vegan': Community Ecology Package. *Version 2.5-7.*

Olenin, S., Elliott, M., Bysveen, I., Culverhouse, P. F., Daunys, D., Dubelaar, G. B., et al. (2011). Recommendations on methods for the detection and control of biological pollution in marine coastal waters. *Marine Pollution Bulletin, 62*(12), 2598-2604. doi:10.1016/j.marpolbul.2011.08.011

Orr, J. C., Fabry, V. J., Aumont, O., Bopp, L., Doney, S. C., Feely, R. A., et al. (2005). Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature, 437*(7059), 681-686. doi:10.1038/nature04095

Osarogiagbon, A. U., Khan, F., Venkatesan, R., & Gillard, P. (2021). Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. *Process Safety and Environmental Protection, 147*, 367-384. doi:10.1016/j.psep.2020.09.038

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., & Aerts, H. J. W. L. (2015). Machine learning methods for quantitative radiomic biomarkers. *Scientific Reports, 5*(1), 1-11. doi:10.1038/srep13087

Parsons, T. R., Takahashi, M., & Hargrave, B. (1977). *Biological oceanographic processes* (2nd ed.). Oxford: Pergamon Press.

Parsons, T. R., Takahashi, M., & Hargrave, B. (1984). *Biological oceanographic processes* (3rd ed.). Oxford: Pergamon Press.

Pavlovska, M., Prekrasna, I., Parnikoza, I., & Dykyi, E. (2021). Soil Sample Preservation Strategy Affects the Microbial Community Structure. *Microbes and Environments, 36*(1), ME20134. doi:10.1264/jsme2.ME20134

Pawlowski, J., Bonin, A., Boyer, F., Cordier, T., & Taberlet, P. (2021). Environmental DNA for biomonitoring. *Molecular Ecology, 30*(13), 2931-2936. doi:10.1111/mec.16023

Pawlowski, J., Bruce, K., Panksep, K., Aguirre, F. I., Amalfitano, S., Apothéloz-Perret-Gentil, L., et al. (2022). Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. *Science of the Total Environment, 818*, 151783. doi:10.1016/j.scitotenv.2021.151783

Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources, 14*(6), 1129-1140. doi:10.1111/1755-0998.12261

Pawlowski, J., Esling, P., Lejzerowicz, F., Cordier, T., Visco, J. A., Martins, C. I. M., Kvalvik, A., Staven, K., & Cedhagen, T. (2016a). Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. *Aquaculture Environment Interactions, 8*, 371-386. doi:10.3354/aei00182

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (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

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., & Esling, P. (2016b). Protist metabarcoding and environmental biomonitoring: time for change. *European Journal of Protistology, 55*, 12-25. doi:10.1016/j.ejop.2016.02.003

Pearman, J. K., Keeley, N. B., Wood, S. A., Laroche, O., Zaiko, A., Thomson-Laing, G., Biessy, L., Atalah, J., & Pochon, X. (2020). Comparing sediment DNA extraction methods for assessing organic enrichment associated with marine aquaculture. *PeerJ, 8*, e10231. doi:10.7717/peerj.10231

Pearman, J. K., Thomson-Laing, G., Howarth, J. D., Vandergoes, M. J., Thompson, L., Rees, A., & Wood, S. A. (2021). Investigating variability in microbial community composition in replicate environmental DNA samples down lake sediment cores. *Plos One, 16*(5), e0250783. doi:10.1371/journal.pone.0250783

Pearson, K., & Henrici, O. M. F. E. (1896). VII. Mathematical contributions to the theory of evolution; III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, 187*, 253-318. doi:10.1098/rsta.1896.0007

Pearson, T., & Rosenberg, R. (1978). Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology, 16*, 229-311. doi:10.2983/035.034.0121u1.10

Phillips, G. R., Anwar, A., Brooks, L., Martina, L. J., Miles, A. C., & Prior, A. (2014). Infaunal quality index: Water Framework Directive classification scheme for marine benthic invertebrates. *Retrieved on 01.01.2022 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/314673/Water_Framework_Directive_classification_scheme_for_marine_benthic_invertebrates_-_report.pdf.*

Pochon, X., Wood, S. A., Keeley, N. B., Lejzerowicz, F., Esling, P., Drew, J., & Pawlowski, J. (2015). Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Marine Pollution Bulletin, 100*(1), 370-382. doi:10.1016/j.marpolbul.2015.08.022

Polinski, J. M., Bucci, J. P., Gasser, M., & Bodnar, A. G. (2019). Metabarcoding assessment of prokaryotic and eukaryotic taxa in sediments from Stellwagen Bank National Marine Sanctuary. *Scientific Reports, 9*(1), e14820. doi:10.1038/s41598-019-51341-3

Polymenakou, P. N., Bertilsson, S., Tselepides, A., & Stephanou, E. G. (2005). Bacterial Community Composition in Different Sediments from the Eastern Mediterranean Sea: a Comparison of Four 16S Ribosomal DNA Clone Libraries. *Microbial Ecology, 50*(3), 447-462. doi:10.1007/s00248-005-0005-6

Prasad, A., Iverson, L., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems, 9*, 181-199. doi:10.1007/s10021-005-0054-1

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *Plos One, 15*(1), e0227434. doi:10.1371/journal.pone.0227434

Prodinger, F., Endo, H., Takano, Y., Li, Y., Tominaga, K., Isozaki, T., et al. (2021). Year-round dynamics of amplicon sequence variant communities differ among eukaryotes, Imitervirales and prokaryotes in a coastal ecosystem. *FEMS microbiology ecology, 97*(12). doi:10.1101/2021.02.02.429489

Prosser, J. I., Bohannan, B. J., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., et al. (2007). The role of ecological theory in microbial ecology. *Nature Reviews Microbiology, 5*(5), 384-392. doi:10.1038/nrmicro1643

Puritz, J. B., & Toonen, R. J. (2011). Coastal pollution limits pelagic larval dispersal. *Nature Communications, 2*(1), 226. doi:10.1038/ncomms1238

Rachinger, N., Fischer, S., Böhme, I., Linck-Paulus, L., Kuphal, S., Kappelmann-Fenzl, M., & Bosserhoff, A. K. (2021). Loss of Gene Information: Discrepancies between RNA Sequencing, cDNA Microarray, and qRT-PCR. *International Journal of Molecular Sciences, 22*(17), e9349. doi:10.3390/ijms22179349

Reavie, E., Jicha, T., Angradi, T., Bolgrien, D., & Hill, B. (2010). Algal assemblages for large river monitoring: Comparison among biovolume, absolute and relative abundance metrics. *Ecological Indicators, 10*, 167-177. doi:10.1016/j.ecolind.2009.04.009

Reish, D. J. (1955). The relation of polychaetous Annelids to harbor pollution. *Public health reports, 70*(12), 1168-1174. doi:10.2307/4589315

Renaud, G., Stenzel, U., Maricic, T., Wiebe, V., & Kelso, J. (2015). deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics, 31*(5), 770-772. doi:10.1093/bioinformatics/btu719

Rissanen, A. J., Kurhela, E., Aho, T., Oittinen, T., & Tiirola, M. (2010). Storage of environmental samples for guaranteeing nucleic acid yields for molecular microbiological studies. *Applied Microbiology and Biotechnology, 88*(4), 977-984. doi:10.1007/s00253-010-2838-2

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., & Rimet, F. (2018). Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia, 807*(1), 37-51. doi:10.1007/s10750-017-3381-2

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ, 4*, e2584. doi:10.7717/peerj.2584

Roguet, A., Eren, A. M., Newton, R. J., & McLellan, S. L. (2018). Fecal source identification using random forest. *Microbiome, 6*(1), 185. doi:10.1186/s40168-018-0568-3

Romanazzi, V., Traversi, D., Lorenzi, E., & Gilli, G. (2015). Effects of freezing storage on the DNA extraction and microbial evaluation from anaerobic digested sludges. *BMC research notes, 8*, 420. doi:10.1186/s13104-015-1407-2

Rosenberg, R. (1985). Eutrophication - The future marine coastal nuisance? *Marine Pollution Bulletin, 16*(6), 227-231. doi:10.1016/0025-326X(85)90505-3

Rubin, B. E. R., Gibbons, S. M., Kennedy, S., Hampton-Marcell, J., Owens, S., & Gilbert, J. A. (2013). Investigating the Impact of Storage Conditions on Microbial Community Composition in Soil Samples. *Plos One, 8*(7), e70460. doi:10.1371/journal.pone.0070460

Rusch, A., Huettel, M., Reimers, C. E., Taghon, G. L., & Fuller, C. M. (2003). Activity and distribution of bacterial populations in Middle Atlantic Bight shelf sands. *Microbial Ecology, 44*(1), 89-100. doi:10.1111/j.1574-6941.2003.tb01093.x

Rygg, B., & Norling, K. (2013). Norwegian Sensitivity Index (NSI) for marine macroinvertebrates, and an update of Indicator Species Index (ISI). *Retrieved on 23.02.2022 from http://hdl.handle.net/11250/216238.*

Sarà, G., Scilipoti, D., Mazzola, A., & Modica, A. (2004). Effects of fish farming waste to sedimentary and particulate organic matter in a southern Mediterranean area (Gulf of Castellammare, Sicily): A multiple stable isotope study. *Aquaculture, 234*, 199-213. doi:10.1016/j.aquaculture.2003.11.020

Sarà, G., Scilipoti, D., Milazzo, M., & Modica, A. (2006). Use of stable isotopes to investigate dispersal of waste from fish farms as a function of hydrodynamics. *Marine Ecology Progress Series, 313*, 261-270. doi:10.3354/MEPS313261

Schander, C., & Willassen, E. (2005). What can biological barcoding do for marine biology? *Marine Biology Research, 1*(1), 79-83. doi:10.1080/17451000510018962

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio, 9*(3), e00525-00518. doi:10.1128/mBio.00525-18

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., & Chen, W. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America, 109*(16), 6241-6246. doi:10.1073/pnas.1117018109

SEPA (2018). Scottish Environmental Protection Agency. Fish farm survey report-evaluation of a new seabed monitoring approach to investigate the impacts of marine cage fish farms. *Retrived on 20.06.2020 from https://consultation.sepa.org.uk/sectorplan/finfishaquaculture/supporting_documents/Fish%20Farm%20Survey%20Report.pdf*

Serrana, J. M., Li, B., Sumi, T., Takemon, Y., & Watanabe, K. (2022). Implications of taxonomic and numerical resolution on DNA metabarcoding-based inference of benthic macroinvertebrate responses to river restoration. *Ecological Indicators, 135*, 108508. doi:10.1016/j.ecolind.2021.108508

Shahidul Islam, M., & Tanaka, M. (2004). Impacts of pollution on coastal and marine ecosystems including coastal and marine fisheries and approach for management: a review and synthesis. *Marine Pollution Bulletin, 48*(7), 624-649. doi:10.1016/j.marpolbul.2003.12.004

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Champaign: University of Illinois Press.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), 591-611. doi:10.1093/biomet/52.3-4.591

Singh, J., Gupta, M., Singh, K. K., Kumar, A., Yadav, D., Wenjing, W., & Singh, P. K. (2021). Chapter 18 - Advancement in bioinformatics and microarray-based technologies for genome sequence analysis and its application in bioremediation of soil and water pollutants. In A. Kumar, V. K. Singh, P. Singh, & V. K. Mishra (Eds.), *Microbe Mediated Remediation of Environmental Contaminants* (pp. 209-225). Sawston: Woodhead Publishing.

Skinner, M., Murdoch, M., Loeza-Quintana, T., Crookes, S., & Hanner, R. (2020). A mesocosm comparison of laboratory-based and on-site eDNA solutions for detection and quantification of striped bass (Morone saxatilis) in marine ecosystems. *Environmental DNA, 2*(3), 298-308. doi:10.1002/edn3.61

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current protocols in molecular biology, 122*(1), e59. doi:10.1002/cpmb.59

Small, C., & Nicholls, R. J. (2003). A Global Analysis of Human Settlement in Coastal Zones. *Journal of Coastal Research, 19*(3), 584-599. doi:10.2307/4299200

Smith, M., B., Rocha, A., M., Smillie, C., S., Olesen, S., W., Paradis, C., Wu, L., et al. (2015). Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. *mBio, 6*(3), e00326-00315. doi:10.1128/mBio.00326-15

Smucker, N. J., Pilgrim, E. M., Nietch, C. T., Darling, J. A., & Johnson, B. R. (2020). DNA metabarcoding effectively quantifies diatom responses to nutrients in streams. *Ecological Applications, 30*(8), e02205. doi:10.1002/eap.2205

Smyth, R. P., Schlub, T. E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., Davenport, M. P., & Mak, J. (2010). Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene, 469*(1), 45-51. doi:10.1016/j.gene.2010.08.009

Sokal, R., & Rohlf, F. (2012). *Biometry: the principles and practice of statistics in biological research* (2nd ed.). London: Macmillan Education.

Steyaert, M., Priestley, V., Osborne, O., Herraiz, A., Arnold, R., & Savolainen, V. (2020). Advances in metabarcoding techniques bring us closer to reliable monitoring of the marine benthos. *Journal of Applied Ecology, 57*(11), 2234-2245. doi:10.1111/1365-2664.13729

Stigebrandt, A. (1983). A model for the exchange of water and salt between the Baltic and the Skagerrak. *Journal of Physical Oceanography, 13*, 411-427. doi:10.1175/1520-0485(1983)013<0411:AMFTEO>2.0.CO;2

Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C. I. M., & Pawlowski, J. (2018a). Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. *Marine Pollution Bulletin, 127*, 139-149. doi:10.1016/j.marpolbul.2017.11.065

Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., & Pawlowski, J. (2018b). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecological Indicators, 85*, 153-164. doi:10.1016/j.ecolind.2017.10.041

Taberlet, P., Bonin, A., Zinger, L., & Coissac, É. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press.

Tait, K., Laverock, B., & Widdicombe, S. (2014). Response of an Arctic Sediment Nitrogen Cycling Community to Increased CO2. *Estuaries and Coasts, 37*(3), 724-735. doi:10.1007/s12237-013-9709-x

Tande, K. S. (1988). The effects of temperature on metabolic rates of different life stages of Calanus glacialis in the Barents Sea. *Polar Biology, 8*(6), 457-461. doi:10.1007/BF00264722

Tang, C., Garreau, D., & Luxburg, U. (2018). When do random forests fail? Conference on Neural Information Processing Systems. *Retrieved on 26.01.2021 from https://www.researchgate.net/publication/328229072_When_do_random_forests_fail.*

Tatangelo, V., Franzetti, A., Gandolfi, I., Bestetti, G., & Ambrosini, R. (2014). Effect of preservation method on the assessment of bacterial community structure in soil and water samples. *FEMS Microbiology Letters, 356*(1), 32-38. doi:10.1111/1574-6968.12475

Thomsen, P. F., Kielgast, J. O. S., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L., & Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology, 21*(11), 2565-2573. doi:10.1111/j.1365-294X.2011.05418.x

Thrush, S., Hewitt, J., Pilditch, C., & Norkko, A. (2021). *Ecology of Coastal Marine Sediments: Form, Function, and Change in the Anthropocene*. Oxford: Oxford University Press.

