High-throughput eDNA surveys for benthic monitoring of salmon farms in Norway: a validation study

FHF project 901092

Report 2019

Prepared by Jan Pawlowski, Tristan Cordier, Kristina Cermakova
(UNIGE/ID-Gene ecodiagnostics), and Tomas Cedhagen (University of Aarhus)

UNIVERSITÉ DE GENÈVE

SOLIDUM PETIT IN PROFUNDIS UNIVERSITAS ARHUSIENSIS

ID-GENE ecodiagnostics

Acknowledgements

The authors of the report thanks Catarina Martins, Group Manager Environment & Sustainability at Mowi ASA for leadership, management and continuous support to the project. We thank the members of Steering Committee, in particular Arne Kvalvik, Knut Staven, Pia Kupka Hansen and Kjell Maroni for guidance and helpful discussions. We also thank all persons who help in organising and conducting sampling campaigns, and who contributed to improve the method through discussing its various aspects. In particular, our thanks go to Arne Kvalvik, Knut Staven, Ingrid Lundamo, Are Andreassen Moe, Jonathan Vaz, and Karstad Oystein at Mowi, Arild Kjerstad, Vegard Langvatn, Dag Slettebo and Dagfinn Breivik Skomso at Akerbla, Alex Lindseth, Christine Klykken, Kristine Brokke, Vidar Strom, Nasir El-Shaik at Aqua Kompetanse, Stein Thon Klem, Thomas Tveit Furset, and Ingeborg Elisabeth Okland at Radgivende-Biologer as well as Anja Hervik at Flatlab, and Reinhold Fieler, and Kristin Saether at Akvaplan-niva.

We also acknowledge help of our colleagues at the University of Geneva and ID-Gene ecodiagnostics for assistance and helpful discussions, in particular Joana Visco, Sofia Wyler, Inès Barrenechea, Laure Apothéloz-Perret-Gentil, Maria Holzmann, Emanuela Reo, Kaya Pawlowska, Stanislaw Pawlowski, Philippe Esling, and Franck Lejzerowicz.

Our special thanks go also to the members of Scientific Panel: Thorsten Stoeck (University of Kaiserlautern, Germany), Anders Lanzen (AZTI, Pasaia, Spain), Thierry Baussant (NORCE, Stavanger, Norway) and Nigel Keeley (IMR, Tromso, Norway) for their fruitful discussions.

Shortlisted results

The project aims at developing and validating eDNA metabarcoding tests for fast, sensitive, and cost-effective benthic monitoring of salmon farms in Norway.

It generated huge amount of metabarcoding data, whose analysis demonstrated the usefulness of eDNA metabarcoding as a tool for assessing the ecological status of benthic community. The project showed the importance of meiofauna, in particular nematodes, as excellent indicators of organic enrichment, and highlighted the limitations of using macrofaunal data in metabarcoding analyses. The project also introduced supervised machine learning to predict benthic indices, demonstrating its effectiveness and accuracy compared to conventional taxonomic methods.

Project in numbers

749 samples collected at 27 sites

500 DNA barcodes for 195 species

170'000'000 metabarcodes

7 papers published

15 presentations at international conferences

Main scientific achievements of the project

- Development of **eDNA metabarcoding** for assessment of benthic biodiversity
- Identification of **new meiofauna indicators** of organic enrichment
- Application of **supervised machine learning** tools to predict biotic indices

Successful steps towards implementation of the method

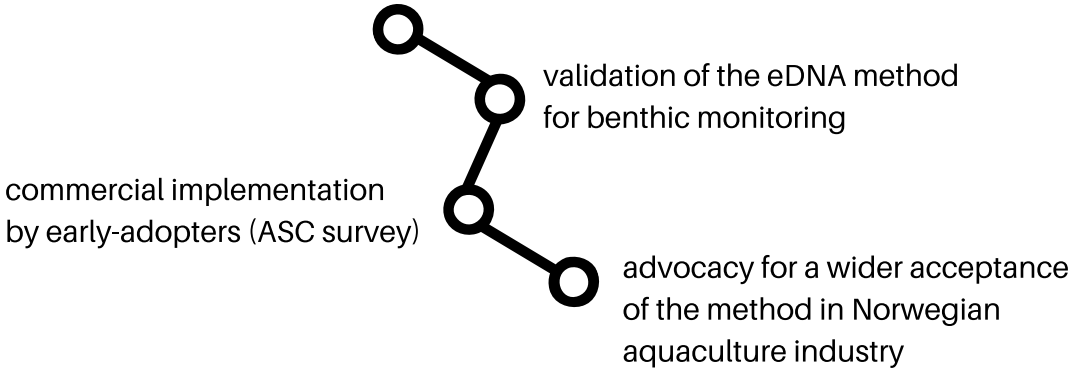
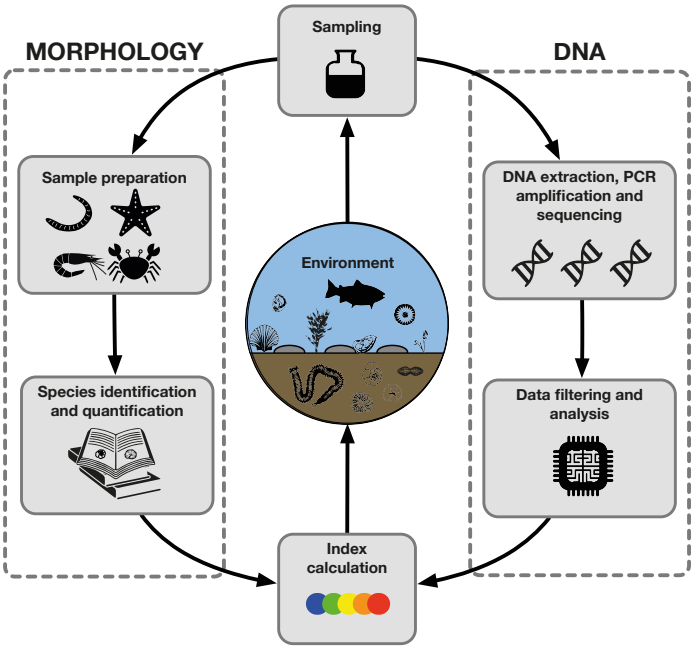


Table of contents

1. Background

Traditionally, the environmental impact of salmon farming on seabed diversity is assessed based on the richness and abundance of benthic macro-invertebrates. However, this traditional approach requires sieving, sorting and morphological identification of specimens, which is time-consuming and demands an excellent taxonomic expertise. This might cause important delays in the analysis of rapidly growing number of samples, seriously limiting the efficiency of benthic monitoring. Moreover, the traditional approach overlooks the morphologically indistinguishable juvenile and life-cycle stages of macrofauna and small-sized organisms (meiofauna, microbes), reducing the accuracy of the assessment of benthic communities.



To overcome the limitations of traditional approach and to take advantages of the development of molecular diagnostic tools, we propose to use environmental DNA (eDNA) metabarcoding to assess the impact of salmon farms on benthic diversity. The metabarcoding approach based on high-throughput sequencing of eDNA has proved to be extremely useful for biodiversity surveys and biomonitoring (Valentini et al. 2016, Taberlet et al. 2018). It has been shown that metabarcoding data can also be used for inferring or predicting biotic indices (Pawlowski et al 2018).

Several studies demonstrate the usefulness of metabarcoding for biomonitoring of marine environment, including surveys of coastal macrofauna (Aylagas et al. 2014, 2016), estuarine (Chariton et al. 2015), and seagrass communities (Cowart et al. 2015) as well as to assess the environmental impact of offshore drilling platforms (Lanzen et al. 2016, Laroche et al. 2016).

Metabarcoding was also applied to assess the impact associated with salmon farming activities in Scotland (Pawlowski et al. 2014), Norway (Pawlowski et al. 2016), New Zealand (Pochon et al. 2015), and Canada (He et al. 2019). Some of these studies target new groups of bioindicators, such as bacteria (Dowle et al. 2015, Stoeck et al. 2018a), foraminifera (Pawlowski et al. 2014, 2016, Pochon et al. 2015) or ciliates (Stoeck et al. 2018b). Others focus on metazoans, showing that the metabarcoding data provide similar evaluation of environmental impact as the morpho-taxonomic analyses (Lejzerowicz et al. 2015).

In this project, we validate metabarcoding as a new tool for benthic monitoring of salmon farms in Norway. The project focuses on completing the barcoding reference database of benthic macrofauna, identifying new meiofauna bioindicators and developing bioinformatic tools to predict the environmental impact using eDNA data.

2. Objectives

The main objective of the project was to develop the eDNA-based survey for a fast, sensitive, and cost-effective benthic monitoring of salmon farms in Norway.

The project comprised three work packages (WP), whose specific aims were:

1. To improve the detection of benthic macroinvertebrates in eDNA samples
2. To identify and validate new meiofaunal bioindicators
3. To develop and validate the eDNA based biotic indices of benthic community

The project workflow involved parallel sampling for macrofauna and eDNA, isolation and molecular identification of macrofauna specimens, analysis of eDNA samples for calibration of biotic indices and identification of meiofauna bioindicators (Figure 1).

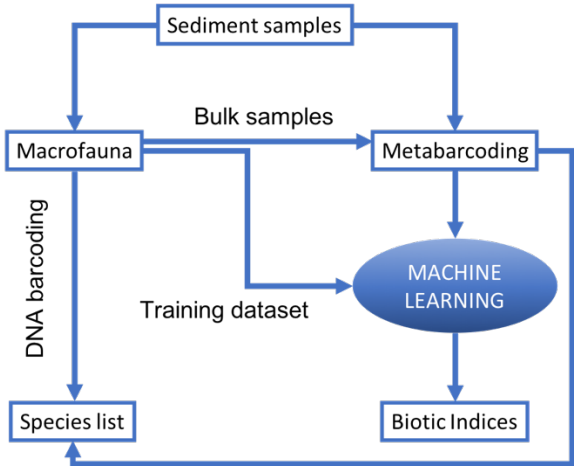


Figure 1. Project workflow

The project started in June 2015 and was initially planned for 1 ½ year. It was extended in 2017 and 2018 in order to expand the sampling of reference sites, complete the barcoding database, and search for meiofaunal bioindicators.