UKMMAS (2014). United Kingdom Marine Monitoring & Assessment Strategy UKMMAS. *Retrieved on 14.02.2022 from https://moat.cefas.co.uk/biodiversity-food-webs-and-marine-protected-areas/benthic-habitats/infaunal-quality-index/.*

UKTAG (2012). United Kingdom Technical Advisory Group. Practitioners guide to the infaunal quality index. Water Framework Directive: Transitional and Coastal Waters. . *Retrieved on 24.02.2022 from https://www.wfduk.org/sites/default/files/Media/Environmental%20standards/Annex%2018%20Transitional%20and%20coastal%20waters%20Invertebrates%20IQI.pdf.*

Vaalgamaa, S., Sonninen, E., Korhola, A., & Weckström, K. (2013). Identifying recent sources of organic matter enrichment and eutrophication trends at coastal sites using stable nitrogen and carbon isotope ratios in sediment cores. *Journal of Paleolimnology, 50*(2), 191-206. doi:10.1007/s10933-013-9713-y

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., et al. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

van de Velde, S., Van Lancker, V., Hidalgo-Martinez, S., Berelson, W. M., & Meysman, F. J. R. (2018). Anthropogenic disturbance keeps the coastal seafloor biogeochemistry in a transient state. *Scientific Reports, 8*(1), 5582. doi:10.1038/s41598-018-23925-y

Verhoeven, J. T. P., Salvo, F., Knight, R., Hamoutene, D., & Dufour, S. C. (2018). Temporal Bacterial Surveillance of Salmon Aquaculture Sites Indicates a Long Lasting Benthic Impact With Minimal Recovery. *Frontiers in Microbiology, 9*, e03054. doi:10.3389/fmicb.2018.03054

Vincent, S. G. T., Jennerjahn, T., & Ramasamy, K. (2021). Chapter 3 - Environmental variables and factors regulating microbial structure and functions. In S. G. T. Vincent, T. Jennerjahn, & K. Ramasamy (Eds.), *Microbial Communities in Coastal Sediments* (1st ed., pp. 79-117). Amsterdam, Oxford, Cambridge: Elsevier.

Wang, K., Zou, L., Lu, X., & Mou, X. (2018). Organic carbon source and salinity shape sediment bacterial composition in two China marginal seas and their major tributaries. *Science of the Total Environment, 633*, 1510-1517. doi:10.1016/j.scitotenv.2018.03.295

Weis, J. S. (2015). *Marine Pollution: What Everyone Needs to Know*. Oxford: Oxford University Press.

Wenqian, C., Meng, W., Zhu, Y., Zhou, J., & Liu, L. (2013). Assessing benthic ecological status in stressed Liaodong Bay (China) with AMBI and M-AMBI. *Chinese Journal of Oceanology and Limnology, 31*, 482-492. doi:10.1007/s00343-013-2177-0

WFD (2000). Water Framework Directive 2000/60/EC of the European Parliament and of The Council of 23 October 2000 establishing a framework for Community action in the field of water policy. *Official Journal of the European Commission, 327*, 1-73.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). New York: Springer.

Wickramasinghe, M. P., Sudarshani, K. A. M., & Wegiriya, H. C. E. (2021). The diversity of marine invertebrate macrofauna in selected rocky intertidal zones of Matara, Sri Lanka. *Asian Journal of Conservation Biology, 10*(1), 15-21. doi:10.53562/ajcb.OZDK5526

Wilhelm, R. C., van Es, H. M., & Buckley, D. H. (2022). Predicting measures of soil health using the microbiome and supervised machine learning. *Soil Biology and Biochemistry, 164*, 108472. doi:10.1016/j.soilbio.2021.108472

Word, J. Q. (1978). The infaunal trophic index. *Southern California Coastal Water Research Project Annual Report. El Segundo*, 19-40.

Wu, P., Christidis, N., & Stott, P. (2013). Anthropogenic impact on Earth's hydrological cycle. *Nature Climate Change, 3*(9), 807-810. doi:10.1038/nclimate1932

Yakimov, M. M., Timmis, K. N., & Golyshin, P. N. (2007). Obligate oil-degrading marine bacteria. *Current opinion in biotechnology, 18*(3), 257-266. doi:10.1016/j.copbio.2007.04.006

Yamaguchi, N., Ichijo, T., Sakotani, A., Baba, T., & Nasu, M. (2012). Global dispersion of bacterial cells on Asian dust. *Scientific Reports, 2*, 525. doi:10.1038/srep00525

Yang, J., Ma, L. a., Jiang, H., Wu, G., & Dong, H. (2016). Salinity shapes microbial diversity and community structure in surface sediments of the Qinghai-Tibetan Lakes. *Scientific Reports, 6*(1), 25078. doi:10.1038/srep25078

Yin, X., Wang, W., Wang, A., He, M., Lin, C., Ouyang, W., & Liu, X. (2022). Microbial community structure and metabolic potential in the coastal sediments around the Yellow River Estuary. *Science of the Total Environment, 816*, e151582. doi:10.1016/j.scitotenv.2021.151582

Yoon, S. J., Hong, S., Kim, S., Lee, J., Kim, T., Kim, B., et al. (2020). Large-scale monitoring and ecological risk assessment of persistent toxic substances in riverine, estuarine, and coastal sediments of the Yellow and Bohai seas. *Environment International, 137*, e105517. doi:10.1016/j.envint.2020.105517

Yukgehnaish, K., Kumar, P., Sivachandran, P., Marimuthu, K., Arshad, A., Paray, B. A., & Arockiaraj, J. (2020). Gut microbiota metagenomics in aquaculture: factors influencing gut microbiome and its physiological role in fish. *Reviews in Aquaculture, 12*(3), 1903-1927. doi:10.1111/raq.12416

Zaiko, A., Greenfield, P., Abbott, C., von Ammon, U., Bilewitch, J., Bunce, M., et al. (2022). Towards reproducible metabarcoding data: Lessons from an international cross-laboratory experiment. *Molecular Ecology Resources, 22*(2), 519-538. doi:10.1111/1755-0998.13485

Zhang, J., Zhang, Z. F., Liu, S. M., Wu, Y., Xiong, H., & Chen, H. T. (1999). Human impacts on the large world rivers: Would the Changjiang (Yangtze River) be an illustration? *Global Biogeochemical Cycles, 13*(4), 1099-1105. doi:10.1029/1999GB900044

Zheng, B., Wang, L., & Liu, L. (2014). Bacterial community structure and its regulating factors in the intertidal sediment along the Liaodong Bay of Bohai Sea, China. *Microbiological Research, 169*(7), 585-592. doi:10.1016/j.micres.2013.09.019

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., et al. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology, 28*(8), 1857-1862. doi:10.1111/mec.15060

Zinger, L., Lionnet, C., Benoiston, A.-S., Donald, J., Mercier, C., & Boyer, F. (2021). metabaR: An r package for the evaluation and improvement of DNA metabarcoding data quality. *Methods in Ecology and Evolution, 12*(4), 586-592. doi:10.1111/2041-210X.13552

Zylstra, P., Rothenfluh, H. S., Weiller, G. F., Blanden, R. V., & Steele, E. J. (1998). PCR amplification of murine immunoglobulin germline V genes: strategies for minimization of recombination artefacts. *Immunology and Cell Biology, 76*(5), 395-405. doi:10.1046/j.1440-1711.1998.00772.x

SUMMARY

Marine coastal ecosystems are exposed to a variety of anthropogenic impacts, which often manifest themselves in the pollution of the surrounding ecosystem. Especially on densely populated coasts or in regions heavily used for aquaculture, changes in the natural marine habitat can be observed. In order to protect nature and thus its ecosystem services for humans, more and more environmental protection laws are coming into force. Exemplary, operators of facilities known to contribute to pollution are obliged to regularly monitor the condition of the surrounding environment. The purpose of such so-called compliance monitoring is to determine whether the prescribed regulations are being followed. The traditional routine involves sampling by ship, during which sediment samples are taken from the seabed below the aquaculture cages and all macrofauna organisms found, such as mussels or worms, are taxonomically determined and quantified by experts. Based on the community of organisms the ecological status of the sample can then be inferred. Since this method is very labor- and time-consuming, a reorientation of the scientific community towards alternative monitoring methods is currently taking place. A bacteria-based eDNA (environmental DNA) metabarcoding system in particular has proven to be a suitable monitoring tool. With this molecular method, the composition of the benthic bacterial community is determined using high-throughput sequencing. The great advantage of this method is that bacteria, due to their short generation times, react rapidly to various environmental influences. The composition of this community can then be used to infer the ecological status of the sample under investigation via sequencing without the need for laborious enumeration and identification of organisms. Additionally, sequencing costs are more and more decreasing, proposing eDNA metabarcoding-based monitoring as a faster and cheaper alternative to traditional monitoring. In order to implement the method in legislation in the long term, standard protocols need to be developed. Once these are sufficiently validated, the novel methodology can be incorporated into regulations to support or even replace traditional monitoring. However, some steps of the eDNA metabarcoding method, from sampling to ecosystem assessment, are not yet sufficiently standardized, which is why the development of this work was necessary.

Since there is no consensus in the scientific community on (i) the preservation of environmental samples during transport, (ii) the reproducibility of ecosystem assessment among different laboratories, (iii) the most appropriate bioinformatic method for ecosystem assessment, and (iv) the minimum sequencing depth required to determine ecosystem status, these sub-steps were investigated. It was found that the most common methods currently used to preserve samples during transport had no discernible effect on the final ecosystem assessment. Furthermore, sample processing in independent laboratories allowed the same ecological interpretations based on the bacterial community, which resulted in concordant ecosystem assessments among laboratories. This indicates the overall reproducibility of the eDNA metabarcoding-based method, thus enabling its implementation in standard protocols. Furthermore, it was shown that corresponding ecosystem assessments can be obtained with the currently used methods for determining ecological status based on eDNA data. Critical to predictive accuracy is not the method itself, but a sufficient number of samples that accounts for the natural spatial and temporal variability of bacterial communities. It was demonstrated that a very shallow sequencing depth per sample can be sufficient to use machine learning to prediction the ecological status of the environmental sample. The quality of this classifications did not depend on the sequencing depth as assumed but was determined by the separability of individual categories. The results and recommendations of this work contribute directly to the standardization of ecological assessment of nearshore marine ecosystems. By establishing these standard protocols, it will be possible to integrate the eDNA metabarcoding-based method for monitoring compliance of coastal marine ecosystems into legislative regulations in the future.

ZUSAMMENFASSUNG

Marine Küstenökosysteme sind einer Vielzahl von durch den Menschen verursachten Einflüssen ausgesetzt. Diese äußern sich häufig in einer Verschmutzung des umgebenden Ökosystems. Vor allem an dicht besiedelten Küsten oder in stark für die Aquakultur genutzten Regionen ist eine Veränderung des natürlichen Lebensraums zu beobachten. Um die Natur und damit ihre Ökosystemleistungen für den Menschen zu schützen, treten immer mehr Naturschutzgesetze in Kraft. Beispielsweise sind Betreiber von Aquakulturanlagen dazu verpflichtet, den Zustand der Umgebung regelmäßig zu prüfen. Dieses so genannte Compliance Monitoring hat den Zweck, festzustellen, ob die geltenden Vorschriften eingehalten werden. Die traditionelle Art dieser Überwachung ist sehr arbeits- und zeitintensiv. Mit Hilfe eines Schiffes werden Sedimentproben unterhalb der Aquakulturkäfige entnommen und alle darin vorkommenden Makrofauna-Organismen, wie Muscheln oder Würmer, von Experten bestimmt und quantifiziert. Anhand der Organismengemeinschaft kann dann auf den ökologischen Zustand der Probe geschlossen werden. Da dieses Verfahren sehr zeitaufwendig ist, findet derzeit eine Umorientierung der wissenschaftlichen Gemeinschaft hin zu alternativen Überwachungsmethoden statt. In der Zwischenzeit hat sich ein auf bakterieller DNA basierendes eDNA (engl. environmental DNA, Umwelt-DNA) Metabarcoding-System als besonders nützliches Instrument erwiesen. Bei dieser molekularen Methode kann die Zusammensetzung der Bakteriengemeinschaft mit Hilfe von Hochdurchsatz-Sequenzierung bestimmt werden. Aus der Zusammensetzung der Bakteriengemeinschaft kann dann auf den ökologischen Status der untersuchten Probe geschlossen werden, ohne dass eine aufwändige Auszählung und Identifizierung der größeren Organismen erforderlich ist. Da Bakterien eine deutlich kürzere Generationszeit als die bisher untersuchten Organismen besitzen, können auch kurzfristige Änderungen von Umwelteinflüssen anhand der Bakterienzusammensetzung detektiert werden. Des Weiteren wird die Anwendung von eDNA Metabarcoding-basiertem Monitoring stetig preiswerter, da die Sequenzierungskosten immer weiter sinken. Um die Methode jedoch langfristig in der Gesetzgebung zu etablieren, müssen zunächst Standardprotokolle entwickelt werden. Sobald diese ausreichend validiert sind, können sie in die Gesetzgebung aufgenommen werden und die herkömmliche Überwachung ergänzen oder sogar ersetzen.

Einige Schritte der eDNA-basierten Methode, von der Probenahme bis zur Ökosystembewertung, sind jedoch noch nicht ausreichend standardisiert, weshalb die Ausarbeitung dieser Arbeit notwendig war. Da es in der wissenschaftlichen Gemeinschaft noch keinen Konsens über (i) die Konservierung von Umweltproben während des Transports, (ii) die Reproduzierbarkeit der Ökosystembewertung durch verschiedene Labore, (iii) die am besten geeignete Methode zur Ableitung der eDNA Daten zur Bewertung von Ökosystemen, und (iv) die für die Bestimmung des Ökosystemstatus erforderliche Mindestsequenziertiefe gibt, wurden diese Teilschritte untersucht. Es wurde gezeigt, dass die derzeit am häufigsten verwendeten Methoden zur Probenkonservierung während des Transports keinen erkennbaren Einfluss auf die endgültige Ökosystembewertung haben. Darüber hinaus ermöglichten unabhängigen Labore die gleiche Interpretation der Bakteriengemeinschaft, was ebenfalls zu einer übereinstimmenden Bewertung des Ökosystems führte. Dies deutet auf die Reproduzierbarkeit der eDNA Metabarcoding-basierenden Methode hin, und erlaubt somit die Implementierung in Standardprotokolle. Darüber hinaus wurde gezeigt, dass mit den derzeit verwendeten Methoden zur Bestimmung des ökologischen Zustands auf der Grundlage von eDNA-Daten die gleichen Ökosystembewertungen erzielt werden können. Entscheidend für die Vorhersagegenauigkeit ist jedoch nicht die Methode selbst, sondern eine ausreichende Anzahl von Proben, die die natürliche räumliche und zeitliche Variabilität von Bakteriengemeinschaften berücksichtigt. Es konnte gezeigt werden, dass eine geringe Sequenziertiefe pro Probe bereits ausreichend war, um mit Hilfe des maschinellen Lernens eine Aussage über den ökologischen Status der Umweltprobe zu treffen. Die Qualität dieser Klassifizierung hing nicht, wie angenommen, von der Sequenzierungstiefe ab, sondern wurde durch die Auftrennbarkeit einzelner Kategorien vorgegeben. Die Ergebnisse und Empfehlungen dieser Arbeit tragen direkt zur Standardisierung der ökologischen Bewertung von küstennahen Meeresökosystemen anhand von eDNA Metabarcoding bei. Durch die Etablierung dieser Standardprotokolle wird es möglich sein, die auf eDNA Metabarcoding basierende Methode zur Überwachung der Einhaltung der Vorschriften für marine Küstenökosysteme in Zukunft in die Gesetzgebung zu integrieren.

APPENDIX

Statement of contributions

Chapter I

Publication: Dully, V., Rech, G., Wilding, T.A., Lanzén, A., MacKichan, K., Berrill, I, & Stoeck, T. (2021). Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems. *Marine Pollution Bulletin 173*, e113129. doi:10.1016/j.marpolbul.2021.113129

Statement: The contribution to this publication includes management and bioinformatic processing of raw and curated data including public provision, creation of all figures, sequence data processing, analysis of sequence quality, alpha diversity, beta diversity, ASV and taxa distribution, and editing and reviewing of the manuscript.

Chapter II

Publication: Dully, V., Balliet, H., Frühe, L., Däumer, M., Thielen, A., Gallie, S., Berrill, I., & Stoeck, T. (2021). Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter laboratory study. *Ecological Indicators 121*, e107049. doi:10.1016/j.ecolind.2020.107049

Statement: The contribution to this publication includes management of raw and curated data including public provision, bioinformatic processing of data, creation and conceptualization of figures, analysis of alpha diversity, beta diversity, taxonomic profiling, conduction of supervised machine learning, and editing and reviewing of the manuscript.

Chapter III (unpublished)

Statement: I was primarily responsible for data curation. Editing and improving the background and results of the manuscript were my responsibilities. The summary and the discussion, including outlook and conclusion, were written by me. I created or improved Figures 13-15, Table 1, and Supplementary Files 3.7- 3.11. All SML analyses were carried out and interpreted by me. The QRS analysis and the preparation of further figures were conducted by Dr. Kleopatra Leontidou.

Chapter IV

Publication: Dully, V., Wilding, T. A., Mühlhaus, T., & Stoeck, T. (2021). Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning. *Computational and Structural Biotechnology Journal 19*, 2256-2268. doi:10.1016/j.csbj.2021.04.005

Statement: My contribution to this publication includes compilation of datasets from previous publication, reprocessing of published datasets, support of conceptualization for the downsampling method, creation and conceptualization of figures, all supervised machine learning analyses, and editing and reviewing of the manuscript

List of figures, tables, and equations

Introduction

Eq. 1) Equation for the calculation of the AMBI following Borja et al., 2000
Figure 1) Schematic representation of the eDNA bioassessment method
Figure 2) Example of a genetic metabarcoding region
Figure 3) Schematic representation of bacterial inventories mirroring aquaculture impact in open net-pen systems
Figure 4) Schematic representation of the RF algorithm
Figure 5) Schematic representation of the k-fold CV approach

Chapter I

Figure 6) ASV richness comparison between non-treated and treated aliquots
Figure 7) Venn diagrams of the shared ASV among non-treated and treated aliquots at the different sites

Chapter II

Figure 8) Schematic representation of the conducted inter-laboratory study
Figure 9) Schematic sketch of the procedure of the applied SML algorithm
Figure 10) Alpha diversity measures among different laboratories
Figure 11) Shared ASVs among technical and biological replicates
Figure 12) Predicted IQI using SML among independent laboratories

Chapter III

Table 1) Classification of IQI values into EGs used for QRS indicator inference
Eq. 2) Equation for the calculation of the IQI version VI from Phillips et al., 2014
Figure 13) Sample EG distribution inferred by IQI values
Figure 14) Indicator EG distribution inferred by QRS
Figure 15) Linear regression plots comparing IQI values
Figure 16) Erroneously classified samples and their respective biotic indices
Figure 17) RF top 20 indicator ASVs per dataset

Chapter IV

Figure 18) A schematic representation of the RF downsampling process
Figure 19) Non-metric multidimensional scaling of the eDNA ‘ScoSa’ dataset

Supplementary files

Chapter I

Supplementary File 1.1)	Proportions of assigned ASVs per taxonomic level and site
Supplementary File 1.2)	Figure of rarefaction curves for the sites DUN, LIS and BASQUE
Supplementary File 1.3)	Quantitative DNA extraction efficiency among preservation methods

Chapter II

Supplementary File 2.1)	Class-level taxonomic bar chart
Supplementary File 2.2)	Family-level taxonomic bar chart

Chapter III

Supplementary File 3.1)	Table of HQ sequences along bioinformatic processing
Supplementary File 3.2)	Figure containing rarefaction curves for the new samples
Supplementary File 3.3)	Table of ASVs identified as indicators by the QRS approach for the Norwegian dataset
Supplementary File 3.4)	Table of ASVs identified as indicators by the QRS approach for the Scottish dataset
Supplementary File 3.5)	QRS plots showing the response of selected ASVs from EGs I to V for the Norwegian salmon farms
Supplementary File 3.6)	QRS plots showing the response of selected ASVs from EGs I to V for the Scottish salmon farms
Supplementary File 3.7)	Table of the top 20 indicator ASV inferred by RF
Supplementary File 3.8)	RF classification accuracies
Supplementary File 3.9)	RF classification results
Supplementary File 3.10)	Shared QRS indicators among datasets
Supplementary File 3.11)	QRS plot of ASV_000011

Chapter IV

Supplementary File 4.1)	Table of samples among the salmon farm production phases
Supplementary File 4.2)	Details of model construction
Supplementary File 4.3)	Frequencies and distribution of sample categorization
Supplementary File 4.4)	Visual representation of separability via ordination analysis
Supplementary File 4.5)	Visual representation of feature specificity via Venn diagrams
Supplementary File 4.6)	Visual representation of feature coefficient of variation

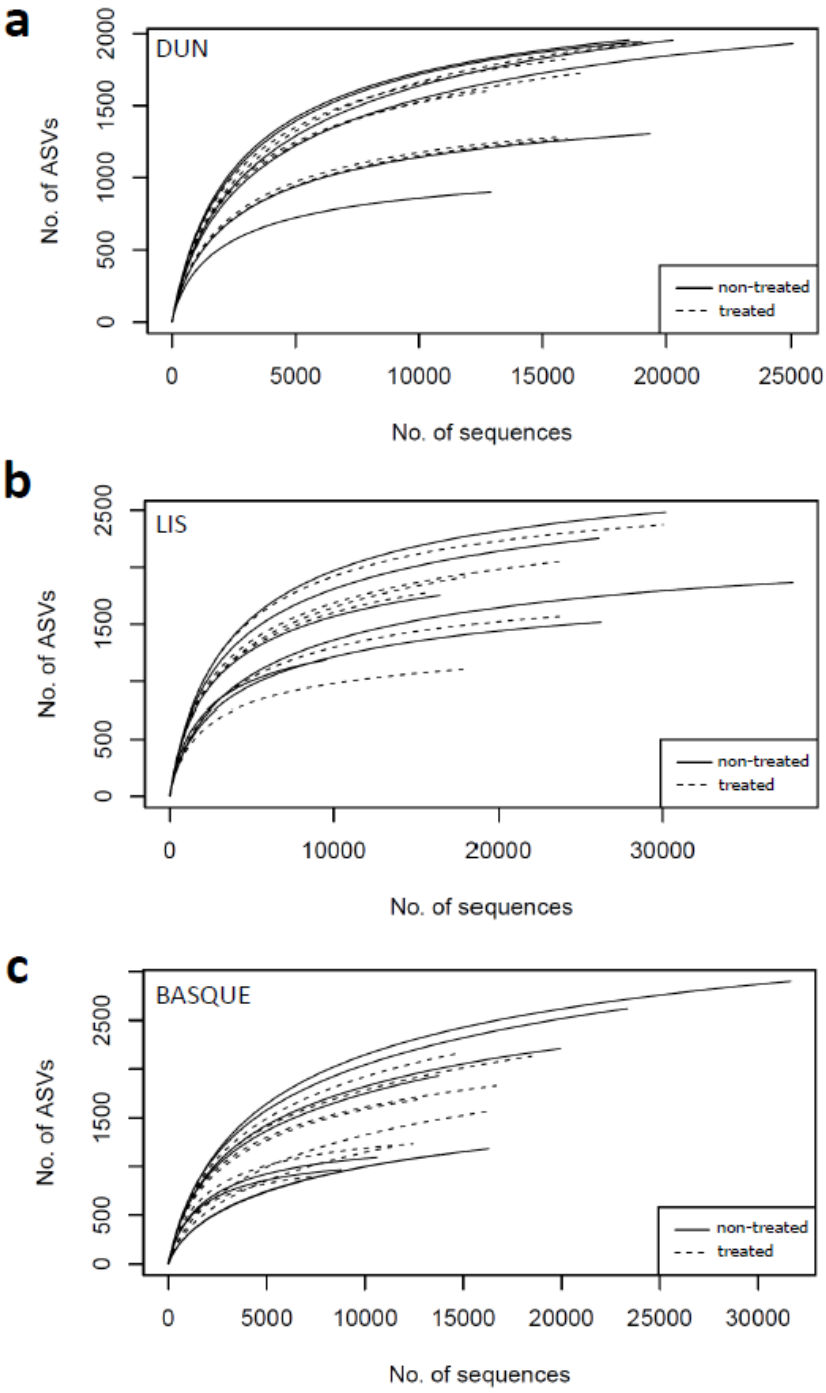
Supplementary files of Chapter I

Supplementary File 1.1)

Supplementary File 1.1) Proportions of assigned ASVs per taxonomic level and site. *The table represents the proportions of assigned ASVs for the sites DUN, LIS and BASQUE per taxonomic level based on vsearch's syntax function assigning sequences deposited in the greengenes database.*

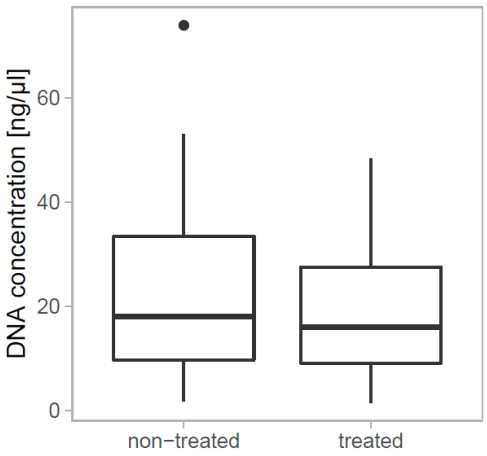
Taxnomic level	DUN assigned ASVs (%)	LIS assigned ASVs (%)	BASQUE assigned ASVs (%)
Phylum	99.68	99.55	99.93
Class	97.95	96.96	99.10
Order	86.85	85.41	89.81
Family	62.20	58.21	67.48
Genus	11.44	9.65	11.06
Species	00.02	0.00	00.19

Supplementary File 1.2)



Supplementary File 1.2) Figure of rarefaction curves for the sites DUN, LIS and BASQUE. *The number of sequences per number of obtained ASVs represented. The dashed lines indicate sample aliquots which were treated with LifeGuard® preservation solution, the solid lines indicate aliquots which were frozen without further treatment.*

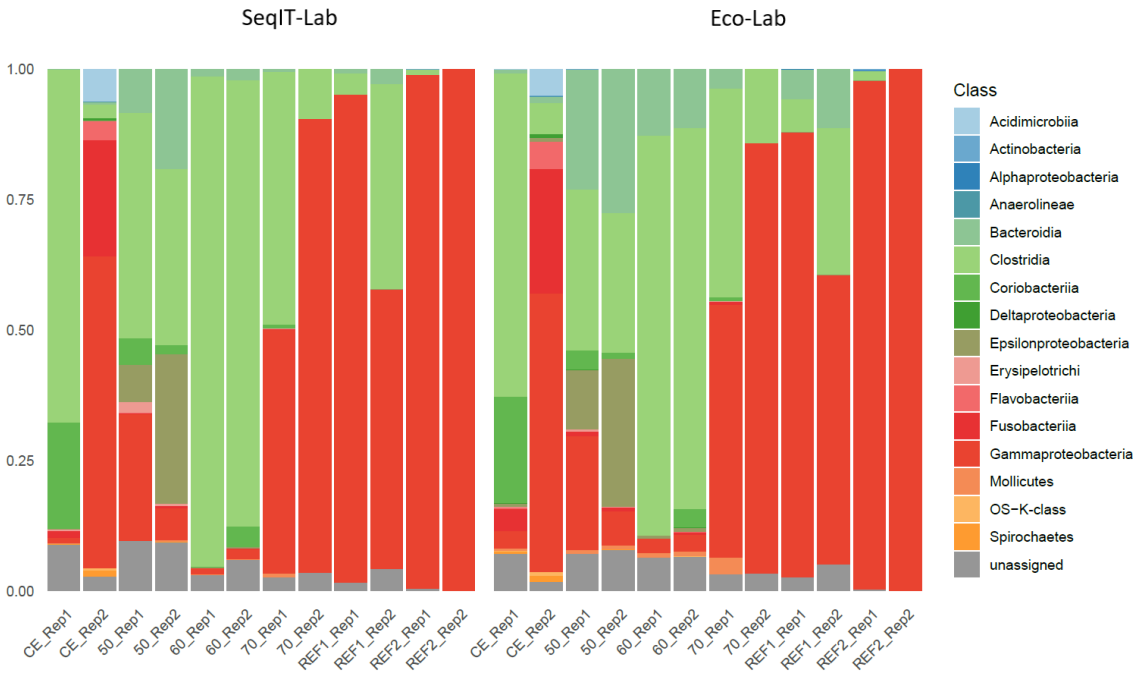
Supplementary File 1.3)



Supplementary File 1.3) Quantitative DNA extraction efficiency among preservation methods. *Figure of quantitative DNA extraction efficiency showing post-extraction DNA concentration among non-treated and treated aliquots. Pairwise testing of aliquot DNA concentration for aliquots deriving from the same sample was conducted using a pairwise Wilcoxon test. The test did not reveal significant differences among the preservation strategies (p = 0.1).*

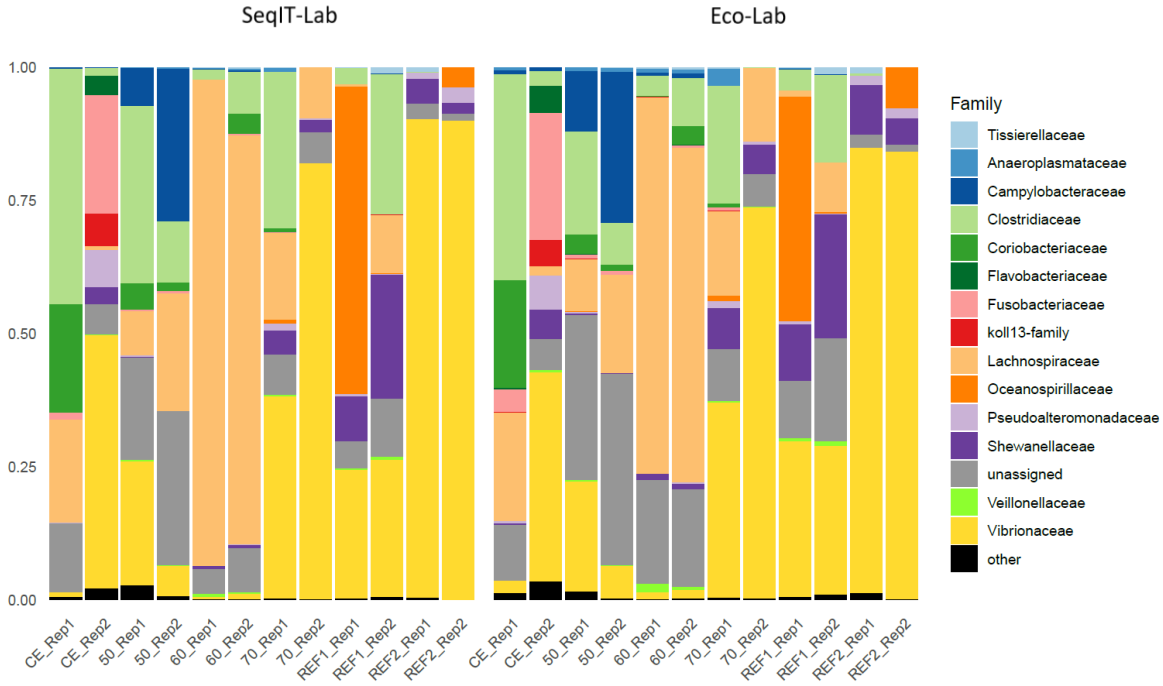
Supplementary files of Chapter II

Supplementary File 2.1)