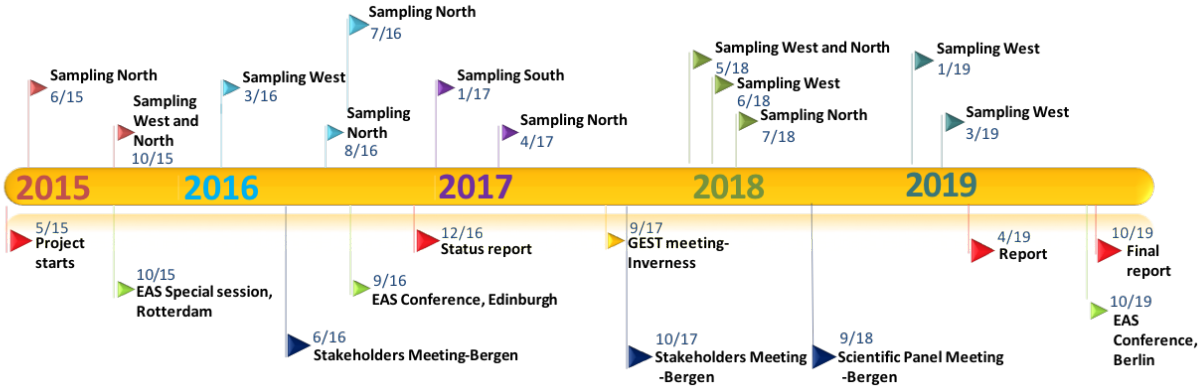
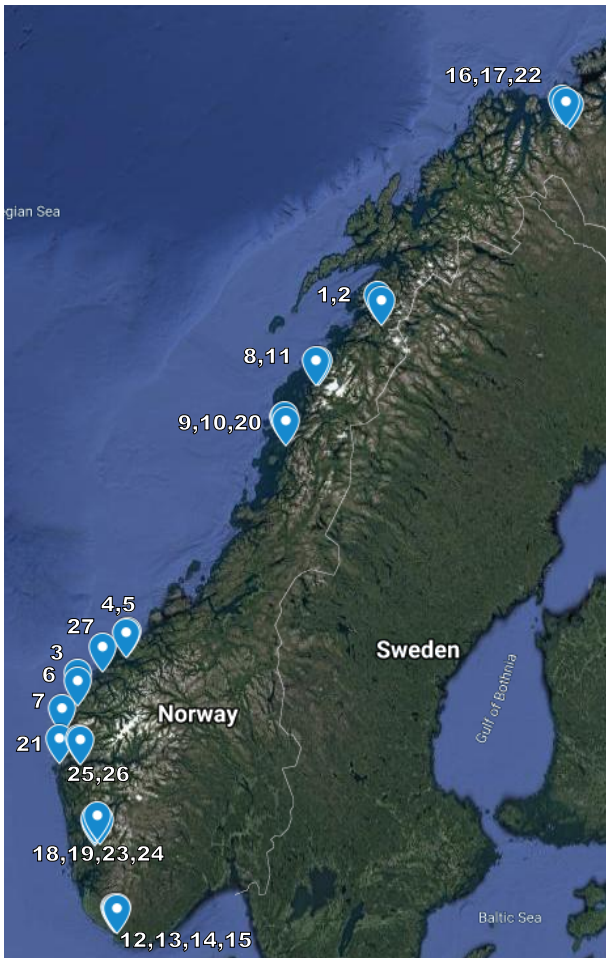


Figure 2. Project timeline

3. Sampling

To fulfil these tasks, the samples were collected between 2015 and 2019 from 27 sites, located along the coast of Norway (see Figure 3).



1. Bjørnsvik (Jun 2015)
2. Nedre Kvarv (Jun 2015)
3. Beitveit (Oct 2015)
4. Storvika (Oct 2015)
5. Aukrasanden (Oct 2015)
6. Rundreimstranda (Mar 2016)
7. Flåtegrunnen (Mar 2016)
8. Kvalvika (Aug 2016)
9. Brattholmen (Aug 2016)
10. Bukkholmen (Jul 2016)
11. Digermulen (Oct 2015)
12. Napp (Jan 2017)
13. Skipningsdalen (Jan 2017)
14. Støytland (Jan 2017)
15. Salvågvika (Jan 2017)
16. Hjellberget (Apr 2017)
17. Karvika (Apr 2017)
18. Tendalsvik (May 2018)
19. Utåker (May 2018)
20. Mefaldskjæret (May 2018)
21. Olderøy (Jun 2018)
22. Rakkenes (Jul 2018)
23. Alsåkervik (Jan 2019)
24. Slåttenes (Jan 2019)
25. Tveit (Mar 2019)
26. Bekksneset (Mar 2019)
27. Røysa (Mar 2019)

Figure 3. Map showing the localisation of sampling sites

The collected material comprises 749 samples from 132 stations. For each station, 2-3 replicates of 5-10 ml of sediment were taken from the surface of two grabs that have been used for macrofaunal study. In addition, for some sites, the 3rd grab was taken for DNA barcoding of macrofauna and analysis of bulk samples (WP1).

Localization of sites, sampling date and number of samples are indicated in Table S1.

4. Results

4.1. DNA barcoding of benthic macrofauna (WP1a)

Analysis of benthic fauna by metabarcoding relies on barcoding reference databases. The two most useful markers for benthic macrofauna are the variable region V1V2 of nuclear 18S rRNA gene and the mitochondrial COI gene.

In the first part of the WP1, our aim was to enrich the barcoding database of benthic macro-invertebrates living in the vicinity of salmon farms at the coast of Norway. In fact, large number of species listed in the morphological reports are not represented in the genetic reference databases such as Genbank or BOLD. This number depends on taxonomic group and on the abundance of species. In general, the common species are well represented, although not always for both markers.

During this project we obtained DNA from over 1000 specimens representing more than 200 species. The number of barcoded specimens for main taxonomic groups is indicated in Table 1. The project considerably increased the number of barcoded species. 82 new COI sequences have been obtained for 41 species, and 91 new 18S sequences have been obtained for 36 species. However, in spite of our contribution, the gap in reference database remains large (Figure 4).

Table 1. Total number of isolated species and specimens of benthic macroinvertebrates, and number of obtained DNA barcodes for 18S and COI markers (species/specimens). The new 18S and COI sequences are under submission to the GenBank database.

	Species	Specimens	18S	18S new	COI	COI new
Annelida	114	567	55/194	**17/44**	58/132	**29/63**
Hydrozoa	7	44	2/4	**1/3**	3/7	**1/3**
Crustacea	38	74	15/30	**8/15**	14/20	**8/13**
Echinodermata	14	32	6/16	**5/13**	2/2	0/0
Mollusca	21	95	15/39	**1/1**	7/12	**1/1**
Nemertea	4	19	4/12	**1/2**	2/2	**1/1**
Others	8	20	11/30	**3/13**	1/1	**1/1**
Total	206	851	108/325	**36/91**	87/176	**41/82**

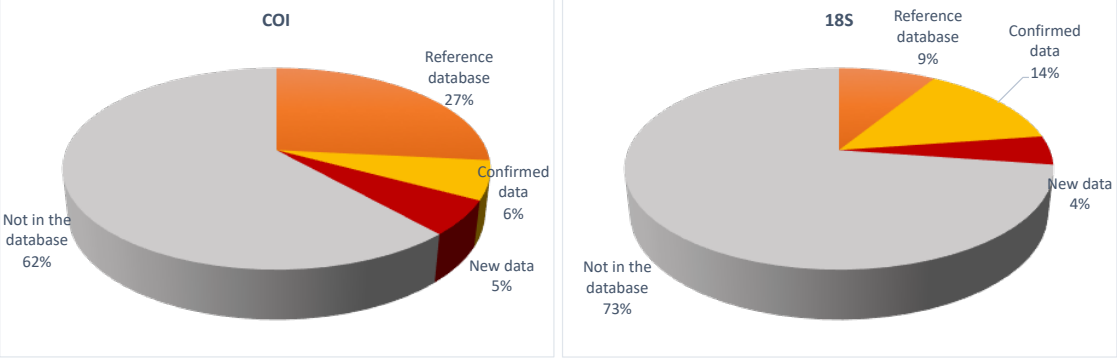


Figure 4. The proportion of macrofauna species present or absent from barcoding reference database for COI and 18S markers.

4.2. Metabarcoding of benthic macrofauna (WP1b)

4.2.1. Macrofauna in bulk samples

In the second part of this work package, we compared the composition of macrofaunal species in sieved, sorted and ethanol fixed samples inferred from eDNA analysis and morphological counts. The analysis was done on 8 samples from Nedre Kvarv and Aukrasanden. In each case, the specimens were sorted, identified and weighted, before being pooled together. Each pool was processed for DNA extraction, amplification and sequencing.

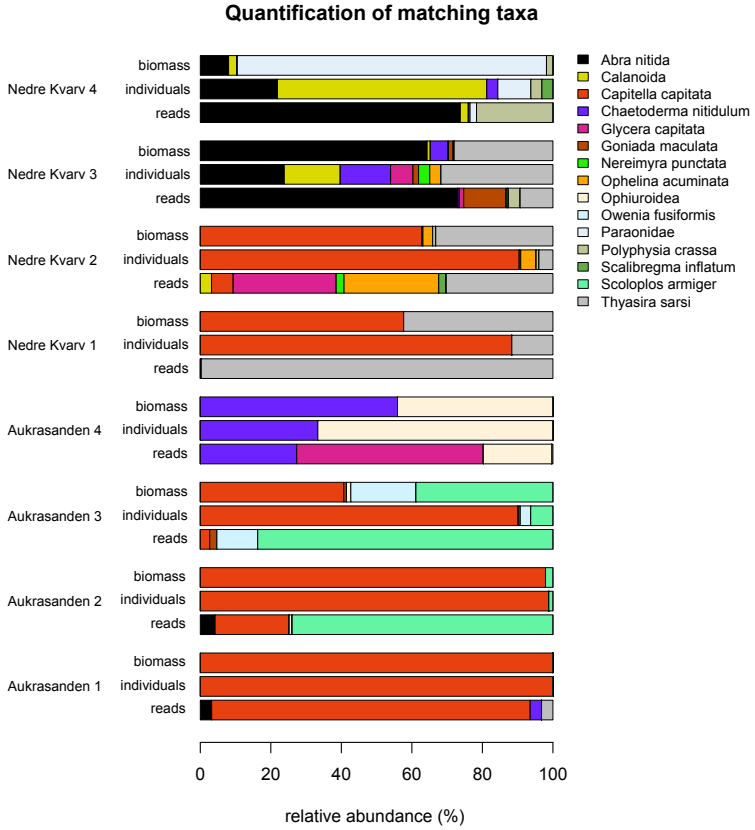


Figure 5. Taxonomic composition of sorted macrofauna samples inferred from sequence data (relative abundance), number of individuals and biomass data for 8 grabs sampled in 8 stations.

The results of our study show important differences in taxonomic composition between morphological and molecular data. In general, the species identified morphologically are also present in metabarcoding data. However, the relative abundance of these species considerably differs between the two approaches. Similar patterns are observed in some low diversity samples (ex. Aukrasanden 1 dominated by Capitella), but no correlation between number of specimens, biomass and number of sequences was found in the majority of samples (Figure 5).

To conclude, the **metabarcoding on fixed bulk samples appears as a useful approach to obtain species list.** However, implementation of this approach would necessitate sieving and sorting macrofauna specimens, thus not facilitating the sampling compared to conventional morphotaxonomic approach. Moreover, the **metabarcoding on fixed bulk samples is not suitable to infer biotic indices,** because the information on species relative abundance is strongly biased by different biological and technical factors.

4.2.2. Macrofauna in sediment samples

We also compared the taxonomic composition of macrofauna as revealed by the metabarcoding of sediment DNA compared to the morphological approach. Presence/absence of the most abundant macrofaunal species in morphological and 18S metabarcoding datasets is shown in Figure 6. Our analyses show limited congruence of the two datasets, which can be explained by several reasons.

First, some species are missing in the metabarcoding data because they have no barcode in the reference sequence database (such as *Galathowenia oculata* and *Chaetozone setosa*, absent in the 18S reference sequence database). These gaps can be filled by barcoding of specimens, as we did in the WP1a. Also, using a second marker might complement the species lists (for example, *Galathowenia oculata* and *Chaetozone setosa* are present in the COI sequence database).

Second, metabarcoding and morphology are two approaches with different sampling strategy. The volume of sediment used for metabarcoding is 5 to 10ml while for morphology, 5 to 10l of sediment are sieved. Morphology counts individuals of macrofauna present in the sampled sediment while metabarcoding counts the DNA sequences of macrofauna present in the sampled sediment, corresponding to their traces left by tissue fragments, organelles, eggs or extracellular DNA molecules. The two approaches provide two not fully identical views on benthic community. On one hand, metabarcoding might not recover the whole assemblage of macrofauna as it is present in the grab. On the other hand, morphology might overlook the morphologically indistinguishable juvenile and life-cycle stages of macrofaunal, detectable in metabarcoding.

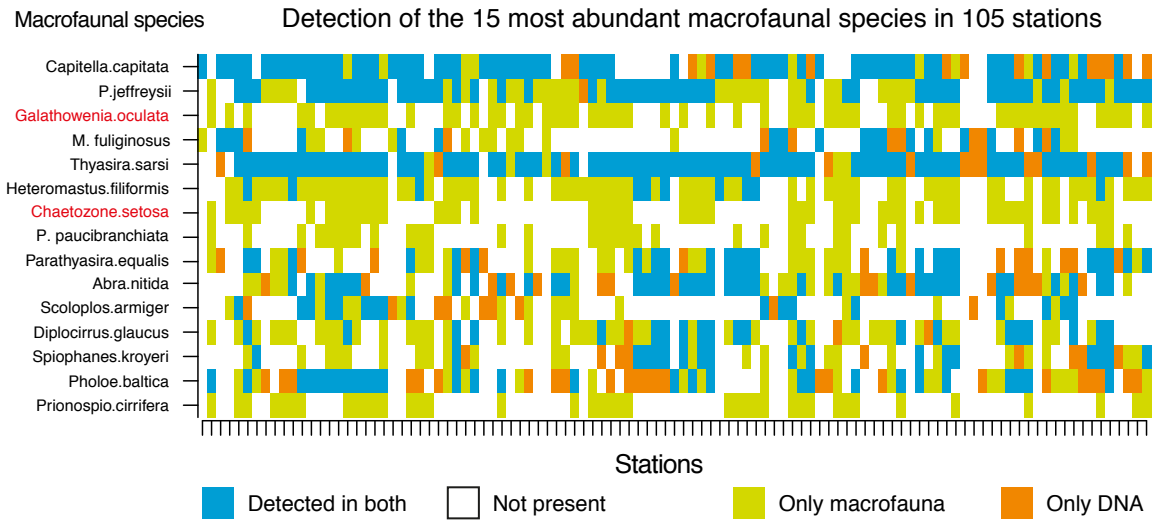
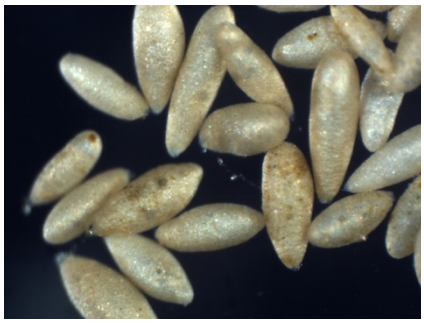


Figure 6. Distribution of the 15 most common species in morphological inventories across 105 stations and their detection in metabarcoding data. The diagram is based on 18S V1V2 dataset. Two species (*Galathowenia oculata, Chaetozone setosa*) absent in the 18S reference sequence database are indicated in red.

4.3. Identification of meiofauna bioindicators (WP2)

The aim of this WP was to explore and identify new benthic bioindicators among meiofaunal taxa present in metabarcoding datasets. Thus, we could expand the list of indicators species with an ascribed ecological category. We focus on two groups of potential meiofaunal bioindicators: foraminifera and metazoans.



Foraminifera

Nematoda

4.3.1. Foraminifera

Foraminifera is a group of unicellular meiofauna that we analysed as potential bioindicators. The foraminiferal datasets have been obtained for 17 sites sampled in 2015-2017 by amplifying and sequencing the hypervariable region 37F of 18S rRNA gene specific to this group. Several studies using foraminiferal metabarcoding to assess the impact of salmon farms in New Zealand, Scotland, Canada and Norway show the potential of this group as bioindicators of organic enrichment (Pawlowski et al. 2014, 2016, Pochon et al. 2015, He et al. 2018).

Our project partly confirmed these results. Relatively good correlation was observed between foraminiferal diversity and biotic indices variation (Figure 7). The foraminiferal OTUs richness, the Shannon diversity and the Chao index tends to decrease with increasing impact, as inferred from NSI, AMBI or Shannon indices.

While foraminifera diversity is consistently less diverse in impacted stations (unlike eukaryotic diversity, see Figure S1), making them promising tool for monitoring, the values of R^2 of linear models (from 0,2 to 0,35) are not sufficiently high for accurate predictions in routine assessment.

Furthermore, the use of foraminiferal communities' composition data to predict biotic indices by supervised machine learning also appears less efficient compared to the whole eukaryotic assemblage (Cordier et al. 2018), which is confirmed here when analysing the full available dataset (see Figure S2).

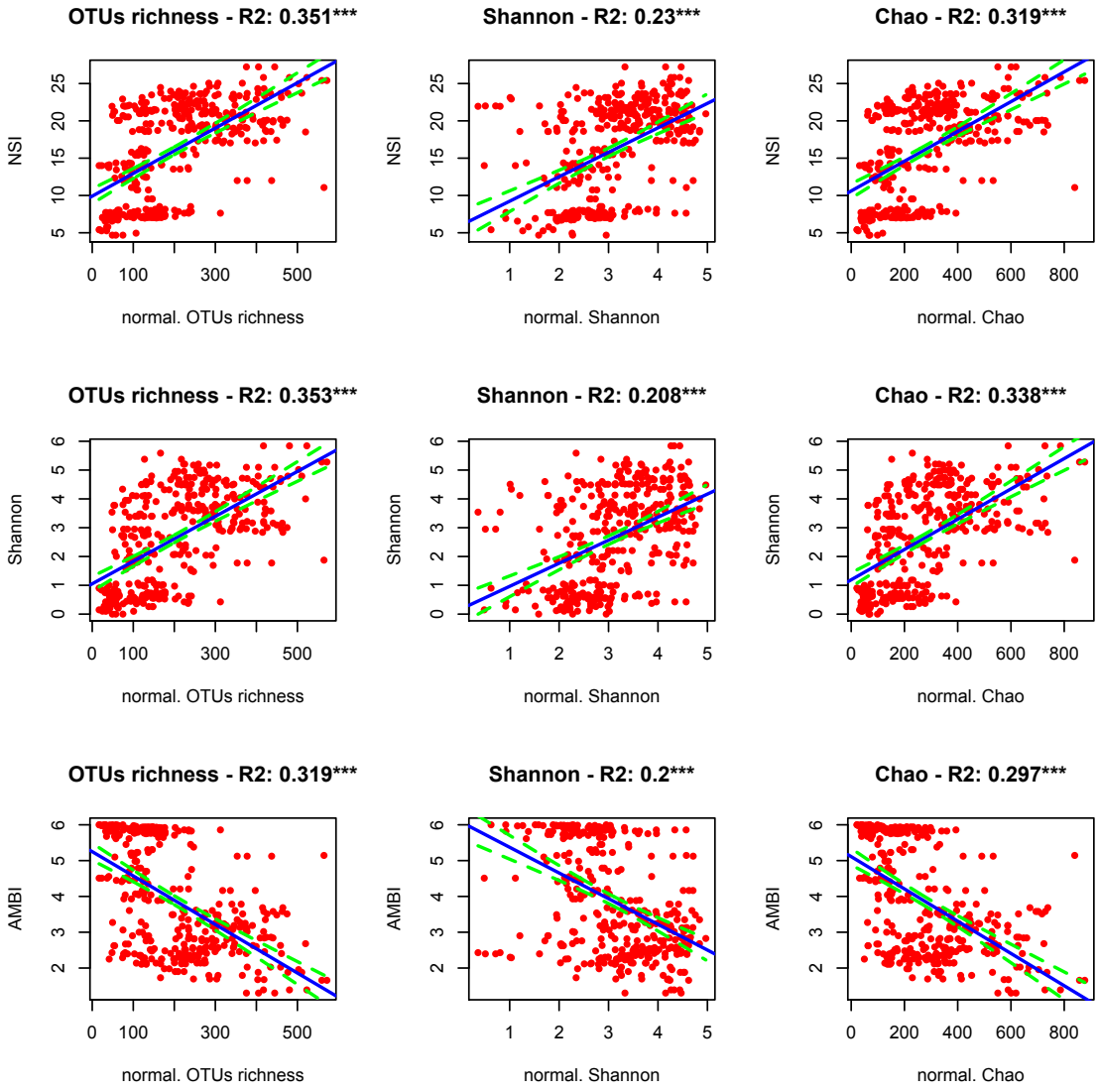


Figure 7. Linear models between foraminiferal diversity metrics (OTUs richness, Shannon diversity and Chao index), with biotic indices variations (AMBI, NSI and Shannon) obtained from macrofauna survey. R^2 and their significance is indicated on the plots. The blue line indicates the fitted linear model and the dashed green lines indicate the 95% confidence intervals.

Moreover, the occurrence of individual foraminiferal OTUs does not seem to be strongly correlated with particular ecological conditions defined by AMBI or NSI indices. We show here two taxa that came up among the best identified bioindicators in our analysis, a rotaliid *Cibicidoides lobatulus* and a monothalamid *Micrometula* sp. *Cibicidoides lobatulus* seems to occur more frequently in impacted zone (Figure 8), yet it is unclear whether this distribution is related to the attached mode of life of this species rather than to its adaptation to organic enrichment. Conversely, the genus *Micrometula* is mainly found in the stations less impacted by organic enrichment (Figure 8). Yet, in both cases, the distribution patterns are not well defined, with DNA profiles not consistently correlated with the gradient of impact.

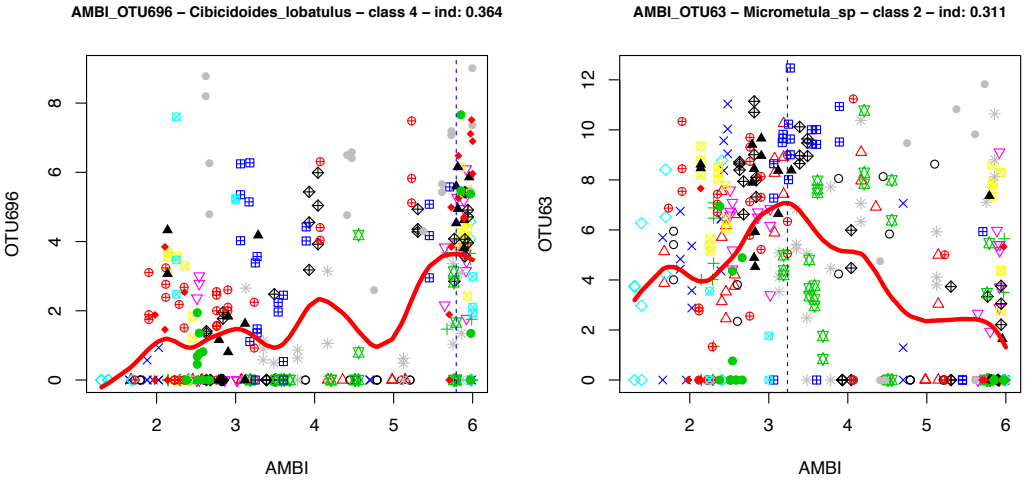


Figure 8. DNA profile along the gradient of impact of two foraminifera taxa identified as good bioindicators in our dataset (*C. lobatulus* and *Micrometula* sp.). Each dot represents the normalized abundance of the taxa (vertical axis) along the gradient of impact (AMBI: horizontal axis) and each colour represents a site. The red curve is a fitted spline to smooth the variation of abundance at a given biotic index value along the horizontal axis. The dashed blue vertical line represents the peak of the spline curve, giving the ecological group of the taxon. Both taxa have a multimodal distribution along the gradient of impact (not a single peak) which means that their ecological preference is not consistently related to organic enrichment.