Supplementary File 2.1) Class-level taxonomic bar chart. *Figure of taxonomic assignment of bacterial ASVs. The bars show the relative proportion of ASVs assigned to each of the different taxonomic entities at class level.*

Supplementary File 2.2)



Supplementary File 2.2) Family-level taxonomic bar chart. *Figure of taxonomic assignment of bacterial ASVs. The bars show the relative proportion of ASVs assigned to each of the different taxonomic entities at family level.*

Supplementary files of Chapter III

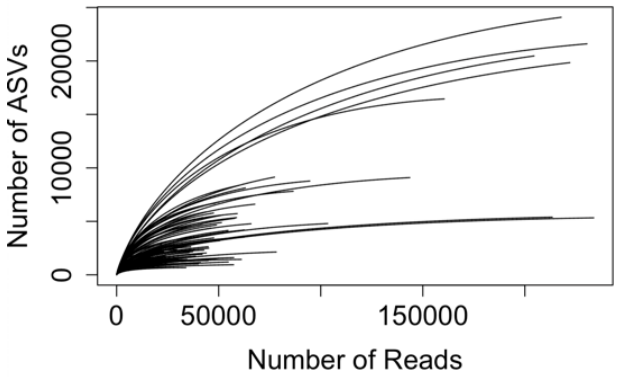
Supplementary File 3.1)

Supplementary File 3.1) Table of HQ sequences along bioinformatic processing. *Table of sequences per samples for the novel dataset obtained from five Scottish salmon farms. Each farm is represented by one row. The number of unprocessed sequences after provisioning by the sequencing facility is indicated in the column 'raw'. Sequences were then processed in the DADA2 workflow, resulting in decreasing numbers at each step. Sequence counts after length- and quality filtering are provided in the column 'filtered'. After merging of forward and reverse reads (column 'merged'), chimera check was conducted, and chimeras were removed (column 'nochim'). The column 'HQ reads' indicates the HQ read counts after the workflow which are clustered into ASVs by the DADA2 algorithm.*

sample	raw	filtered	merged	nochim	HQ reads	HQ ASVs
FIS_0_RepA	177759	79883	44095	14156	13510	1553
FIS_0_RepB	174760	84622	52812	23768	23108	1674
FIS_25SE_RepA	170545	91408	72217	45339	44886	1333
FIS_25SE_RepB	219565	108108	76948	43672	42824	1397
FIS_37_RepA	154370	83059	63366	37426	36805	1662
FIS_37_RepB	184450	99386	67945	32747	31818	2221
FIS_47_RepA	186443	105421	70540	35174	34426	2151
FIS_47_RepB	169582	90794	60004	31616	31187	1668
FIS_50NE_RepA	232459	92687	47765	19368	18666	1655
FIS_50NE_RepB	187434	88069	72119	51308	50551	1433
FIS_50SE_RepA	218239	105271	93882	63999	62617	1492
FIS_50SE_RepB	172111	89709	71253	45544	45158	1294
FIS_57_RepA	167124	88526	71880	52624	51389	2438
FIS_57_RepB	161790	80827	65299	48241	47601	1649
FIS_Ref1_RepA	145936	78175	77689	52481	51298	2444
FIS_Ref1_RepB	156368	87825	374364	234180	217793	2633
FIS_Ref2_RepA	223869	93444	66316	49012	48504	1540
FIS_Ref2_RepB	743998	436281	74406	55176	54618	2017
KIN_0T1_RepA	174719	73259	56096	34332	33954	634
KIN_0T1_RepB	227171	92095	74708	45544	44871	870
KIN_0T2_RepA	200723	95789	80235	44715	43946	1010
KIN_0T2_RepB	192646	71656	60654	35006	34620	994
KIN_17T2_RepA	178010	86816	72085	36356	35466	1529
KIN_17T2_RepB	185368	94438	71712	34929	34033	1681
KIN_27T2_RepB	164822	77238	62377	29431	28499	1610
KIN_37T2_RepA	177249	97910	82618	57737	57312	554
KIN_37T2_RepB	223452	79588	60371	40379	39933	846
KIN_50T1_RepA	210316	81062	56334	29650	29225	991
KIN_50T1_RepB	277831	96554	69292	33578	32636	1372
KIN_60T1_RepA	166872	84967	58309	29097	28632	1003
KIN_60T1_RepB	247516	93616	56722	20038	19457	1273
KIN_70T1_RepA	193828	98923	79525	45498	44767	664
KIN_70T1_RepB	245621	95095	82574	61996	61130	768
KIN_Ref1_RepA	159286	59021	38606	18270	17960	900
KIN_Ref1_RepB	803469	423790	356260	177443	160518	2573

KIN_Ref2_RepA	186657	102766	92796	69934	67726	2532
KIN_Ref2_RepB	202366	109596	96280	67341	65865	2063
MAN_0_RepA	369710	359490	325627	216075	213411	1303
MAN_0_RepB	597490	383201	358321	236238	233734	1252
MAN_25N_RepA	226823	145865	117298	63786	62560	1435
MAN_25N_RepB	616921	219025	185682	105294	103430	1474
MAN_25S_RepA	777639	112798	71728	40920	39887	2736
MAN_25S_RepB	478060	247949	173170	89801	86523	2896
MAN_50N_RepA	551960	336803	253305	147135	143651	2603
MAN_50N_RepB	747964	90590	61065	37152	36352	2000
MAN_50S_RepA	534681	168182	109807	60527	58789	2998
MAN_50S_RepB	346921	253183	173023	97953	94676	3295
MAN_75S_RepA	774034	163255	105312	60906	59108	3389
MAN_75S_RepB	865605	77452	51266	25761	25121	2006
NOS_0_RepA	189725	86567	63520	35057	34521	1131
NOS_0_RepB	194786	85147	62211	32525	31822	1218
NOS_44_RepA	183838	99804	81931	55128	54747	573
NOS_44_RepB	189684	100550	77229	48844	48189	782
NOS_54_RepA	200683	110440	92513	58044	57380	820
NOS_54_RepB	218369	135074	115025	78957	78199	809
NOS_64_RepA	203245	100948	71325	39458	38736	1406
NOS_64_RepB	224232	114760	77719	42715	41931	1133
NOS_Start2_RepA	158583	72396	59237	41489	41098	615
SCA_0_RepA	193069	96170	78296	49906	49249	703
SCA_0_RepB	175274	75690	56263	29960	29520	1133
SCA_100SE_RepA	235993	95860	59692	22975	21913	2099
SCA_100SE_RepB	989222	539838	433073	219880	204526	2187
SCA_37_RepA	175557	92421	69689	37827	37327	1536
SCA_37_RepB	214065	114614	94586	60005	57961	2474
SCA_47_RepA	167267	81549	55695	23656	23012	1430
SCA_47_RepB	195925	99892	76121	41204	40689	951
SCA_57_RepA	178039	93603	74771	46979	46203	2063
SCA_57_RepB	161939	107188	94969	65111	63087	2629
SCA_75SE_RepA	184333	96432	63354	27690	26891	2000
SCA_75SE_RepB	224784	84164	53752	24666	24119	1294
SCA_Ref1_RepA	240498	129674	115443	79566	77389	2992
SCA_Ref1_RepB	589352	352140	324213	232674	221988	2223
SCA_Ref2_RepA	163616	90250	74579	49746	48964	2246
SCA_Ref2_RepB	144221	79224	61443	39613	38575	3640

Supplementary File 3.2)



Supplementary Figure 3.2) Figure containing rarefaction curves for the new samples. *Rarefaction curves for the 74 Scottish samples newly sequenced for this study showing the number of sequence reads against the number of ASVs.*

Supplementary File 3.3)

Supplementary File 3.3) Table of ASVs identified as indicators by the QRS approach for the Norwegian dataset. *For each ASV, the highest resolution of taxonomic assignment is indicated in the column 'Taxon'. The inferred Eco-Group according to the QRS calculations is indicated at the column 'Eco-Group'. The relative abundance of the ASV among all samples is indicated in the column 'Relative abundance'.*

ASV	Taxon	Eco-Group	Relative abundance
ASV_000031	Myxococcales	III	1.03
ASV_000024	Alteromonadales	IV	0.92
ASV_000043	Myxococcales	III	0.9
ASV_000040	Myxococcales	II	0.86
ASV_000016	Helicobacteraceae	IV	0.83
ASV_000013	Helicobacteraceae	IV	0.72
ASV_000022	Helicobacteraceae	IV	0.54
ASV_000100	*Psychromonas*	IV	0.51
ASV_000054	Syntrophobacteraceae	II	0.49
ASV_000039	Bacteroidales	IV	0.48
ASV_000036	Bacteroidales	IV	0.46
ASV_000038	Thiohalorhabdales	IV	0.46
ASV_000063	Rhodospirillales	II	0.45
ASV_000055	Alteromonadales	IV	0.41
ASV_000126	*Nitrospina*	II	0.4
ASV_000076	Desulfobulbaceae	IV	0.36
ASV_000106	Myxococcales	III	0.36
ASV_000044	Flavobacteriaceae	IV	0.35
ASV_000052	Helicobacteraceae	IV	0.3
ASV_000114	*Desulfococcus*	II	0.29
ASV_000026	Acidimicrobiales	IV	0.27
ASV_000066	Caldilineaceae	IV	0.27
ASV_000046	Helicobacteraceae	IV	0.25
ASV_000173	*Psychromonas*	IV	0.25
ASV_000233	Myxococcales	I	0.24
ASV_000198	Gammaproteobacteria	II	0.23
ASV_000086	Bacteroidales	IV	0.22
ASV_000033	Enterobacteriaceae	III	0.22
ASV_000155	Myxococcales	III	0.22
ASV_000186	Nitrospiraceae	II	0.22
ASV_000098	Caldilineaceae	IV	0.2
ASV_000085	Flavobacteriaceae	IV	0.2
ASV_000149	Myxococcales	II	0.2
ASV_000128	Unknown	IV	0.2
ASV_000123	Alteromonadales	IV	0.19
ASV_000090	Thiohalorhabdales	IV	0.19
ASV_000238	Piscirickettsiaceae	III	0.18
ASV_000014	Helicobacteraceae	IV	0.17
ASV_000296	Helicobacteraceae	IV	0.17
ASV_000153	Lachnospiraceae	IV	0.17

ASV_000271	Myxococcales	III	0.17
ASV_000165	Desulfobacteraceae	IV	0.16
ASV_000080	Desulfosarcina	IV	0.16
ASV_000276	Gammaproteobacteria	II	0.16
ASV_000385	Piscirickettsiaceae	II	0.16
ASV_000268	*Psychromonas*	IV	0.16
ASV_000208	*Psychromonas*	IV	0.16
ASV_000134	Ruminococcaceae	IV	0.16
ASV_000456	Acidimicrobiales	II	0.14
ASV_000217	Acidobacteria	III	0.14
ASV_000437	Alphaproteobacteria	II	0.14
ASV_000273	Myxococcales	II	0.14
ASV_000357	Nitrospiraceae	II	0.14
ASV_000377	Piscirickettsiaceae	II	0.14
ASV_000776	Alteromonadales	II	0.12
ASV_000214	Desulfarculaceae	IV	0.12
ASV_000053	*Lutimonas*	IV	0.12
ASV_000348	Bacteroidales	IV	0.11
ASV_000422	Betaproteobacteria	I	0.11
ASV_000243	Chromatiales	III	0.11
ASV_000281	Desulfobacteraceae	IV	0.11
ASV_000420	Gammaproteobacteria	II	0.11
ASV_000472	Gammaproteobacteria	II	0.11
ASV_000488	Myxococcales	I	0.11
ASV_000275	Desulfobulbaceae	IV	0.1
ASV_000438	Flammeovirgaceae	III	0.1
ASV_000171	Gammaproteobacteria	IV	0.1
ASV_000494	Gemmatimonadetes	II	0.1
ASV_000050	Helicobacteraceae	IV	0.1
ASV_000486	Piscirickettsiaceae	II	0.1
ASV_000662	Piscirickettsiaceae	II	0.1
ASV_000325	Alteromonadales	II	0.09
ASV_000255	Alteromonadales	IV	0.09
ASV_000188	Desulfobulbaceae	IV	0.09
ASV_000064	*Lutimonas*	IV	0.09
ASV_000528	Myxococcales	II	0.09
ASV_001328	Piscirickettsiaceae	II	0.09
ASV_000323	Verrucomicrobia	IV	0.09
ASV_000573	Acidimicrobiales	III	0.08
ASV_000254	Acidimicrobiia	III	0.08
ASV_000458	Alteromonadales	IV	0.08
ASV_000361	Alteromonadales	IV	0.08
ASV_000340	Bacteroidales	V	0.08
ASV_000344	Bacteroidales	IV	0.08
ASV_000342	Desulfobacteraceae	IV	0.08
ASV_000345	Desulfobacteraceae	IV	0.08
ASV_000306	Desulfobulbaceae	II	0.08

ASV_000262	Desulfuromonadaceae	II	0.08
ASV_000482	Myxococcales	II	0.08
ASV_000334	Myxococcales	II	0.08
ASV_000504	Pirellulaceae	I	0.08
ASV_000696	Piscirickettsiaceae	II	0.08
ASV_000366	*Sulfurimonas*	IV	0.08
ASV_000449	Acidimicrobiales	II	0.07
ASV_000423	Acidobacteria	III	0.07
ASV_000556	Alphaproteobacteria	II	0.07
ASV_000250	Alteromonadales	IV	0.07
ASV_000725	Alteromonadales	II	0.07
ASV_000207	Alteromonadales	IV	0.07
ASV_000314	Alteromonadales	IV	0.07
ASV_000354	Bacteroidales	IV	0.07
ASV_000393	Desulfobulbaceae	III	0.07
ASV_000122	Flavobacteriaceae	IV	0.07
ASV_000710	Gammaproteobacteria	II	0.07
ASV_000591	Gemmatimonadetes	II	0.07
ASV_000218	Helicobacteraceae	IV	0.07
ASV_000105	Helicobacteraceae	IV	0.07
ASV_000491	Hyphomicrobiaceae	II	0.07
ASV_000167	*Lutimonas*	IV	0.07
ASV_000809	Myxococcales	I	0.07
ASV_000748	Myxococcales	II	0.07
ASV_000735	Myxococcales	II	0.07
ASV_000600	Piscirickettsiaceae	III	0.07
ASV_000786	Piscirickettsiaceae	III	0.07
ASV_000763	*Psychromonas*	III	0.07
ASV_000414	Unknown	IV	0.07
ASV_000368	Verrucomicrobia	IV	0.07
ASV_001190	Acidimicrobiales	II	0.06
ASV_000708	Acidimicrobiales	II	0.06
ASV_000586	Acidobacteria	II	0.06
ASV_001120	Acidobacteria	II	0.06
ASV_000521	Acidomicrobiales	II	0.06
ASV_000723	Alphaproteobacteria	II	0.06
ASV_000749	Alteromonadales	II	0.06
ASV_000355	Alteromonadales	IV	0.06
ASV_000646	Alteromonadales	II	0.06
ASV_000799	Alteromonadales	III	0.06
ASV_000351	Anaerolineae	II	0.06
ASV_000168	Bacteroidales	III	0.06
ASV_000443	Chromatiales	III	0.06
ASV_000324	Deltaproteobacteria	III	0.06
ASV_000593	Deltaproteobacteria	III	0.06
ASV_000082	Desulfobulbaceae	IV	0.06
ASV_000990	Flammeovirgaceae	II	0.06