4.3.2. Metazoan meiofauna

Metazoan sequences have been extracted from the 18S metabarcoding datasets and analysed focusing on the four groups of benthic meiofauna (nematodes, gastrotriches, platyhelminths, and to lesser extent Xenacoelomorpha).

Our first observation is that these four meiofauna groups by far dominate benthic metazoan community in metabarcoding data in the majority of sites. The proportion of meiofauna sequences decreases with the distance from the cage (Figure 9).

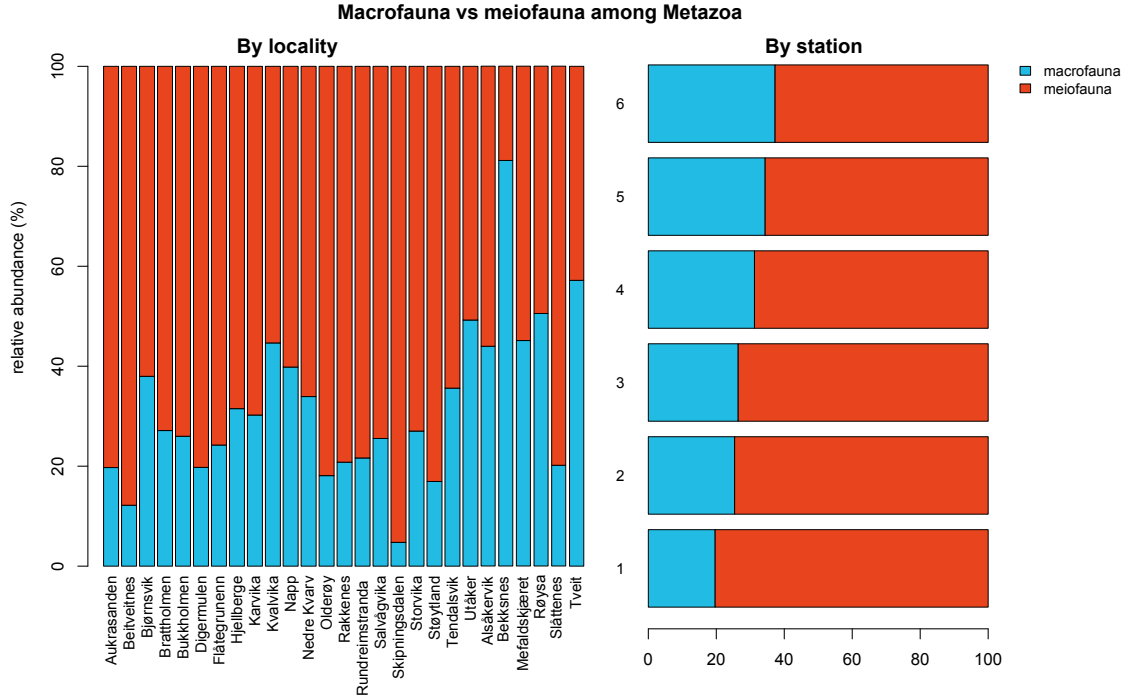


Figure 9. Relative abundance of macro- and meiofauna in metazoan metabarcoding data for 27 sites (left panel) and as a function of station, considered as a proxy for distance to the cages (right panel). The copepods were excluded from analyses because of the difficulty to distinguish planktonic from benthic species in metabarcoding data.

The quantitative analysis of sequences assigned to the groups of benthic meiofauna shows clearly that nematodes dominate the assemblage (red in Figure 10). Their relative abundance increases close to the cages, confirming previous observations that many nematodes are opportunistic species dominating in organic polluted areas. The abundance of gastrotrichs in metabarcoding data increases with the distance from cages, while the abundance of platyhelminths remains more or less stable.

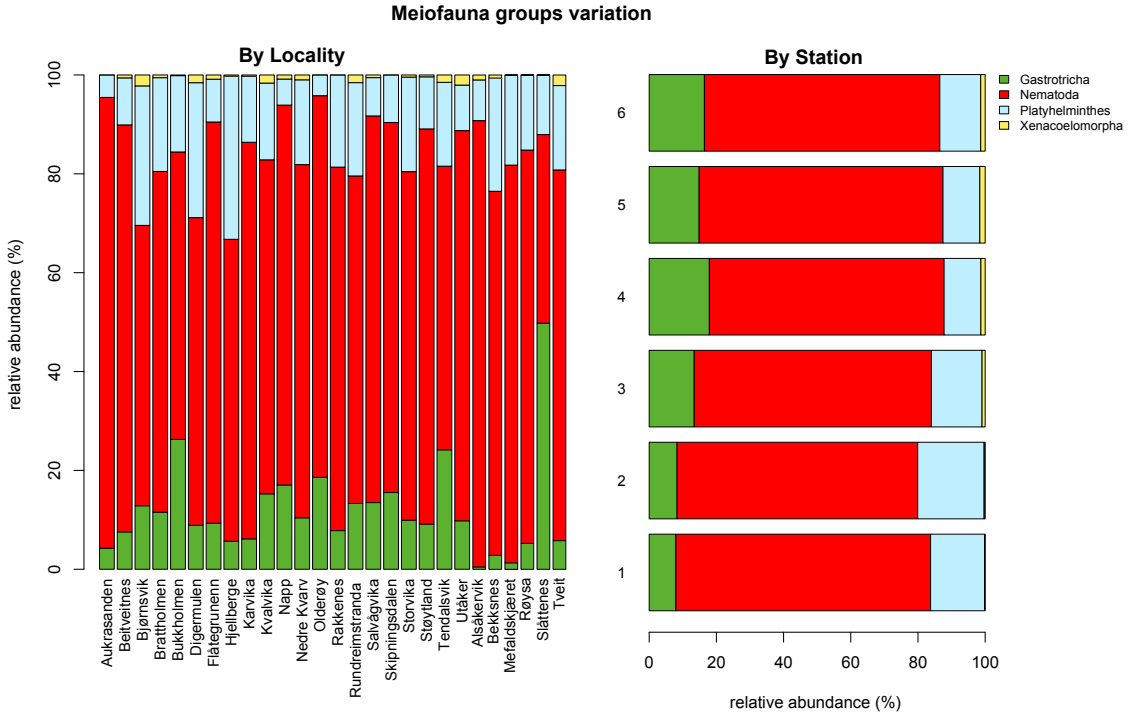


Figure 10. Relative abundance of gastrotriches, nematodes and platyhelminths in metabarcoding data for 27 sites (left panel) and as function of station, considered as a proxy for distance to the cages (right panel).

The response to organic enrichment is clearly visible in the spline representation of the distribution of particular nematode taxa (Figure 11).

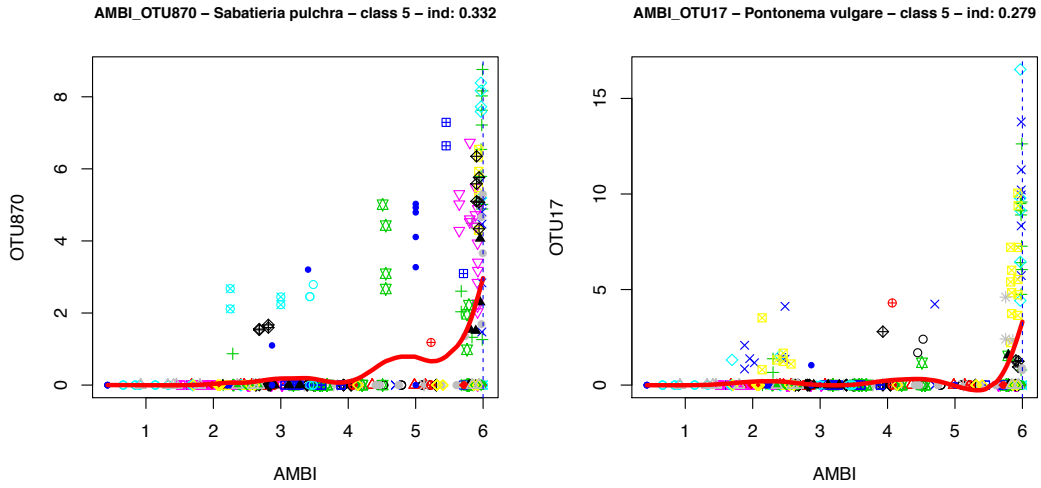


Figure 11. DNA profile along the gradient of impact of two nematode species identified as good bioindicators of organic enrichment in our dataset (*S. pulchra* and *P. vulgare*). Each dot represents the normalized abundance of the taxa (vertical axis) along the gradient of impact (AMBI: horizontal axis) and each colour represents a site. The red curve is a fitted spline to smooth the variation of abundance at a given biotic index value along the horizontal axis. The dashed blue vertical line represents the peak of the spline curve, giving the ecological group of the taxa.

As opposed to the foraminifera taxa analysed above, these two nematodes have a unimodal distribution along the gradient of impact (a single, well identified peak) which means that their ecological preference is consistently related to organic enrichment. This makes them robust bioindicators of the impact associated with fish farming.

Several nematodes occur almost exclusively in the zone of high organic enrichment (AMBI/NSI group 5) – see Table 2. Some of these species have long been recognized as indicators of organic enrichment. *Sabatieria pulchra* is placed among opportunistic species belonging to ecological group 5 of AMBI index. The highly consistent occurrence of *Pontonema vulgare* in organically polluted fjords has been reported by Lorenzen et al. (1987). Yet, until now, none of these species are included in the current biotic indices (Rygg and Norling , 2013), mainly due to the difficulties of identifying them morphologically.

Table 2. Top 10 species of nematodes selected as indicators of organic enrichment using spline analysis of metabarcoding data.

SPECIES	**OTU**	**AMBI group**	**AMBI value**	**NSI group**	**NSI value**
Anticoma sp.	OTU51	5	5.94	5	9.02
Enoplolaimus sp. WUS5	OTU349	5	5.79	5	7.78
Halomonhystera disjuncta	OTU1300	5	6.00	5	6.56
Molgolaimus demani	OTU124	5	6.00	5	4.68
Oncholaimellinae sp. AS71	OTU61	5	5.71	5	7.66
Paracanthonchus caecus	OTU112	5	6.00	5	6.71
Pontonema vulgare	OTU17	5	6.00	5	6.56
Prochaetosoma sp. 2	OTU212	5	5.81	5	7.16
Sabatieria pulchra	OTU30	5	6.00	5	5.41
Spirinia parasitifera	OTU138	5	5.65	4	14.11

In general, all nematodes taxa are considered as a homogeneous indicator group, as exemplified by the AMBI index in which all the taxa are classified in group 3. Yet, according to our data, some nematode species consistently occur in non-impacted stations, which means that these nematode taxa have a clear ecological preference for organically non-enriched sites (Figure 12). Hence, our results suggest that nematodes are highly variable group that cannot be ascribed to a single ecological category.