ASV_000866	Myxococcales	II	0.06
ASV_000527	Myxococcales	III	0.06
ASV_000930	Nitrosomonadaceae	I	0.06
ASV_000985	*Nitrospina*	I	0.06
ASV_001204	Piscirickettsiaceae	II	0.06
ASV_000898	Piscirickettsiaceae	III	0.06
ASV_001010	Piscirickettsiaceae	II	0.06
ASV_000721	*Psychromonas*	IV	0.06
ASV_000532	Verrucomicrobia	IV	0.06
ASV_000874	Deltaproteobacteria	II	0.05
ASV_000716	Desulfobulbaceae	II	0.05
ASV_000499	Desulfobulbaceae	IV	0.05
ASV_000782	Piscirickettsiaceae	III	0.05

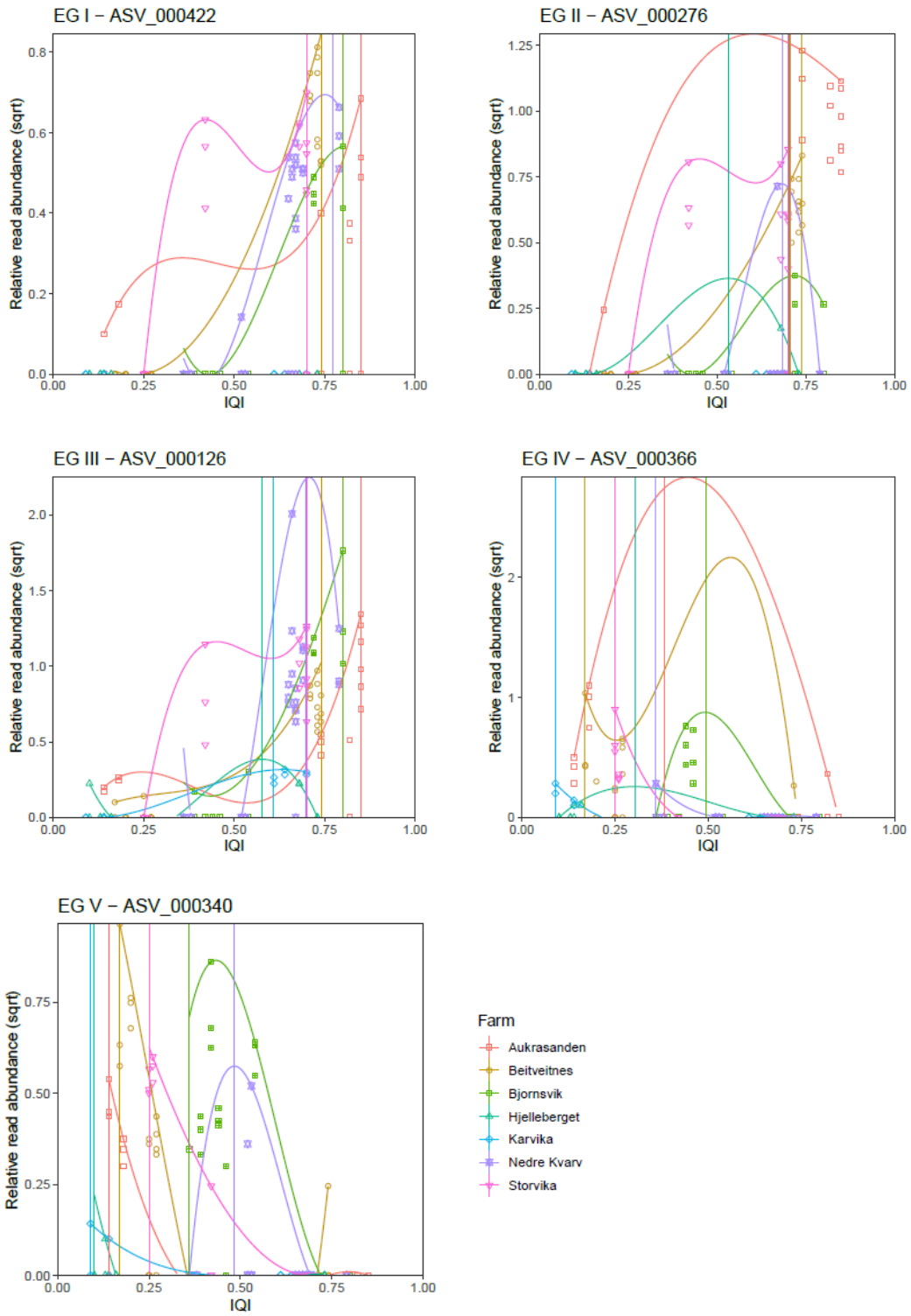
Supplementary File 3.4)

Supplementary File 3.4) Table of ASVs identified as indicators by the QRS approach for the Scottish dataset*. For each ASV, the highest resolution of taxonomic assignment is indicated in the column 'Taxon'. The inferred Eco-Group according to the QRS calculations is indicated at the column 'Eco-Group'. The relative abundance of the ASV among all samples is indicated in the column 'Relative abundance'.*

ASV	Taxon	Eco-Group	Relative abundance
ASV_000008	*Psychrilyobacter*	IV	2.81
ASV_000014	Helicobacteraceae	IV	1.53
ASV_000004	Planococcaceae	III	1.38
ASV_000003	Alteromonadales	IV	1.12
ASV_000013	Helicobacteraceae	IV	0.7
ASV_000017	Bacillales	III	0.67
ASV_000018	Acidimicrobiales	IV	0.62
ASV_000060	Alteromonadales	IV	0.47
ASV_000079	Alteromonadales	IV	0.44
ASV_000078	Helicobacteraceae	IV	0.43
ASV_000016	Helicobacteraceae	IV	0.4
ASV_000028	Acidimicrobiales	III	0.38
ASV_000064	*Lutimonas*	IV	0.34
ASV_000026	Acidimicrobiales	IV	0.3
ASV_000053	*Lutimonas*	IV	0.3
ASV_000067	Helicobacteraceae	IV	0.27
ASV_000092	Actinomycetales	IV	0.25
ASV_000051	Helicobacteraceae	IV	0.25
ASV_000023	Helicobacteraceae	IV	0.22
ASV_000024	Alteromonadales	IV	0.21
ASV_000059	Desulfobulbaceae	IV	0.21
ASV_000082	Desulfobulbaceae	IV	0.18
ASV_000175	Desulfobulbaceae	IV	0.17
ASV_000087	Helicobacteraceae	IV	0.17
ASV_000096	Helicobacteraceae	IV	0.17
ASV_000154	*Lutimonas*	IV	0.17
ASV_000044	Flavobacteriaceae	IV	0.16
ASV_000050	Helicobacteraceae	IV	0.16
ASV_000039	Bacteroidales	IV	0.15
ASV_000301	Anaerolineae	IV	0.12
ASV_000350	Chromatiales	IV	0.12
ASV_000191	Helicobacteraceae	IV	0.12
ASV_000098	Caldilineaceae	IV	0.1
ASV_000066	Caldilineaceae	IV	0.1
ASV_000091	Desulfobacteraceae	IV	0.1
ASV_000331	Chromatiales	IV	0.09
ASV_000137	Helicobacteraceae	III	0.09
ASV_000239	Helicobacteraceae	IV	0.09
ASV_000468	Helicobacteraceae	IV	0.09
ASV_000292	*Lutimonas*	IV	0.09

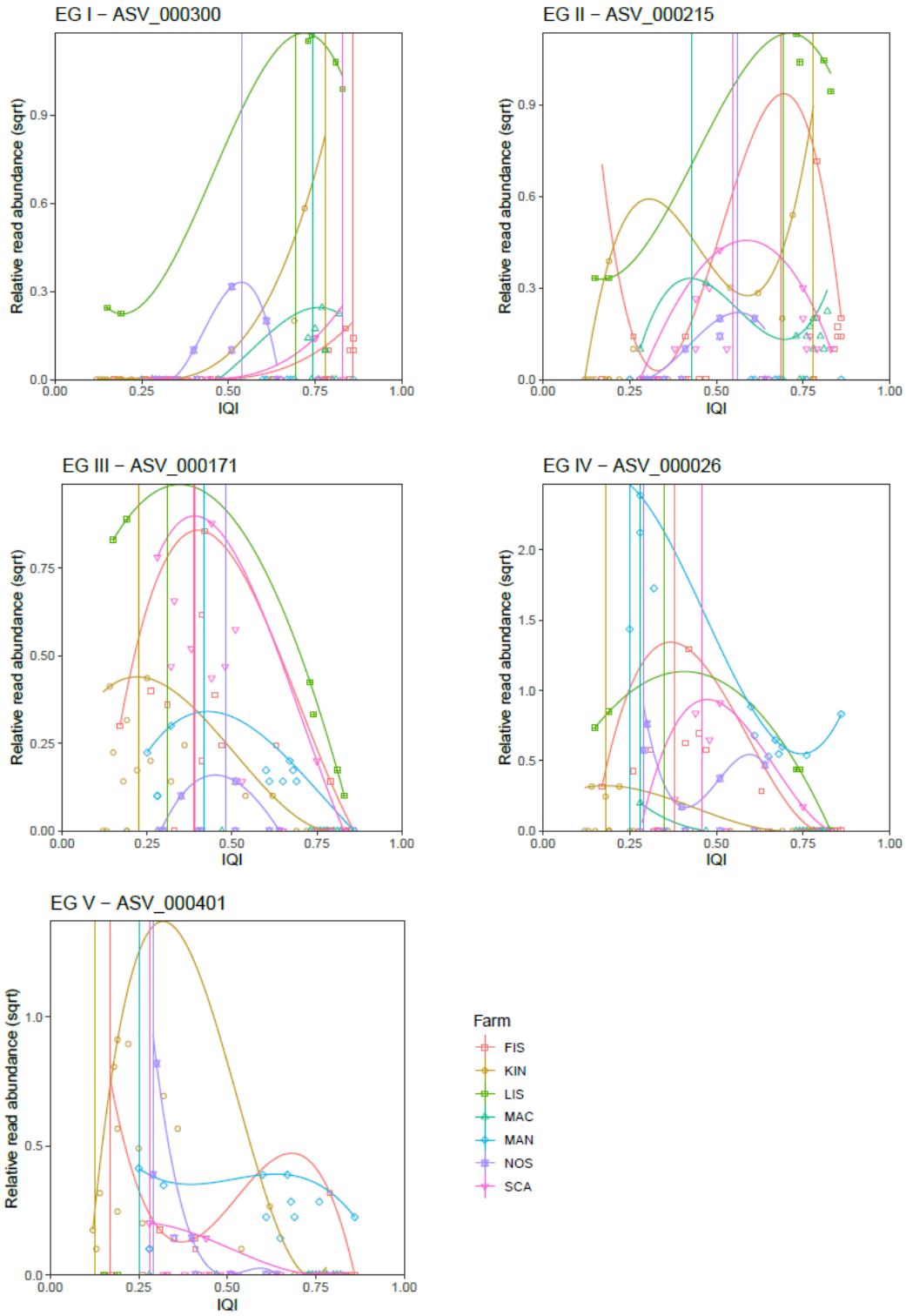
ASV_000421	*Lutimonas*	IV	0.09
ASV_000127	*Persicirhabdus*	IV	0.09
ASV_000038	Thiohalorhabdales	IV	0.09
ASV_000090	Thiohalorhabdales	III	0.09
ASV_000171	Alteromonadales	IV	0.08
ASV_000036	Bacteroidales	IV	0.08
ASV_000215	Desulfuromonadaceae	III	0.08
ASV_000485	Helicobacteraceae	IV	0.08
ASV_000303	*Lutimonas*	IV	0.08
ASV_000031	Myxococcales	III	0.08
ASV_000107	Bacteroidales	IV	0.07
ASV_000122	Flavobacteriaceae	III	0.07
ASV_000473	Helicobacteraceae	IV	0.07
ASV_000401	*Lutimonas*	V	0.07
ASV_000356	Myxococcales	II	0.07
ASV_000110	Unknown	III	0.07
ASV_000200	Acidimicrobiales	III	0.06
ASV_000041	Bacteroidales	IV	0.06
ASV_000582	Chromatiales	IV	0.06
ASV_000781	Chromatiales	IV	0.06
ASV_000300	Desulfobulbaceae	I	0.06
ASV_000306	Desulfobulbaceae	I	0.06
ASV_000262	Desulfuromonadaceae	II	0.06
ASV_000415	Flavobacteriaceae	IV	0.06
ASV_000120	*Lutimonas*	II	0.06
ASV_000111	Actinomycetales	IV	0.05
ASV_000168	Bacteroidales	IV	0.05
ASV_000086	Bacteroidales	IV	0.05
ASV_000566	Chromatiales	III	0.05
ASV_000699	Chromatiales	IV	0.05
ASV_000247	Desulfobulbaceae	I	0.05
ASV_000114	*Desulfococcus*	II	0.05
ASV_000080	*Desulfosarcina*	IV	0.05
ASV_000641	Flavobacteriaceae	IV	0.05
ASV_000412	*Persicirhabdus*	IV	0.05
ASV_000180	Thiotrichaceae	IV	0.05
ASV_000630	Thiotrichales	IV	0.05
ASV_000660	Acidimicrobiales	IV	0.04
ASV_000552	Caldilineaceae	IV	0.04

Supplementary Figure 3.5)



Supplementary Figure 3.5) QRS plots showing the response of selected ASVs from EGs I to V for the Norwegian salmon farms*. For each ASV, relative abundance information is plotted against the sample IQI. Quantile Regression Splines are calculated for each farm individually, which is indicated by the different colors. For each ASV, the inferred Eco-Group is indicated by 'EG' following the respective roman numeral.*

Supplementary File 3.6)



Supplementary Figure 3.6) QRS plots showing the response of selected ASVs from EGs I to V for the Scottish salmon farms. *For each ASV, relative abundance information is plotted against the sample IQI. Quantile Regression Splines are calculated for each farm individually, which is indicated by the different colors. For each ASV, the inferred Eco-Group is indicated by 'EG' following the respective roman numeral.*

Supplementary File 3.7)

***Supplementary Table 3.7) Table of the top 20 indicator ASV inferred by RF**. For each ASV, RF calculates a variable importance according to the influence of the ASVs regarding the right prediction. For regression analysis, variable importance can be described as the percentage in increase of the root mean squared error. Additionally, for each ASV the best taxonomic assignment possible is indicated in the column 'Taxon'.*