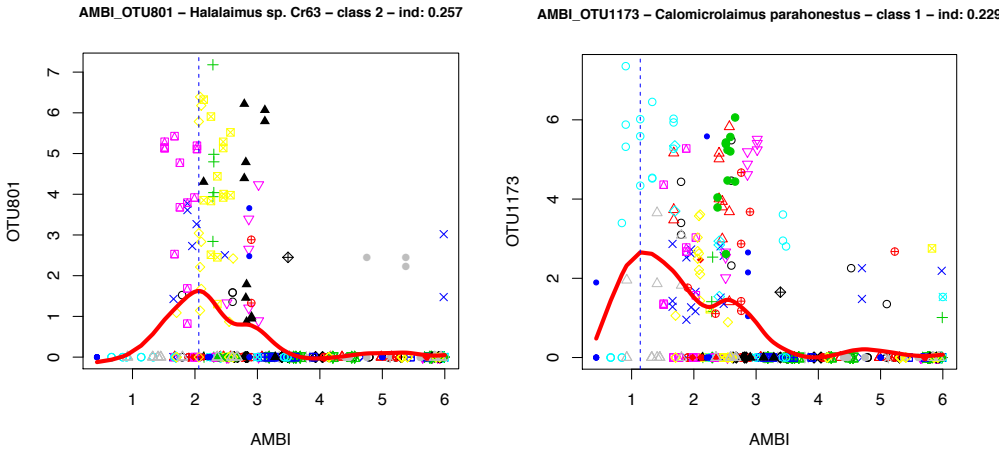


Figure 12. DNA profile along the gradient of impact of two nematode species identified as potential non-organic enrichment indicators in our dataset (*Halalaimus sp. and C. parahonestus*). A consistent peak is observed in non-impacted stations, which means that these two nematodes taxa shall be classified in AMBI category 1-2.

Overall, we conclude that **nematodes could be useful for DNA-based benthic monitoring of salmon farms**. First, they comprise large diversity of species and these species have relatively well-defined ecological preferences, including species present exclusively in organically enriched zone and species present mostly outside this zone.

Second, nematodes are well represented in metabarcoding data, as they represent the most abundant group of metazoan sequences in our dataset. Third, the nematode species can be accurately discriminated using standard eukaryotic 18S barcodes, even if many of them are absent from reference database.

In addition to nematodes, we also analysed two other meiofauna groups: the gastrotriches and the platyhelminths. The gastrotriches comprise much lower number of species compared to nematodes and are rarely present in organically enriched zone (Table S2). The diversity of platyhelminths is also quite low, with a variety of species associated with different ecological groups, sometimes even within the same genus (ex. *Microstomum*) (Table S3). Overall, both groups seem much less useful as bioindicators of organic enrichment compared to nematodes.

4.4. Revision of benthic indices (WP 3)

The aim of this WP was to develop a novel category of ecogenomic indices that would fit best the specificity of metabarcoding data. The metabarcoding approaches discussed in this section rely on extracting environmental DNA from sediment.

4.4.1. Different metabarcoding approaches for inferring biotic indices

In our preliminary study, the values of classical benthic indices (AMBI, ITI) were inferred from **metabarcoding data assigned to macrofauna** species with known ecological values (Lejzerowicz et al. 2015). The results of this study were very promising showing a good congruence between index scores inferred from morphological and metabarcoding data. Yet, a very small proportion (1-2%) of DNA sequences could be assigned to macrofaunal taxa which prompted us to look for more holistic and robust method.

During this project, we have developed an alternative method to predict benthic indices using a **supervised machine learning (SML)** approach. This approach consists in directly predicting the values of biotic indices from the whole assemblage of sequence data, regardless of the taxonomic affiliation of the sequences. Using a training set of samples with metabarcoding data combined with morphologically inferred indices, a predictive model is trained to predict biotic indices values for new upcoming samples (Figure 13).

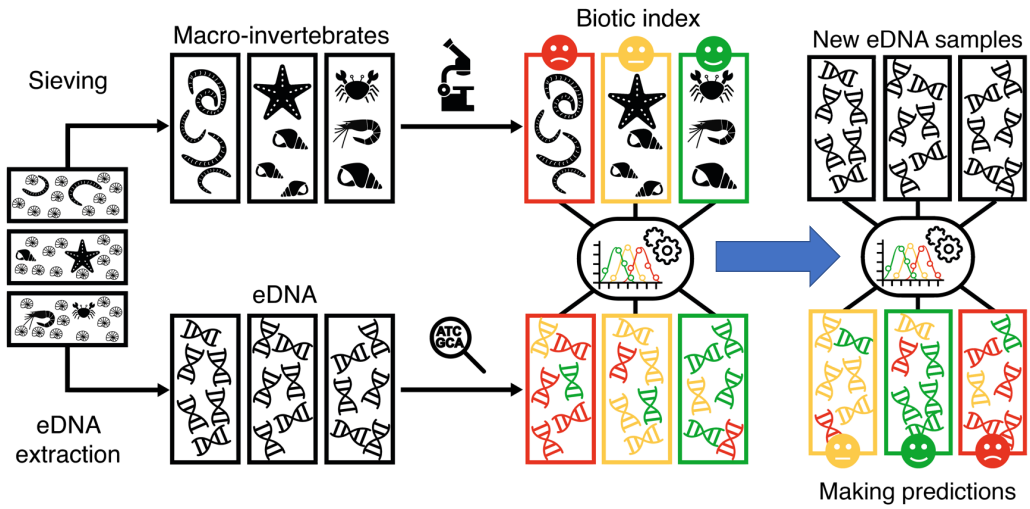


Figure 13. The workflow of the supervised machine learning approach. The training dataset contains both morphological inventories of macro-invertebrates and metabarcoding data from the same grabs (left side). The training of the predictive model by the machine learning algorithm (middle) consists in automatically finding correlation (linear or not) and association rules that explain the biotic indices variation. Finally, the trained model can be used to make prediction on new eDNA samples (right side).

In the proof of concept study, we compared the co-occurrence and two SML approaches (Random Forest and Self-Organizing Map) applied to foraminiferal metabarcoding data for five salmon farming sites. The results of this study show high

level of agreement between the molecular and morphological data for four common benthic indices (AMBI, ISI, NSI, NQI1) (Cordier et al. 2017, Table S4).

Given these promising results, we further investigated the potential of SML approaches for biomonitoring by comparing the performance of using metabarcoding data assigned to known macrofaunal bioindicators with the performance of SML inferences from the whole metabarcoding datasets. We also compared the results using five different genetic markers, that target different groups, i.e. three eukaryotic, one foraminiferal and one bacterial (Cordier et al. 2018). We found that for all tested genetic markers, the performance of SML was better than macrofaunal inference of biotic indices (Table 3). Most importantly, the macrofauna-based inference uses only a small fraction of the generated data (from 10,8% with V4 up to 20,8% with V1V2), while SML uses all the available data, hence being more holistic. Moreover, the SML provides much more accurate predictions (Kappa between 0,21 and 0,569 for macrofauna-based inference and above 0,8 for SML (Table 3)).

Table 3. Results of biotic index value predictions performance between metabarcoding targeting macrofaunal bioindicators and SML-based predictions, by genetic marker. The three eukaryotic markers allow to assign sequences to macrofaunal species. The average R^2 and K (Kappa) are the arithmetic mean over four biotic indices (AMBI, NSI, ISI, NQI1).

Genetic marker	Inference	Taxonomic group	OTUs	Reads	% reads	Average R^2	Average K
Eukaryotes V1V2	BI Inference	Macrofauna	405	2803510	20,8	0.394 (±0.15)	0.477 (±0.16)
Eukaryotes V4	BI Inference	Macrofauna	199	956980	10,8	0.513 (±0,37)	0.569 (±0,38)
Eukaryotes V9	BI Inference	Macrofauna	191	5005194	11,7	0.22 (±0.06)	0.21 (±0.1)
Foraminifera 37F	SML	All	1594	7989253	-	0.782 (±0.09)	0.815 (±0.02)
Bacteria V3V4	SML	All	3630	1707760	-	0.862 (±0.05)	0.877 (±0.04)
Eukaryotes V1V2	SML	All	2680	13460533	-	0.867 (±0.07)	0.859 (±0.03)
Eukaryotes V9	SML	All	3588	42913527	-	0.897 (±0.07)	0.887 (±0.04)
Eukaryotes V4	SML	All	2031	8850988	-	0.875 (±0.11)	0.852 (±0.07)

All these results were also confirmed using the complete V1V2 18S dataset obtained during the course of the project, consisting of 27 sites (Figure 14). Both the R^2 and Kappa statistics are significantly higher when using SML compared to those obtained when using only the sequences assigned to macrofaunal taxa to calculate biotic indices (Figure 14).

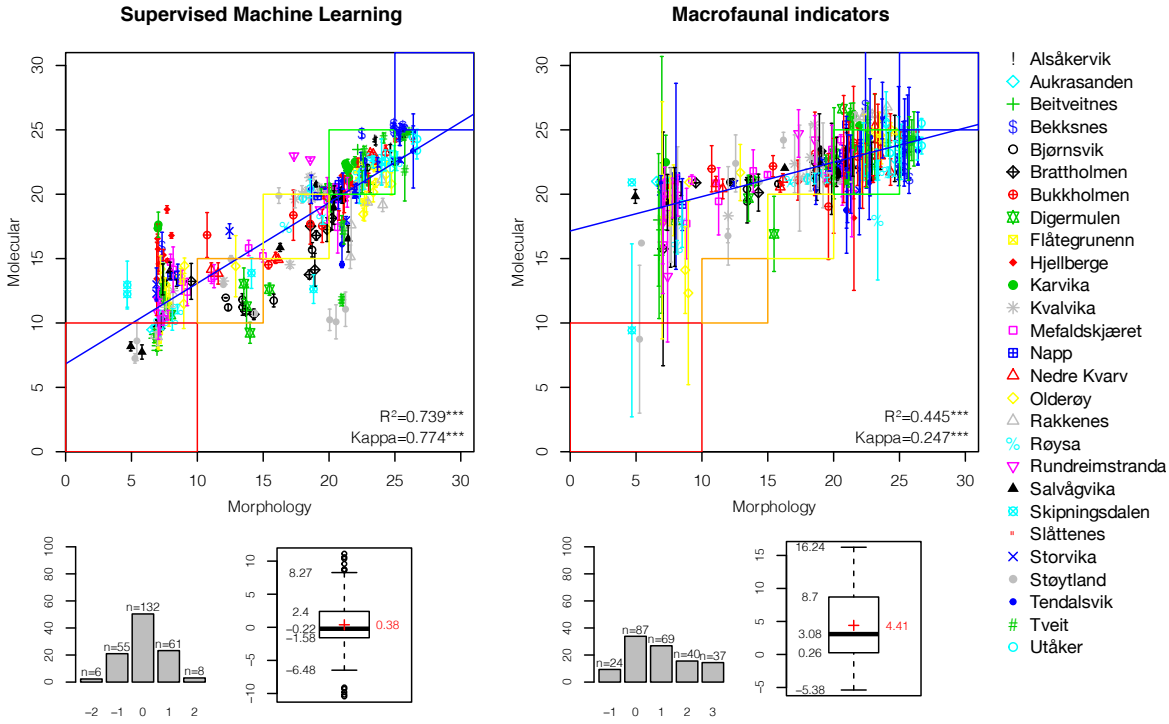


Figure 14. NSI predictions performance comparison between the SML (Random Forest) and the macrofauna-based 18S metabarcoding approaches for the full 27 farms dataset. Both R^2 and kappa statistics with significances are indicated as inset in each plot. The histogram on the bottom left of each plot indicates the congruence between the reference discrete status value and the predicted one (number indicates the mismatches, zero being perfect classification). The bottom right boxplots indicate the distribution of errors (the difference between the reference NSI values and the predicted ones) with medians and quantiles in black and average in red.