Dataset	ASV	Taxon	Variable importance
Norway	ASV_000011	Stramenopiles	13.25
Norway	ASV_000198	Gammaproteobacteria	12.56
Norway	ASV_000013	Helicobacteraceae	9.19
Norway	ASV_000126	*Nitrospina*	7.47
Norway	ASV_000080	*Desulfosarcina*	7.26
Norway	ASV_000240	Stramenopiles	7.23
Norway	ASV_000044	Flavobacteriaceae	7.08
Norway	ASV_000063	Rhodospirillales	6.99
Norway	ASV_000105	Helicobacteraceae	6.37
Norway	ASV_000054	Syntrophobacteraceae	6.17
Norway	ASV_000296	Helicobacteraceae	5.73
Norway	ASV_000212	Desulfobulbaceae	5.56
Norway	ASV_000040	Myxococcales	5.47
Norway	ASV_000217	Unknown	5.03
Norway	ASV_000107	Bacteroidales	5
Norway	ASV_000348	Bacteroidales	4.91
Norway	ASV_000024	Unknown	4.84
Norway	ASV_000031	Myxococcales	4.81
Norway	ASV_000036	Bacteroidales	4.68
Norway	ASV_000469	Gammaproteobacteria	4.54
Scotland	ASV_000064	*Lutimonas*	14.67
Scotland	ASV_000014	Helicobacteraceae	8.33
Scotland	ASV_000078	Helicobacteraceae	6.65
Scotland	ASV_000008	*Psychrilyobacter*	6.56
Scotland	ASV_000067	Helicobacteraceae	6.13
Scotland	ASV_000195	*Psychrilyobacter*	5.63
Scotland	ASV_000079	Alteromonadales	5.59
Scotland	ASV_000154	*Lutimonas*	5.52
Scotland	ASV_000421	*Lutimonas*	5.1
Scotland	ASV_000418	*Psychrilyobacter*	4.95
Scotland	ASV_000107	Bacteroidales	4.77
Scotland	ASV_000039	Bacteroidales	4.67
Scotland	ASV_000114	*Desulfococcus*	4.64
Scotland	ASV_000053	*Lutimonas*	4.56
Scotland	ASV_000003	Alteromonadales	4.51
Scotland	ASV_000760	*Psychrilyobacter*	4.5
Scotland	ASV_000211	Chromatiales	4.44
Scotland	ASV_000513	*Psychrilyobacter*	4.31
Scotland	ASV_000175	Desulfobulbaceae	4.29
Scotland	ASV_000292	*Lutimonas*	4.18

Supplementary File 3.8)

Supplementary Table 3.8) RF classification accuracies. *Results of the RF classification approach are shown including the number of samples for correct and incorrect classifications for A) Norway and B) Scotland. Additionally, the classification accuracy and the kappa values are given.*

A) Norway

Result of classification	Number of samples
Correct classifications adequate to adequate	63
Correct classifications inadequate to inadequate	67
Misclassifications adequate to inandequate	6
Misclassifications inadequate to adequate	2

Accuracy	94.2%
Kappa	0.88

B) Scotland

Result of classification	Number of samples
Correct classifications adequate to adequate	36
Correct classifications inadequate to inadequate	48
Misclassifications adequate to inandequate	4
Misclassifications inadequate to adequate	4

Accuracy	91.3%
Kappa	0.82

Supplementary File 3.9)

Supplementary Table 3.9) RF classification results. *Results of the RF classification are compared to the RF regression and the inferred QRS results for the two IQI categories 'adequate' and 'inadequate' samples. The column reference indicates the inferred category according to reference macrofauna data. Only misclassified samples are shown.*

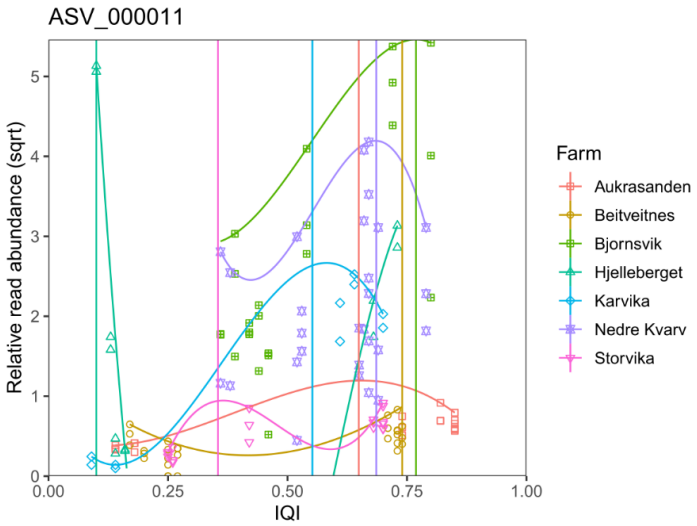
Dataset	Sample	Reference	QRS	RF regression	RF classification
Norway	HJ_588_RepA	adequate	inadequate	inadequate	inadequate
Norway	HJ_588_RepB	adequate	inadequate	inadequate	inadequate
Norway	ST_68_G1_RepB	inadequate	adequate	adequate	adequate
Norway	AK_630_G1_RepA	adequate	inadequate	adequate	adequate
Norway	BJ_280_G2_RepA	adequate	inadequate	adequate	adequate
Norway	BV_1110_G2_RepC	adequate	inadequate	adequate	adequate
Norway	BV_430_G1_RepB	adequate	inadequate	adequate	adequate
Norway	BV_430_G2_RepB	adequate	inadequate	adequate	adequate
Norway	NK_640_G2_RepA	adequate	inadequate	adequate	adequate
Norway	ST_68_G1_RepA	inadequate	inadequate	adequate	adequate
Norway	ST_68_G1_RepC	inadequate	inadequate	adequate	adequate
Norway	NK_140_G2_RepB	adequate	inadequate	inadequate	adequate
Norway	NK_640_G2_RepC	adequate	inadequate	inadequate	adequate
Norway	HJ_50_RepB	inadequate	adequate	inadequate	adequate
Norway	KA_290_RepA	inadequate	adequate	inadequate	adequate
Norway	KA_290_RepB	inadequate	adequate	inadequate	adequate
Norway	BV_430_G1_RepC	adequate	adequate	inadequate	adequate
Norway	BV_430_G2_RepC	adequate	adequate	inadequate	adequate
Norway	HJ_488_RepA	adequate	adequate	inadequate	adequate
Norway	HJ_488_RepB	adequate	adequate	inadequate	adequate
Norway	KA_428_RepA	adequate	adequate	inadequate	adequate
Norway	KA_428_RepB	adequate	adequate	inadequate	adequate
Norway	KA_504_RepB	adequate	adequate	inadequate	adequate
Norway	NK_140_G1_RepB	adequate	adequate	inadequate	adequate
Norway	NK_140_G1_RepC	adequate	adequate	inadequate	adequate
Norway	ST_68_G2_RepB	adequate	adequate	inadequate	adequate
Scotland	FIS_50NE_RepA	adequate	inadequate	inadequate	inadequate
Scotland	FIS_50SE_RepA	inadequate	inadequate	adequate	adequate
Scotland	FIS_Ref2_RepB	adequate	inadequate	adequate	adequate
Scotland	MAN_25S_RepA	inadequate	adequate	adequate	adequate
Scotland	MAN_50N_RepB	adequate	inadequate	inadequate	inadequate
Scotland	MAN_50S_RepA	adequate	adequate	inadequate	adequate
Scotland	NOS_54_RepA	inadequate	inadequate	inadequate	adequate
Scotland	NOS_64_RepA	inadequate	adequate	inadequate	inadequate
Scotland	NOS_Start2_RepA	adequate	inadequate	inadequate	inadequate
Scotland	SCA_100SE_RepB	adequate	inadequate	inadequate	inadequate
Scotland	SCA_37_RepB	inadequate	inadequate	inadequate	adequate
Scotland	SCA_57_RepB	adequate	inadequate	adequate	adequate
Scotland	SCA_Ref1_RepA	adequate	inadequate	adequate	adequate
Scotland	SCA_Ref2_RepA	adequate	inadequate	adequate	adequate
Scotland	SCA_Ref2_RepB	adequate	inadequate	adequate	adequate

Supplementary File 3.10)

Supplementary Table 3.10) Shared QRS indicators among datasets. *The table presents shared indicator ASVs among the two datasets derived from Norway and Scotland which were inferred by QRS. The assigned Eco-Group per ASV for the respective dataset is indicated. For four indicator ASVs, a discrepancy regarding the assigned Eco-Group was detected.*

ASV	Eco-Group Norway	Eco-Group Scotland
ASV_000013	IV	IV
ASV_000014	IV	IV
ASV_000016	IV	IV
ASV_000024	IV	IV
ASV_000026	IV	IV
ASV_000031	III	III
ASV_000036	IV	IV
ASV_000038	IV	IV
ASV_000039	IV	IV
ASV_000044	IV	IV
ASV_000050	IV	IV
ASV_000053	IV	IV
ASV_000064	IV	IV
ASV_000066	IV	IV
ASV_000080	IV	IV
ASV_000082	IV	IV
ASV_000086	IV	IV
ASV_000090	IV	III
ASV_000098	IV	IV
ASV_000114	II	II
ASV_000122	IV	III
ASV_000168	III	IV
ASV_000171	IV	IV
ASV_000262	II	II
ASV_000306	II	I

Supplementary File 3.11)



Supplementary Figure 3.11) QRS plot of ASV_000011. *ASV_000011 was found in all Norwegian sediments. For different farms, the ASV shows a different abundance pattern along the enrichment gradient represented by the IQI. The QRS peak for Hjelleberget farm is visible at a poor environmental quality (IQI = 0.1, turquoise) while the QRS abundance peak for Bjornsvik farm is visible at high environmental quality (IQI = 0.77, green). Therefore, ASV_0000011 does not meet the QRS quality score requirements (maximal deviation = 0.2) and is therefore not considered as an indicator ASV.*

Supplementary files of Chapter IV

Supplementary File 4.1)

Supplementary File 4.1) Table of samples among the salmon farm production phases. *The table represent all samples from the novel 'ScoSa' dataset including their respective station, sampling month and salmon production phase. Station and salmon production phase were targeted for classification prediction.*

Sample	Station	Sampling Month	Salmon production phase
ScoSa18_Apr_AZE_RepA	AZE	April_2018	pre-production phase
ScoSa18_Apr_AZE_RepB	AZE	April_2018	pre-production phase
ScoSa18_Apr_AZE_RepC	AZE	April_2018	pre-production phase
ScoSa18_Apr_CE_RepA	CE	April_2018	pre-production phase
ScoSa18_Apr_CE_RepB	CE	April_2018	pre-production phase
ScoSa18_Apr_REF_RepA	REF	April_2018	pre-production phase
ScoSa18_Apr_REF_RepB	REF	April_2018	pre-production phase
ScoSa18_Apr_REF_RepC	REF	April_2018	pre-production phase
ScoSa18_Aug_AZE_RepA	AZE	August_2018	early production phase
ScoSa18_Aug_CE_RepA	CE	August_2018	early production phase
ScoSa18_Aug_CE_RepB	CE	August_2018	early production phase
ScoSa18_Aug_CE_RepC	CE	August_2018	early production phase
ScoSa18_Aug_REF_RepA	REF	August_2018	early production phase
ScoSa18_Aug_REF_RepB	REF	August_2018	early production phase
ScoSa18_Aug_REF_RepC	REF	August_2018	early production phase
ScoSa18_Jul_AZE_RepA	AZE	July_2018	early production phase
ScoSa18_Jul_AZE_RepB	AZE	July_2018	early production phase
ScoSa18_Jul_AZE_RepC	AZE	July_2018	early production phase
ScoSa18_Jul_CE_RepA	CE	July_2018	early production phase
ScoSa18_Jul_CE_RepB	CE	July_2018	early production phase
ScoSa18_Jul_CE_RepC	CE	July_2018	early production phase
ScoSa18_Jul_REF_RepA	REF	July_2018	early production phase
ScoSa18_Jul_REF_RepB	REF	July_2018	early production phase
ScoSa18_Jul_REF_RepC	REF	July_2018	early production phase
ScoSa18_Jun_AZE_RepB	AZE	June_2018	early production phase
ScoSa18_Jun_AZE_RepC	AZE	June_2018	early production phase
ScoSa18_Jun_CE_RepA	CE	June_2018	early production phase
ScoSa18_Jun_CE_RepB	CE	June_2018	early production phase
ScoSa18_Jun_CE_RepC	CE	June_2018	early production phase
ScoSa18_Jun_REF_RepA	REF	June_2018	early production phase
ScoSa18_Jun_REF_RepB	REF	June_2018	early production phase
ScoSa18_Jun_REF_RepC	REF	June_2018	early production phase
ScoSa18_Mar_AZE_RepA	AZE	March_2018	pre-production phase
ScoSa18_Mar_AZE_RepB	AZE	March_2018	pre-production phase
ScoSa18_Mar_AZE_RepC	AZE	March_2018	pre-production phase
ScoSa18_Mar_CE_RepA	CE	March_2018	pre-production phase
ScoSa18_Mar_CE_RepB	CE	March_2018	pre-production phase
ScoSa18_Mar_CE_RepC	CE	March_2018	pre-production phase

ScoSa18_Mar_REF_RepA	REF	March_2018	pre-production phase
ScoSa18_Mar_REF_RepC	REF	March_2018	pre-production phase
ScoSa18_May_AZE_RepA	AZE	May_2018	pre-production phase
ScoSa18_May_AZE_RepB	AZE	May_2018	pre-production phase
ScoSa18_May_AZE_RepC	AZE	May_2018	pre-production phase
ScoSa18_May_CE_RepA	CE	May_2018	pre-production phase
ScoSa18_May_CE_RepB	CE	May_2018	pre-production phase
ScoSa18_May_CE_RepC	CE	May_2018	pre-production phase
ScoSa18_May_REF_RepA	REF	May_2018	pre-production phase
ScoSa18_May_REF_RepB	REF	May_2018	pre-production phase
ScoSa18_May_REF_RepC	REF	May_2018	pre-production phase
ScoSa18_Oct_AZE_RepA	AZE	October_2018	late production phase
ScoSa18_Oct_AZE_RepC	AZE	October_2018	late production phase
ScoSa18_Oct_CE_RepA	CE	October_2018	late production phase
ScoSa18_Oct_CE_RepB	CE	October_2018	late production phase
ScoSa18_Oct_CE_RepC	CE	October_2018	late production phase
ScoSa18_Sep_AZE_RepA	AZE	September_2018	late production phase
ScoSa18_Sep_AZE_RepB	AZE	September_2018	late production phase
ScoSa18_Sep_CE_RepA	CE	September_2018	late production phase
ScoSa18_Sep_CE_RepB	CE	September_2018	late production phase
ScoSa18_Sep_CE_RepC	CE	September_2018	late production phase
ScoSa18_Sep_REF_RepA	REF	September_2018	late production phase
ScoSa18_Sep_REF_RepC	REF	September_2018	late production phase
ScoSa19_Feb_CE_RepA	CE	February_2019	late production phase
ScoSa19_Feb_REF_RepA	REF	February_2019	late production phase
ScoSa19_Jan_AZE_RepA	AZE	January_2019	late production phase
ScoSa19_Jan_AZE_RepB	AZE	January_2019	late production phase
ScoSa19_Jan_CE_RepB	CE	January_2019	late production phase
ScoSa19_Jan_REF_RepA	REF	January_2019	late production phase
ScoSa19_Mar26_AZE_RepA	AZE	March_2019	late production phase
ScoSa19_Mar26_AZE_RepB	AZE	March_2019	late production phase
ScoSa19_Mar26_AZE_RepC	AZE	March_2019	late production phase
ScoSa19_Mar26_CE_RepA	CE	March_2019	late production phase
ScoSa19_Mar26_CE_RepB	CE	March_2019	late production phase
ScoSa19_Mar26_CE_RepC	CE	March_2019	late production phase
ScoSa19_Mar26_REF_RepA	REF	March_2019	late production phase
ScoSa19_Mar26_REF_RepB	REF	March_2019	late production phase
ScoSa19_Mar26_REF_RepC	REF	March_2019	late production phase

Supplementary File 4.2)

Supplementary File 4.2) Details of model construction. *This supplemental document is intended to clarify how Random Forest (RF) models were constructed during the analysis. The novel ScoSa dataset was used as an example. In the first step (Step 1), full models (FM) are constructed. For this RF models, the complete ScoSa dataset containing all available sequences was used. In the second step (Step 2), the RF models are built on a reduced, down-sampled ScoSa dataset.*

STEP 1

1a.) A full RF model is constructed using all available sequences of the dataset (on average 37,642 sequences per sample). For this FM, *mtry* is set to default value (here: *mtry*=55). This results in one model for step 1a.

1b.) Using the same dataset with *mtry* set to:

default value +1 (*mtry*=56),

default value +2 (*mtry*=57),

default value +3 (*mtry*=58).

This results in tree models for step 1b.

1c.) Using the same dataset with *mtry* set to:

default value -1 (*mtry*=54),

default value -2 (*mtry*=53),

default value -3 (*mtry*=52).

This results in tree models for step 1c.

1d.) Repeat the 7 models from step 1a-1c two times. This results in 21 RF models for step 1 in total.

STEP 2

2a.) The dataset is downsampled to the minimum sequence number per sample (here: 15,177 sequences). This step is conducted using the *rrarefy* function.

2b.) Using this rarefied dataset for new RF models as described in 1b-1d.

This results in 21 models for step 2b.

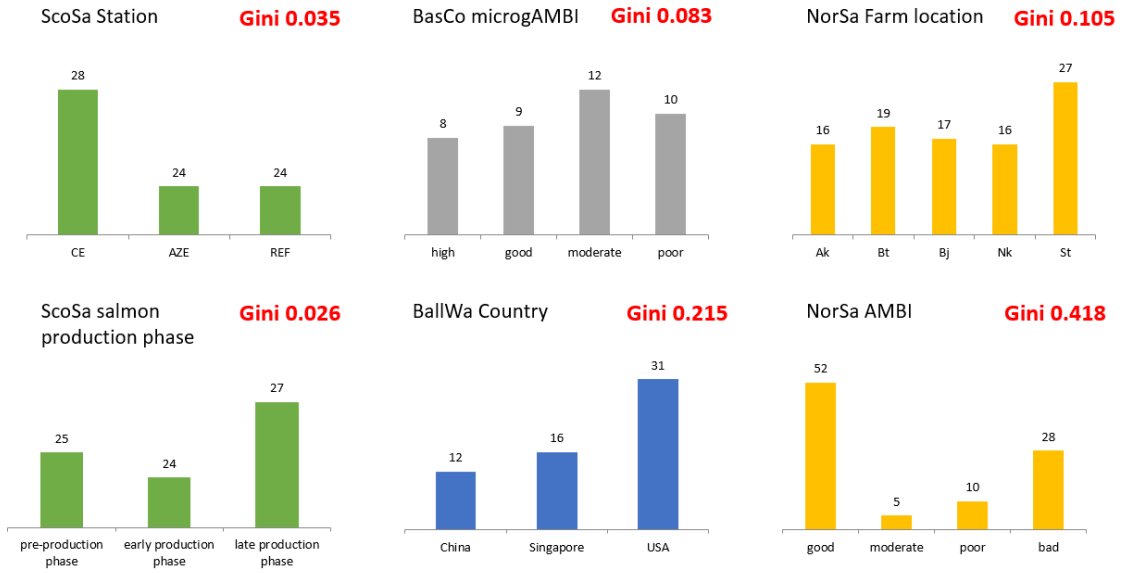
2c.) Deeper rarefaction to given sequence numbers. The dataset is downsampled to 12 different sequence numbers per sample: 12,500, 10,000, 7500, 5000, 2500, 1000, 500, 400, 300, 200, 100 and 50 sequences.

Using these 12 rarefied datasets for new RF models as described in 1a, 1b and 1c with seven models each. This results in 12*7 = 84 models for step 2c.

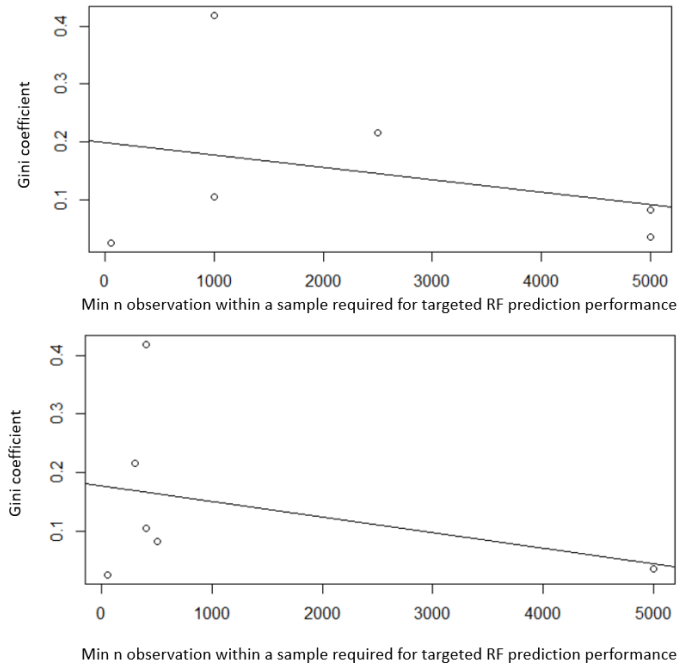
2d.) All of the 84 models from step 2c are repeated two times using different base trees. This results in 168 models for step 2d. This results in 273 RF models for step 2 in total.

Supplementary File 4.3)

A)



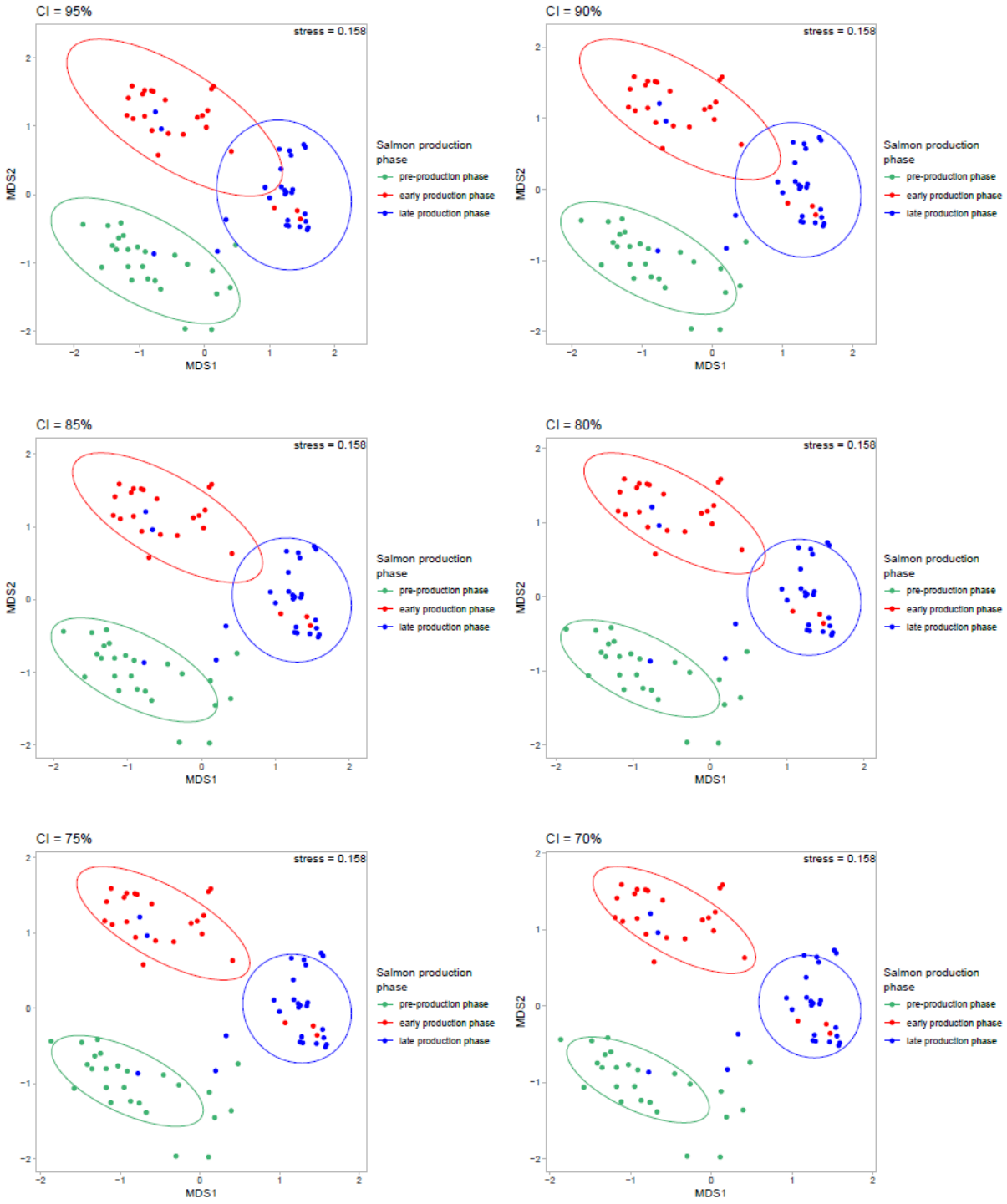
B)



Supplementary File 3.4) Frequencies and distribution of sample categorization. *A) Class frequency distribution in individual datasets visualized in bar plots and calculated with Gini coefficient (in red color). Higher Gini coefficient values indicate a higher imbalance (max Gini coefficient = 1). Despite both ScoSa datasets (Station and salmon production phase) have highly similar Gini coefficients, the difference in the minimum number of observations required for targeted RF prediction performance is maximal (50 sequences/sample versus 5000 sequences per sample). B) Plotting class frequency distributions (Gini coefficient) against minimum number of observations required for targeted RF prediction performance (as measured by kappa and out-of-bag error) shows a lack of correlation between these two parameters (p = 0.56 and p = 0.51)*

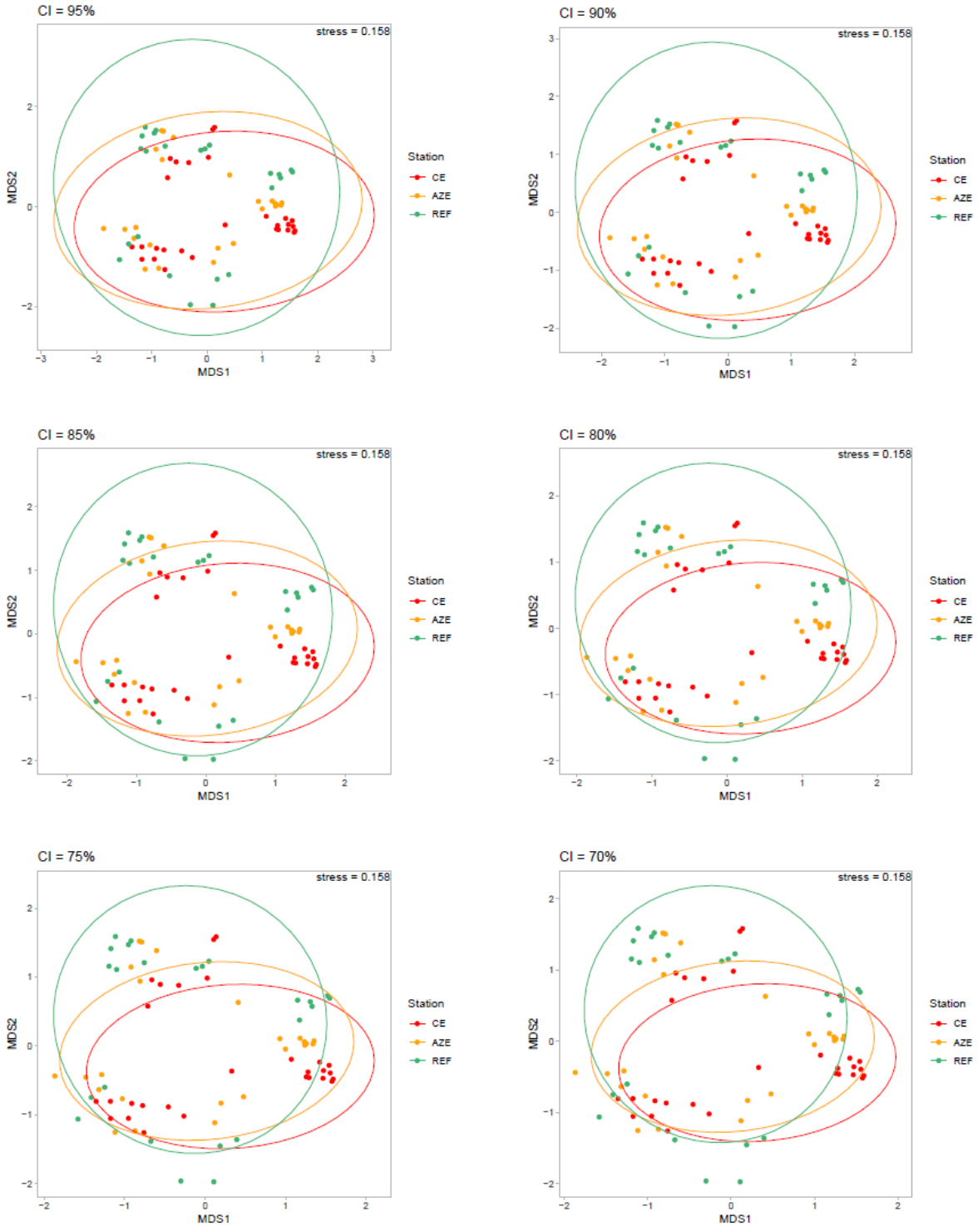
Supplementary File 4.4)

A)