To conclude, our studies suggest that **supervised machine learning represent the best approach at hand to implement metabarcoding for predicting benthic indices in routine biomonitoring**.

4.4.2. Congruence of SML-predicted and morphology-inferred biotic indices

In addition to NSI predictions performance of SML using Random Forest algorithm on V1V2 18S metabarcoding data depicted in Figure 14, its performance for predicting ISI, Shannon, AMBI and NQI1 indices is shown in Figure S3.
Congruence of SML-predicted and morphology-inferred AMBI values can be appreciated farm by farm on Figure 15. The majority of stations are assessed to the same ecological status or differ by one class. There is no evidence that metabarcoding data consistently provide lower assessment of ecological status than morphology.

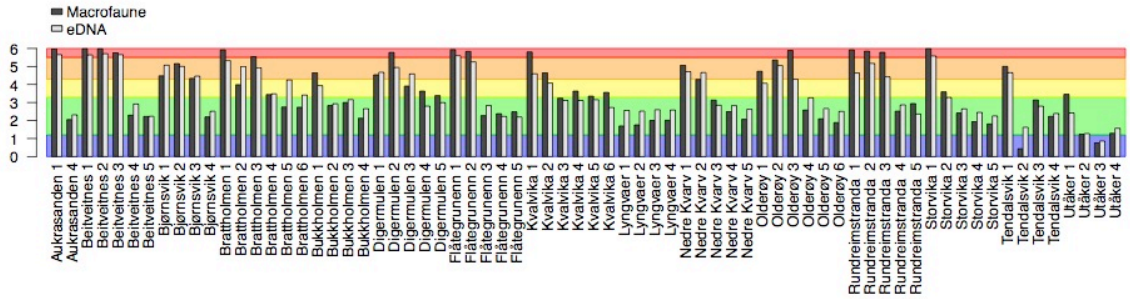


Figure 15. AMBI values inferred with morphology (black) and eDNA (white) for 70 stations from 15 farms in region Mid. As shown in the figure **only 2 out of 70 stations** (Olderøy 3, Brattholmen 5) are classified differently (more than 1 class) and **only 2 out of 32** stations (Brattholmen 5-6) are underscored by SML from good to moderate status.

Nevertheless, the metabarcoding tends to reduce the extreme values. In Figure 16, NSI, Shannon and AMBI indices values are sorted according to ecological categories. In very bad and bad classes, the values of SML-predicted indices are always lower than for morphologically inferred indices, while in good and very good classes, the values of SML-predicted indices are higher than for morphologically inferred indices. The best congruence between the two approaches is obtained in good (green) class, while the less congruent results are observed in moderate class, with the majority of stations assigned by SML to better conditions than based on morphology.

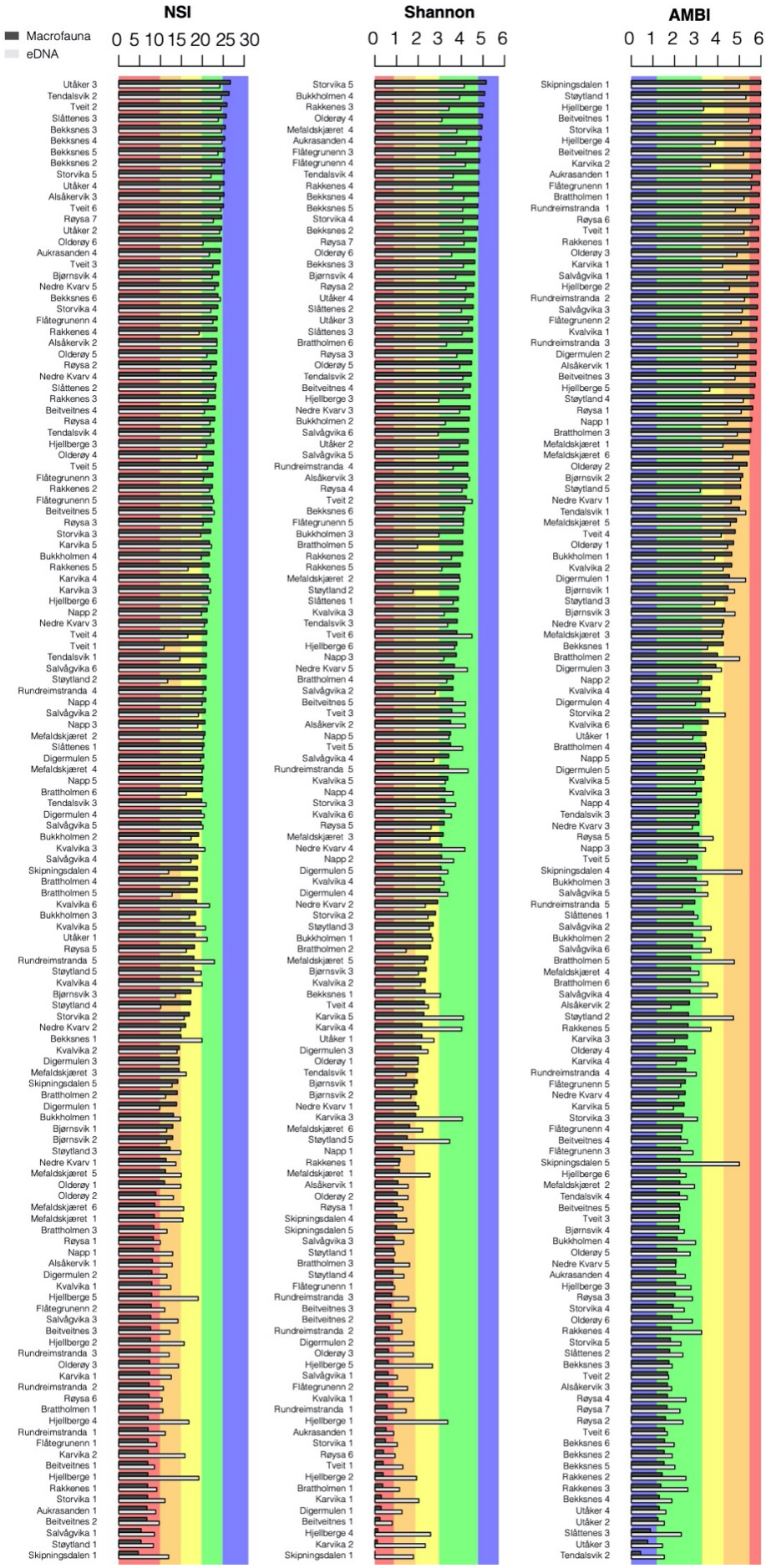


Figure 16. NSI, Shannon and AMBI indices values inferred with morphology (black) and eDNA (white) for 132 stations from 27 farms, sorted according to ecological categories.

Introduction of new class limits

In 2018, Norwegian directive (Veileder 02:2018) updated the ecological class limits for NSI, NQI1, ISI and Shannon indices for every specific water type. The water types are classified according to different coastal regions and the exposure to the current flows, to freshwater outflow, etc. Table 4 shows the 27 sites examined in this project are distributed unevenly between different water types. Water type-wise choice of reference sites for future extension of the training data will help to cover the underrepresented water types.

Table 4. Distribution of the 27 sites between different water types.

Water type	Marine region	Number of sites	Site names
B1-5	Barentshavet	3	Hjellberget, Karvika, Rakkenes
G1-3	Norskehavet Nord	1	Bjørnsvik
G4-5	Norskehavet Nord	1	Nedre Kvarv
H1-3	Norskehavet Sør	9	Aukrasanden, Beitveit, Brattholmen, Bukkholmen, Digermulen, Kvalvika, Mefaldskjæret, Røysa, Storvika
H4-5	Norskehavet Sør	0	
M1-2	Nordsjøen Nord	3	Flåtegrunnen, Olderøy, Rundreimstranda
M3-5	Nordsjøen Nord	2	Bekksneset, Tveit
N1-2	Nordsjøen Sør	2	Skipningsdalen, Napp
N3-5	Nordsjøen Sør	6	Alsåkervik, Salvågvika, Slåttenes, Støytland, Tendalsvik, Utåker
S1-3	Skagerrak	0	
S5	Skagerrak	0	

As shown in figure 17, the new class limits have some impact on ecological status. Overall, the new "very good" class limits are less stringent than the previous ones. Indeed, more stations are assigned to "very good" ecological status with both morphology and eDNA inferences than with the previous class limits.

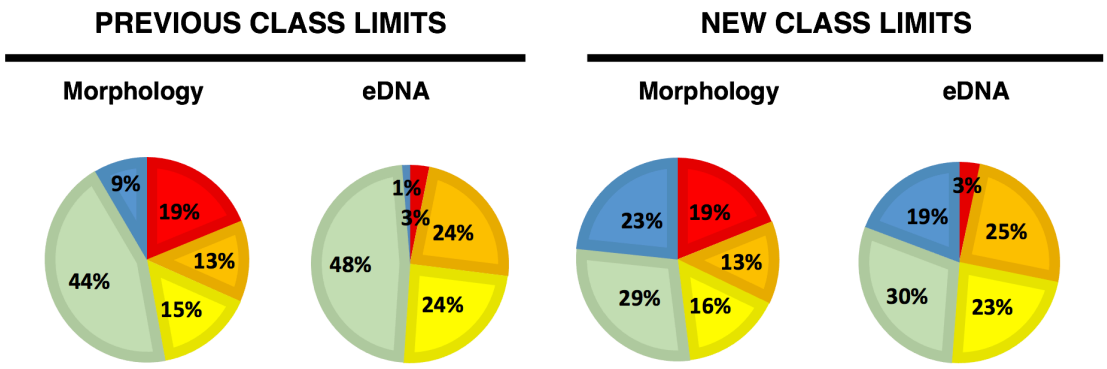


Figure 17. Distribution of the five ecological classes as inferred with morphology and eDNA for all five indices together, according to previous and new class limits. Table S5 details this distribution for each index separately.

Inter-variability of macrofaunal indices

Furthermore, we analysed the inter-variability of macrofaunal indices for a given station, both inferred from morphology and predicted by eDNA (Figure 18). We classified the stations according to the ecological status indicated by five indices. When all indices indicate the same status, the station is represented by 1 class. When indices show 2 or 3 different ecological statuses for the same station, the station is classified in 2 or 3 classes, respectively.

As shown in figure 18, the new class limits have some impact on the congruence between indices. In the previous class limits, all five indices give the same ecological status in 24% of morphology inferences and in 43% for eDNA-based predictions. When applying the new class limits, we observe a decrease of the number of stations assigned to the same ecological status with the five indices. However, for both previous and new class limits, the proportion of stations assigned to a single ecological status is always higher in eDNA compared to morphological data.