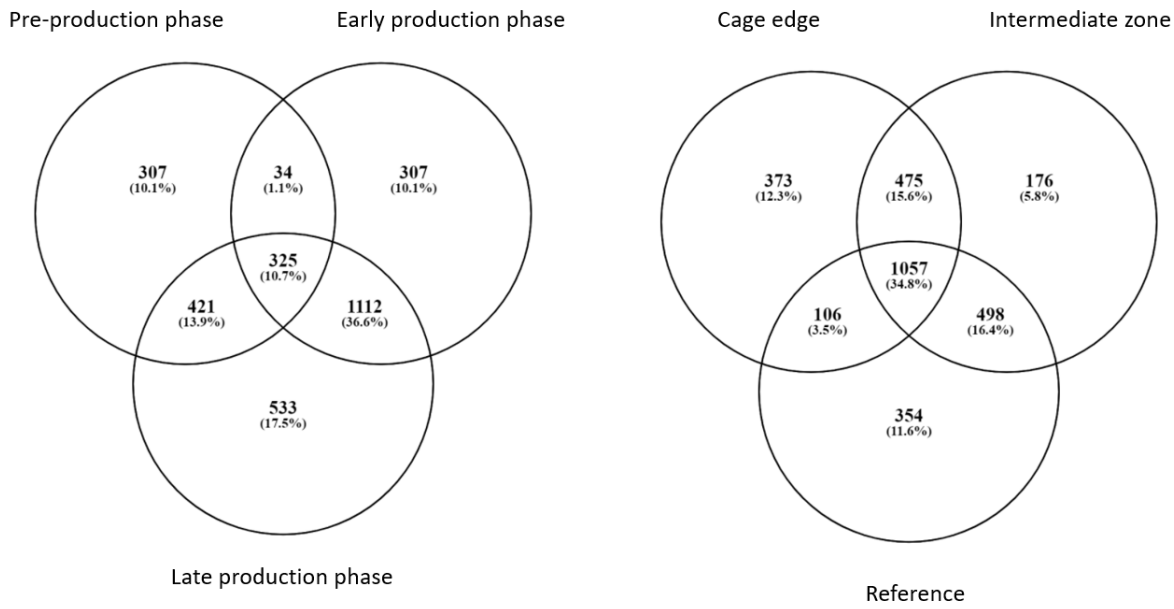
Supplementary File 4.4) Visual representation of separability via ordination analysis*. NMDS of bacterial community ASVs of the ScoSa dataset. Confidence intervals are for classes 'salmon production phase' (A) and 'distance' (B). At different CI-thresholds (95%-70%), salmon production phase classes are substantially better separated than distance classes, corroborating well with the lower number of observations in a sample required to achieve targeted RF prediction performance (50 sequences for salmon production phase prediction versus 5000 sequences for distance prediction)*

B)



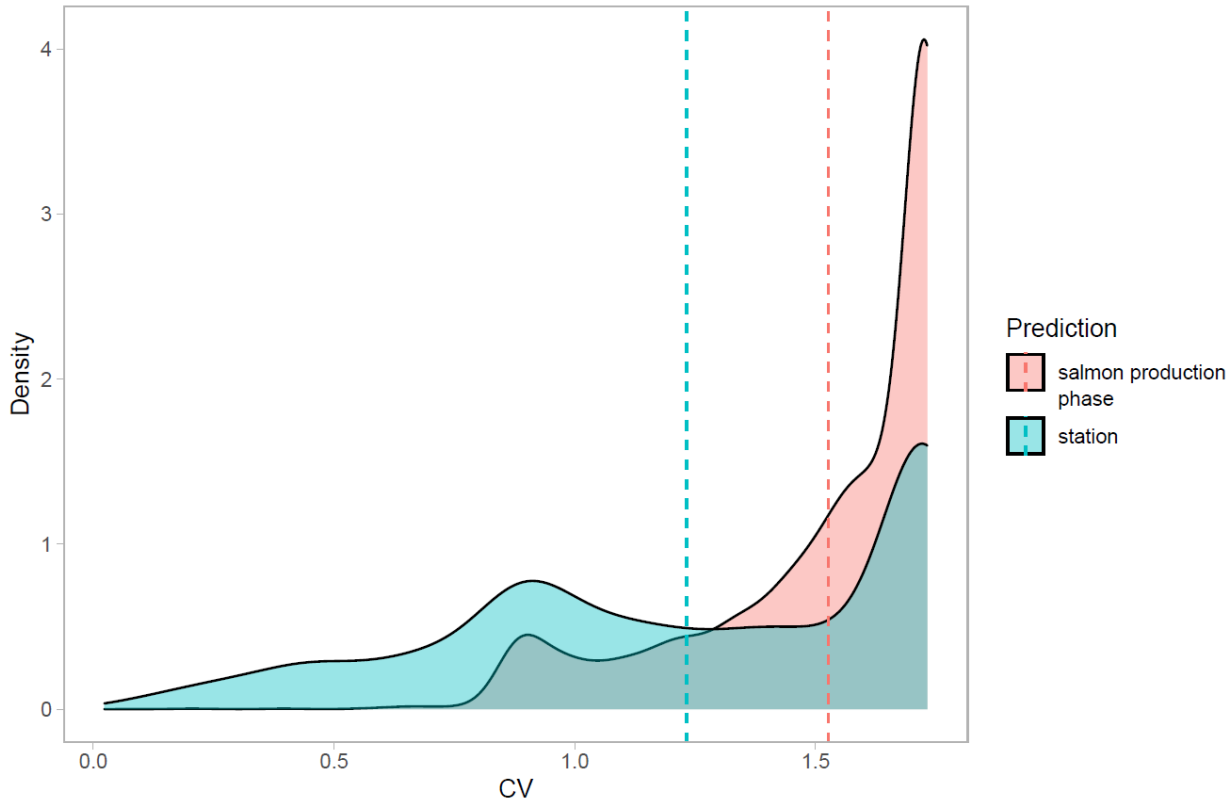
Supplementary File 4.4 *(continued)*

Supplementary File 4.5)



Supplementary File 4.5) Visual representation of feature specificity via Venn diagrams. *Feature specificities for prediction classes. Venn diagrams show ScoSa datasets for salmon production phase (left) and distance (right). The higher the proportion of ASVs (features) exclusive to individual classes and the lower the proportion of ASVs that are common to all classes, the harder are boundaries between prediction classes (see also Supplementary File 4.4 for boundaries between classes).*

Supplementary File 4.6)



Supplementary File 4.6) Visual representation of feature coefficient of variation. *Density plot of the coefficient of variation (CV) for each feature (= ASV) of the ScoSa distance and the ScoSa salmon production phase dataset. To infer the coefficient of variation (CV) for each feature of the ScoSa dataset, the cv function of the EnvStats package was used. The coefficient of variation measures how widely scattered the distribution is relative to the size of the mean. The CV was calculated using samples grouped by 'sampling season' or 'salmon production phase' respectively. The respective CVs per group were then plotted against each other in a multiple density plot. The density plot shows the kernel density estimations per group. The construction of probability density functions is based on the idea that the area between the function and the x-axis (the integral) from a point a to a point b corresponds to the probability of obtaining a value between a and b. The dashed lines indicate the CV group mean (station= 1.23, salmon production phase = 1.53). Thus, when downsampling the ScoSa salmon production phase dataset, it can be expected that the removal of a substantial number of ASVs from the full dataset still leaves more informative features with a high CV than is the case for the ScoSa distance dataset.*

CURRICULUM VITAE

Personal information

Name:	Verena Nicola Rubel (née Dully)
Place of birth:	Landau i.d. Pfalz, 76829, Germany

Education and jobs

03/2019-04/2022: *Doctoral program*
Technical University of Kaiserslautern, Germany. Research and teaching activities, employed in the field of molecular ecology as scientific staff. Mainly working on marine environmental monitoring and how to optimize it using molecular (eDNA-based) and computational methods (bioinformatics and machine learning). Bachelor and master student instructor, teaching assistant for the practical course 'Limnology and Microbial Ecology' with field work and laboratory work. Lecturer of theoretical course 'Aquatic Microbial Ecology'.

10/2018-01/2019: *Scientific employee*
Department of Ecology, Technical University of Kaiserslautern, Germany. Teaching of bachelor and master students regarding field work and laboratory work. Teaching assistant for the practical course 'Limnology and Microbial Ecology'

10/2016-12/2018 *Master of Science, M.Sc.,*
Technical University of Kaiserslautern, Germany. Master program 'Ecology and Microbial Biodiversity', Grade: 1.1. Master thesis: *'Inferring Indicator Values from ciliate eDNA metabarcodes for ecological status assessment of salmon farms'.* Grade: 1.0

06/2017-03/2018 *Scientific employee*
Department of urban water management (Siedlungswasser-wirtschaft), Technical University of Kaiserslautern, Germany. Sampling at wastewater treatment plants, microscopy, evaluation, and assessment of purification services. Digitalization, hydrographic work & digital mapping. Planning and execution of conferences.

10/2013-05/2017 *Bachelor of Science, B.Sc.,*
Technical University of Kaiserslautern, Germany. Bachelor program ‘Biowissenschaften‘. Bachelor thesis: ‘*Anwendbarkeit geläufiger Schlamm-Indices in kommunalen Kläranlagen in Kaiserslautern und Umgebung*‘. Grade: 1.3

Further education

2019
- International training ‘*Artificial Intelligence and supervised machine learning in biology*’, VLIZ Flanders Marine Institute, Ostend, Belgium
- PhD course ‘*Traditional and molecular methods to assess biodiversity, focus on DNA metabarcoding & aquatic organisms*’, Swedish University of Agricultural Sciences, Uppsala, Sweden
- Python programming course, German Research Center for Artificial Intelligence DFKI, Kaiserslautern, Germany

2018 Internship at the Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research. Department of Molecular genetics and polar biological oceanography, Bremerhaven, Germany

2017 Internship ‘*biological wastewater treatment*’, Department of Ecology in cooperation with the sewage treatment plant Kaiserslautern, Germany

Voluntary commitment

2021
- Supervisor and coach of the children's vacation camp by the German table soccer federation DTFB in Westernohe, Germany
- Official D-license coach for children and youth work at the German table soccer federation DTFB in Minden, Germany

Since 2018 Treasurer and member of the board of directors in the non-profit sports club with youth promotion and captain of the women’s team at 1. Kicker Club Kaiserslautern e.V., Kaiserslautern, Germany

2015-2018 Member of the board of directors in the non-profit sports club with youth promotion at 1. Kicker Club Kaiserslautern e.V., Kaiserslautern, German

List of publications

Journal articles

2021

- **Dully V**, Rech G, Wilding TA, Lanzén A, MacKichan K, Berrill I & Stoeck T. Comparing sediment preservation methods for genomic biomonitoring of coastal marine ecosystems. *Marine Pollution Bulletin* 173: 113129, doi:10.1016/j.marpolbul.2021.113129
- **Dully V**, Wilding TA, Mühlhaus T & Stoeck T. Identifying the minimum amplicon sequence depth to adequately predict classes in eDNA-based marine biomonitoring using supervised machine learning. *Computational and Structural Biotechnolgy Journal* 19: 2256-2268, doi:10.1016/j.csbj.2021.04.005
- Frühe L, **Dully V**, Forster D, Keeley NB, Laroche O, Pochon X, Robinson SMC, Wilding TA & Stoeck T. Global trends of benthic bacterial diversity and community composition along organic enrichment gradients of salmon farms. *Frontiers in Microbiology (section Aquatic Microbiology)* 12: 637811, doi:10.3389/fmicb.2021.637811
- **Dully V**, Balliet H, Frühe L, Däumer M, Thielen A, Gallie S, Berril I & Stoeck T. Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture - An inter-laboratory study. *Ecological Indicators*, doi:10.1016/j.ecolind.2020.107049

2020

Frühe L, Cordier T, **Dully V**, Breiner HW, Lentendu G, Pawlowski J, Martins C, Wilding TA & Stoeck T. Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology*, doi:10.1111/mec15434

2019

Kahlert M, Alfjorden A, Apunte-Ramos K, Bailet B, Pérez Burillo J, Carrera Gonzalez AG, Castro D, Di Bernardi C, **Dully V**, Fekete J, Frühe L, González R, Gratsia E, Hanjalić J, Kamberović J, Kelly AM, Meriggi C, Nousiainen, I, Ørberg S B, Orr J, Quintana CO, Papatheodoulou A, Sargac J, Shahbaz M, Tapolczai K, Tosic K, Wallin I, Zupančič M, Bohman P, Buttigieg P L, Häubner N, Leese F, Macher JN, Peura S, Roslin T, Strand M, Terenius O, Vasselon V & Weigand AM. New molecula methods to assess biodiversity. Potentials and pitfalls of DNA metabarcoding: a workshop report. *Research Ideas and Outcomes*, 5:e38915, doi: 10.3897/rio.5.e38915

2018 Stoeck T, Pan H, **Dully V**, Forster D & Jung T. Towards an eDNA metabarcode-based performance indicator for full-scale municipal wastewater treatment plants. *Water Research,* doi: 10.1016/j.watres.2019.07.051

Presentations

2021 - Predicting classifications in marine biomonitoring with supervised machine learning: how much data is required?
1. DNAqua International Conference, Evian, France
- Towards a standard protocol in coastal aquaculture biomonitoring: an interlaboratory study to assess reproducibility of the wet lab protocol and of Illumina sequencing (poster)
1. DNAqua International Conference, Evian, France

2020 Inter-laboratory reproducibility of machine learning predictions in applied environmental coastal monitoring
Faculty meeting, University of Kaiserslautern

Awards

2021 Award for the best student oral presentation
1. DNAqua International Conference, Evian, France

2020 Award for an outstanding Master thesis, Grade: 1.0
Kreissparkassen-Stiftung Kaiserslautern

ACKNOWLEDGEMENTS

First, I would like to thank Prof. Dr. Thorsten Stoeck for providing me with the opportunity to prepare this dissertation in his research group. Thank you very much for always pointing out the right path towards science, and for your support over the last years since the start of my bachelor thesis.

I would also like to thank Jun. Prof. Dr. Timo Mühlhaus, who agreed to be part of the dissertation committee as a second referee. Also, a big 'Thank You' for your collaboration in research projects and the helpful scientific contributions.

Also, I would like to thank Prof. Dr. Nicole Frankenberg-Dinkel for agreeing to chair the examination committee.

I thank all cooperation partners and co-authors who have always been available with advice and support. A special thanks goes to Dr. Kleopatra Leontidou for collaborating by conducting the QRS analysis. I would also like to thank all funders who made the creation of this work possible.

I say 'Thank You' to the whole Ecology and Microbial Ecology working group, and big praise to all of you! This goes especially to M.Sc. Maren Nothof, M.Sc. Sven Katzenmeier, Hans-Werner Breiner, and Jun. Prof. Dr. Sabine Filker. I have rarely met such competent and motivated people! Your former, recent, and future; institutions, bosses, and colleagues; can only be grateful for having you! Furthermore, I would like to thank my former co-workers Dr. Guillaume Lentendu and Dr. Dominik Forster who helped me become the scientist I am today. A special thanks also goes to Dr. Larissa Frühe, who happily welcomed me into her office at the beginning of my master's thesis. Common interests have fortunately led to many corporations, which we can look back onto in the future. I hope we will meet each other again soon to share some freshly made popcorn.

An dieser Stelle möchte ich meiner Familie danken. Ohne Eure Unterstützung in vielerlei Hinsicht wäre nichts hiervon nur ansatzweise möglich gewesen. Danke Mama, Danke Eva, dass Ihr immer Extrawürste für mich gekocht und gebacken habt. Danke, dass Ihr immer für mich da wart. Man kann sich seine Familie zwar nicht aussuchen, aber hätte ich die Wahl gehabt, hätte ich Euch ausgesucht.

A special thanks also goes to Christoph Rubel, who always kept me on the right track. Thank you for your understanding in difficult situations during my bachelor's and master's studies and during my time as a doctoral student. Thank you for your understanding for nights studied and worked through over the years, even if it was not always easy. Sometime from now, I'll manage to finally take the long-awaited vacation with you. I am looking forward to our upcoming time together.