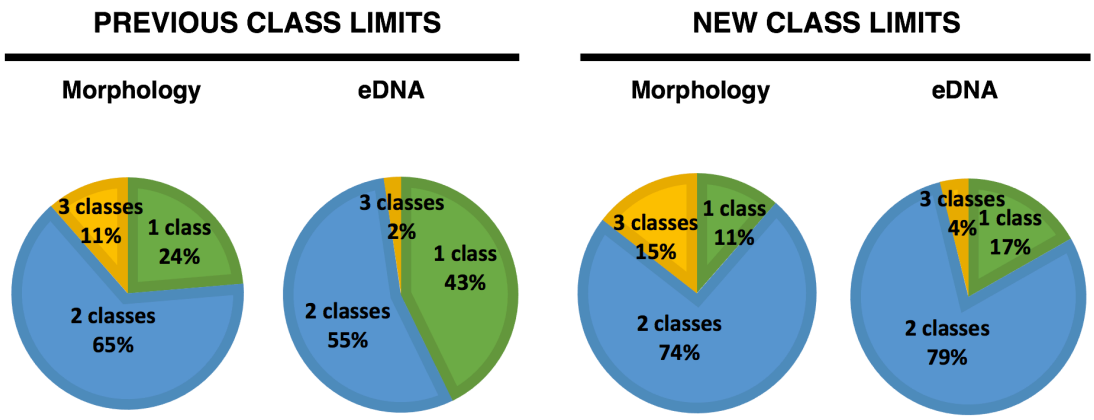


Figure 18. The congruence of different indices for a given the 132 stations, illustrated by the proportion of stations assigned the same ecological class with all 5 indices (green), to two ecological classes (blue) and to three ecological classes (yellow). the congruence of different indices for a given station

The inter-variability of macrofaunal indices for each of 132 station is illustrated in Figure 19. Overall, we observe a relatively good congruence between indices, especially for eDNA datasets. Still, some indices seem more congruent than others. In particular, NSI and AMBI usually provide the same ecological status, while Shannon (H') index often differs from both of them, indicating the status similar to NQI1 and ISI.

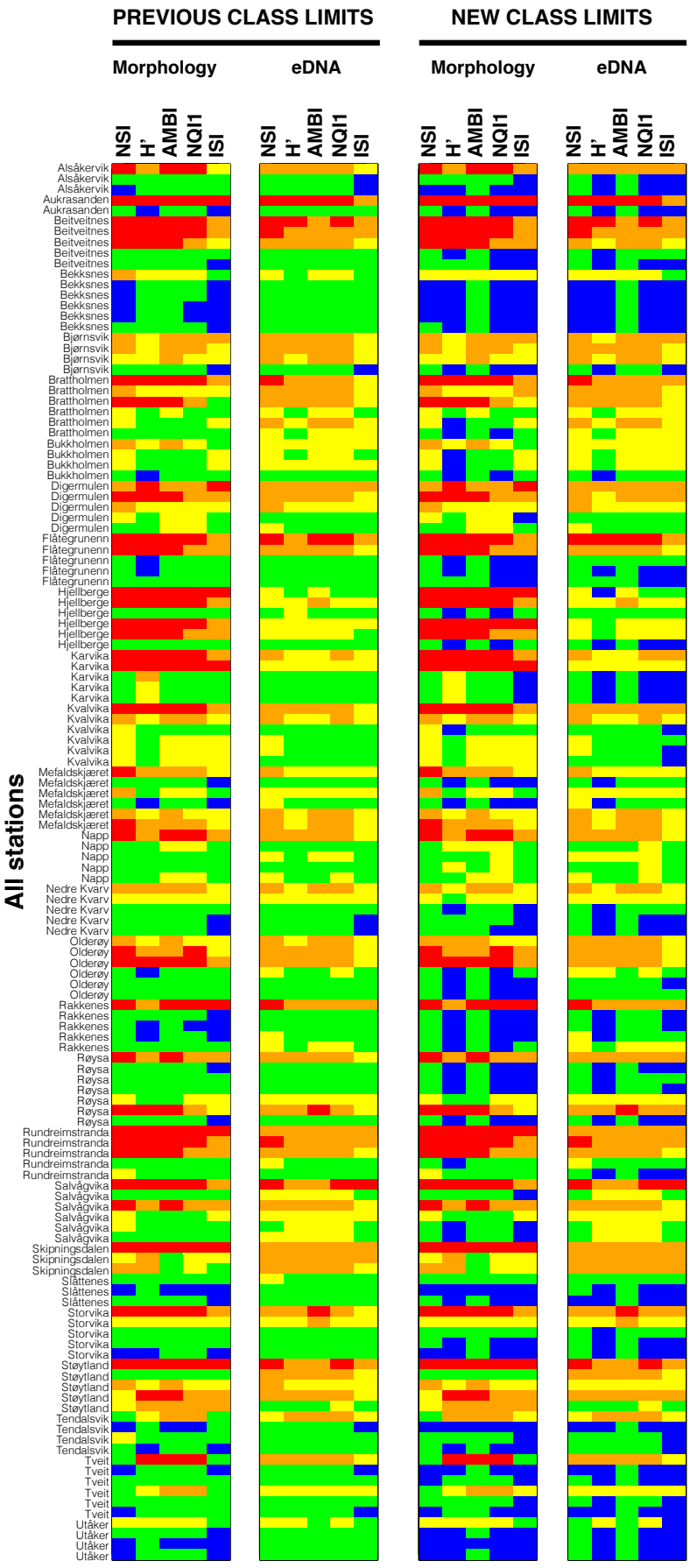


Figure 19. Congruence of the ecological status for NSI, Shannon, AMBI, NQI1 and ISI indices for all the stations (n=132) of the 27 farms, according to the previous class limits and the limits updated in 2018.

Impact of training dataset size

We investigated the effect of the training dataset size on the accuracy of predictions. We used the full dataset (27 farms from the Norwegian coast) or a random subset as training dataset and calculated the error as a function of the number of sites included (Figure 20). The trend shows a consistent decrease in the predictions error with increasing number of sites included in the training dataset. Even if the improvement seems to reach a plateau with the full dataset of 27 sites, adding more sites might further improve the accuracy, notably by accounting for more possible environmental variations, seasonality effects, and local sites peculiarities.

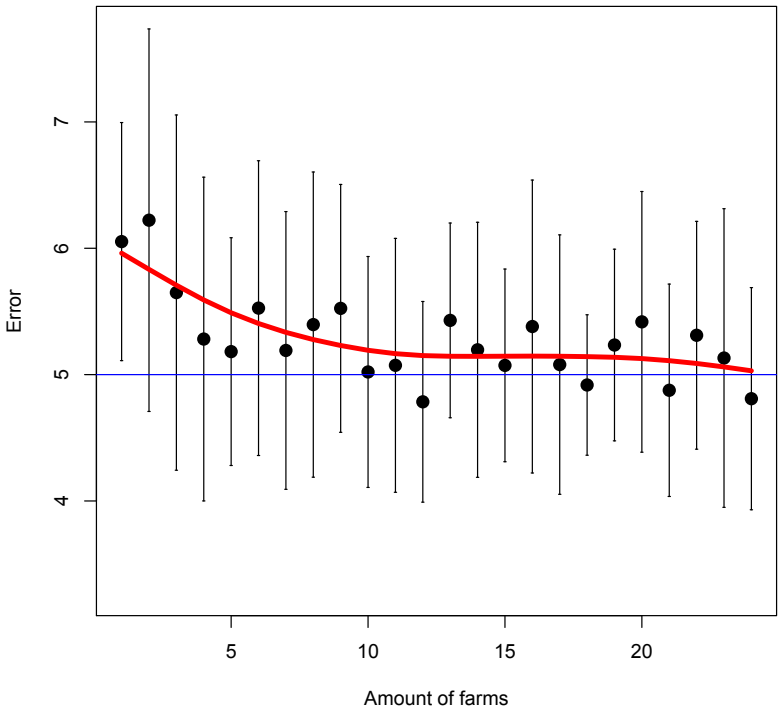


Figure 20. Error of NSI prediction estimation as a function of the number of farms in the training dataset. Each dot represents the mean error and the bars represent the 95% confidence intervals. The blue line represents an error of 5 in NSI, which corresponds roughly to one NSI class size.

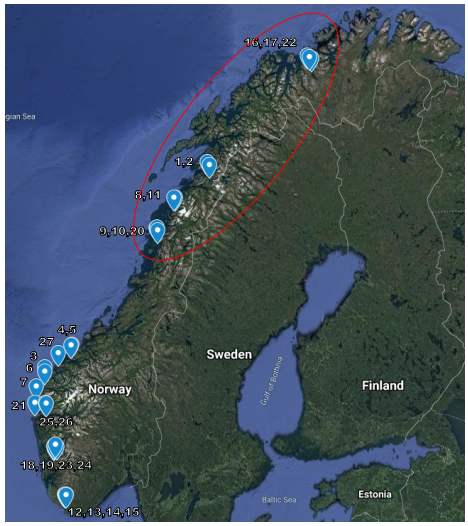
In conclusion, SML-predicted and morphology-inferred indices values have a good congruence, especially for the good ecological class, and the congruence between indices is significantly higher with the SML approach than with morphology. Further improvements could be obtained by completing the training dataset with reference localities from under-sampled regions and water types. Indeed, the error in predictions shows a consistent decrease with increasing number of reference sites.

4.4.3. Correlation of metabarcoding data and environmental parameters

We examined the correlation between the sediment properties, e.g. normalized Total Organic Content (N.TOC), granulometry (Fine particles), phosphorus (P), copper (Cu) and zinc (Zn), and the structure of benthic communities inferred with metabarcoding, as well as the effect of latitude and distance from the cages (Table 5, Figure 21). Our analyses reveal that the location of the site along the latitudinal gradient is the strongest determinant of the benthic community's structure (R^2 = 0,7, Table 5). The distance to the cage and the granulometry were also significant (Table 5), while noteworthy none of chemical parameters, including the N.TOC, was significantly correlated with the structure of benthic communities.

Table 5. Environmental parameters correlation with eukaryotic community structure

Environmental parameter	R^2	p-value
Fine particles	**0.2213**	**0.001*** **
normalized TOC	0.0109	0.158
Copper	0.0036	0.565
Phosphorus	**0.0192**	**0.039***
Zinc	0.0021	0.713
Latitude	**0.7012**	**0.001*** **
Distance from cages	**0.1358**	**0.001*** **



Benthic communities structure inferred with metabarcoding for the 27 sites was analyzed by non-metric multidimensional scaling analysis in Figure 21. Each point on the plot corresponds to the benthic community of a given sample (grab in this case) colored either according to the sampling site (first plot) or to the NSI value inferred with morphology (three last plots). All the sampled benthic communities tend to cluster in two distinctive groups. The group encircled in red on the first plot corresponds to the samples coming from the region encircled in red on the map. Thus, the benthic communities revealed by metabarcoding reflect the biogeography of sampled sites covering about 4 latitude degrees on the south of Norwegian coast and about 5 latitude degrees on the north of the Norwegian coast. An unsampled gap of about 3 latitude degrees is in the middle.

We also represented the values of three environmental parameters (N.TOC, fine particles and phosphorus) as lines on the plot of benthic communities' structure. Low values of N.TOC (below 10mg/g) and high percentage of fine particles (above 60%) correlate with higher NSI values. Relationship between phosphorus and the NSI values of the benthic communities seems to be more complex.

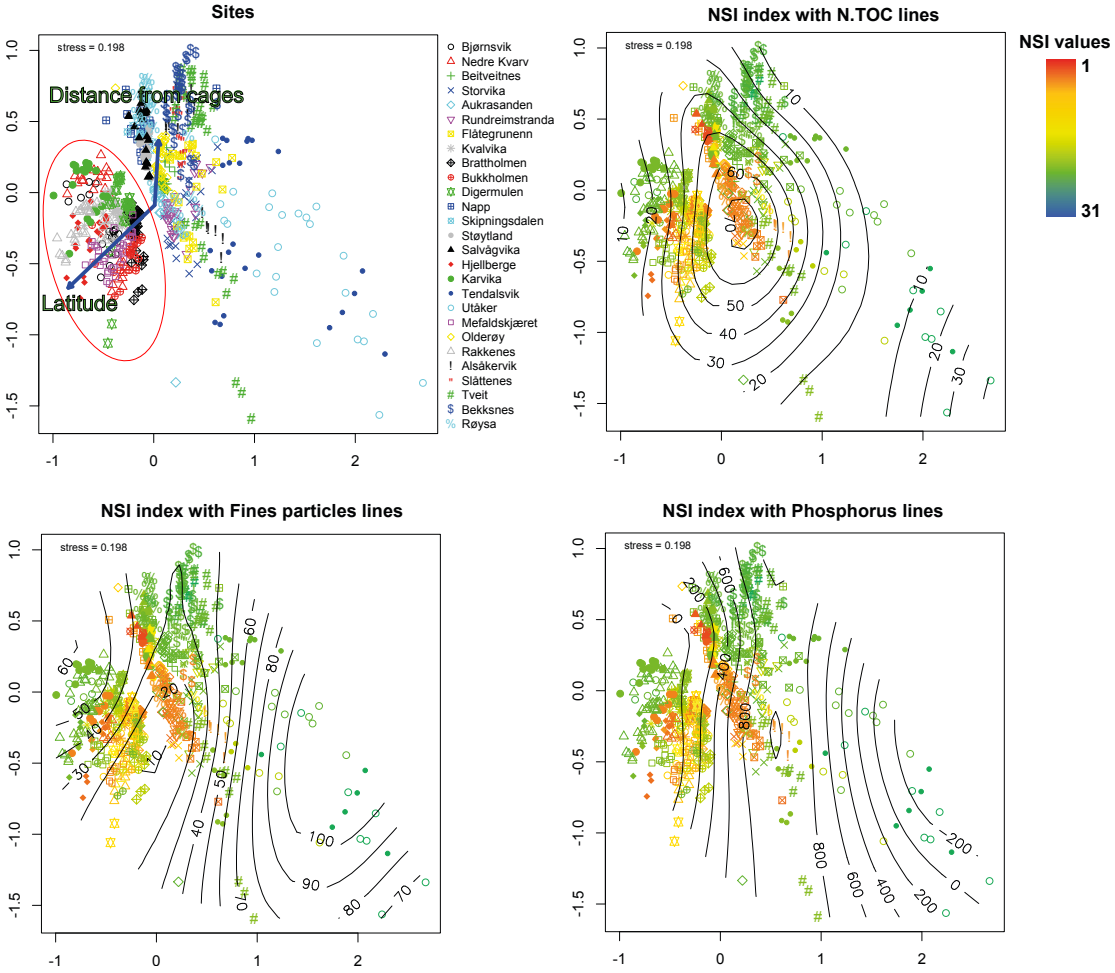


Figure 21. Non-metric multidimensional scaling analysis of the benthic communities' structure. On the top left, the colors indicate the farms localities and the arrows indicates the direction of the correlation with the distance from the cages and the latitude. On the top-right, bottom-left and bottom-right, the colors gradient indicates the values of NSI inferred with morphology and the grey lines indicates a smooth fitted surface that represent the N.TOC (mg/g), fine particles (%) and phosphorus (mg/kg) environmental parameters (see also Table 5).

In conclusion, these results illustrate that **benthic communities are shaped mainly by biogeography**, but other geochemical parameters might also have impact on their structure. Therefore, higher will be the number of reference sites in the training dataset, better will be the coverage of environmental and spatio-temporal variations, improving the performance of the SML predictions of biotic indices.

5. Implementation of the method

5.1. Stakeholders involvement

The development and validation of new tools for benthic monitoring require a challenging mix of science, political support and stakeholders involvement. Throughout the duration of the project, the project team deployed consistent efforts to inform, consult and collaborate with the stakeholders interested in the new method.

The Steering Committee led by Dr Catarina Martins, Group Manager Environment and Sustainability at MOWI organized several meetings with regional managers, operators and stakeholders. Participation of MOWI regional managers in Steering committee at an early stage helped to optimize plans for sampling trips by bringing up a thorough knowledge of local conditions. The discussions further allowed to familiarize future users with methodological issues regarding eDNA-based biotic indexes. Four general meetings have been organized in MOWI headquarters, including a meeting with representatives of different governmental agencies and consulting companies involved in environmental monitoring of salmon farms. The stakeholders' collaboration framework put in place proved to be an efficient mechanism to build up high acceptance level for new method. MOWI also promoted the project through Annual Reports and by organizing a special session during the EAS conference in 2017.

Consulting Companies. In parallel, several meetings have been organized with the Norwegian consulting companies involved in environmental monitoring, including Akerbla, Aqua Kompetanse, and Akvaplan-niva. These direct contacts led to building trust in application of new method in routine biomonitoring and to conducting independent validation tests comparing the morphological and molecular approaches. The companies brought into consultative process their knowledge of the complex certification processes at national and international level, and their input regarding technical solutions. On their recommendation also, the eDNA assessment reports were considerably simplified and information synthetized. In 2017-2018, more than 30 sites have been assessed using eDNA method, mainly as a part of the ASC surveys (see below). Recently, the authors of the project have been invited to participate in the construction of eDNA lab that could conduct the tests in Norway.

Scientific community. Significant effort was made to ensure robust and transparent process of assessing the scientific readiness of the method. The best experts in the eDNA field have been invited to join the **Scientific Panel**. Prof. Thorsten Stoeck (University of Kaiserlautern, Germany), Dr. Anders Lanzen (AZTI, Spain), Dr. Nigel Keeley (IMR, Tromso) and Dr. Thierry Baussant (NORCE Research AS, Stavanger) accepted the invitation and participated in the meeting in Bergen, in September 2018. The members of Scientific Panel drafted a report, which presented the method and discussed its advantages and main challenges. In conclusion, the authors of the report consider the method as sufficiently mature to be implemented in routine biomonitoring. They suggest that the method should be introduced gradually with certain number of farms tested in parallel for molecular and morphological analyses.

5.2. ASC compliance

One of the main steps towards the implementation of the method was the acceptance of metabarcoding as an alternative method to assess the impacts on benthic diversity in compliance with Aquaculture Stewardship Council (ASC) salmon standard requirements.

In 2017, the ASC approved the variance request proposing to use molecular techniques such as metabarcoding for benthic monitoring of salmon farms in Norway. According to the VR nr.226, the metabarcoding data can be used instead of macrofauna to comply with indicators 2.1.2 and 2.1.3 of the ASC salmon standard. In the case of indicator 2.1.2, the threshold values of biotic indices inferred or predicted from molecular data should be same as required by traditional approach. The compliance with indicator 2.1.3. is verified by demonstrating the presence of DNA sequences assigned to non-pollution indicators among the abundant species.

So far, 33 sites with total of 182 stations have used the eDNA bioassessment test for an ASC survey. ASC audit reports of certified sites using this test are publicly available on ASC webpage.

When applied to the ASC surveys, the metabarcoding analyses are conducted on two markers: nuclear 18S rRNA gene and mitochondrial COI gene. The variable region V1V2 of 18S gene is used to predict biotic indices. Currently, the ASC reports provide values of six biotic indices (AMBI, NSI, NQI 1, ISI, Shannon-Wiener and nEQR). The values are predicted for each of two grabs collected at each station and the two values are averaged into a final value for a station (Table 6).

Table 6. Values of six biotic indices predicted based on 18S V1V2 metabarcoding data. The colours correspond to ecological status (blue = very good, green = good, yellow = moderate, orange = bad, red = very bad). The ± sign indicates standard deviation. According to the ASC indicator 2.1.2, the ecological quality in sediment outside the AZE should be good to very good, which corresponds to AMBI score ≤ 3.3, or Shannon score > 3.

	ASC1-AZE	ASC2-AZE	ASC3-outside AZE	ASC4-outside AZE	ASC Reference
AMBI	5.11±0.2	4.99±0.2	2.54±0.35	2.41±0.07	2.23±0.1
NSI	11.6±0.52	11.24±0.82	21.55±1.7	22.41±0.28	23.23±0.08
ISI	6.87±0.69	7.1±0.17	9.13±0.42	9.44±0.29	9.9±0.13
NQI1	0.38±0.02	0.39±0.03	0.73±0.04	0.73±0.01	0.76±0
Shannon	1.76±0.11	1.94±0.29	3.68±0.4	3.85±0.15	4.22±0.09
nEQR	0.338±0.11	0.356±0.12	0.696±0.04	0.712±0.04	0.746±0.04

In addition to the score of biotic indices, the ASC reports also comprise species lists inferred from metabarcoding data for both 18S and COI markers. Only, macrofauna species with defined ecological category in AMBI, NSI or ITI (as in the reference papers of Rygg and Norling, 2013 and Word, 1980) are currently included in the list. The taxa are listed with their relative frequencies within the 18S and COI eDNA datasets of each station (Table 7).

The species lists for both markers are not the same due to the specificity of each marker. Both markers are complementary. The 18S marker is known to target particularly meiofaunal groups, such as nematodes, while the COI marker covers mainly macrofauna, such annelids, echinoderms, and molluscs. The taxonomic resolution of COI marker is higher than the 18S, which means that the COI sequences can be easier identified to species level. However, the 18S marker has larger taxonomic range, which means more species can be detected in eDNA samples.

Table 7. Example of species list for COI marker. The pollution indicator species classified in ecological category 5 are marked in red.

Taxon	Ecol.group			Relative frequency
	AMBI	ITI	NSI	18S marker
Ciona intestinalis	3	NA	NA	61,8
Sabatieria pulchra	5	NA	NA	31,0
Ascidiella sp.	3	1	NA	4,6
Malacoceros fuliginosus	5	4	5	1,3
Capitellida undet.	5	NA	3	1,3
Pectinaria koreni	4	NA	4	<0,5
Priapulus caudatus	3	3	3	<0,5
Corymorpha nutans	1	NA	NA	<0,5
Phyllodocida undet.	NA	3	2	<0,5
Capitella sp.	5	4	3	<0,5
Actiniaria undet.	2	3	1	<0,5

Taxon	Ecol.group			Relative frequency
	AMBI	ITI	NSI	COI marker
Priapulus caudatus	3	3	3	24,1
Malacoceros fuliginosus	5	4	5	23,2
Evadne nordmanni	NA	1	NA	18,3
Cribrilina mutabilis	1	NA	NA	13,8
Pectinaria koreni	4	NA	4	3,8
Caprella mutica	2	NA	3	3,1
Mytilus sp.	3	NA	NA	3,1
Akera bullata	1	NA	NA	2,2
Tubificoides benedii	5	4	NA	2,2
Ophiocomina nigra	1	1	NA	1,2
Phyllodoce groenlandica	4	NA	3	0,6

According to ASC indicator 2.1.3, at least 2 highly abundant taxa that are not pollution indicator species should be present in the sediment within the AZE. When using morphological analyses, these taxa should be represented by at least 100 organisms per square meter (or equally high to reference site(s) if natural abundance is lower than this level). When using molecular technology, compliance is verified by demonstrating the presence of DNA sequences assigned to non-pollution indicators among the abundant species. The analysis of eDNA data does not allow to infer the number of specimens present in the grab samples, as provided by classical macrofauna benthic monitoring. Nevertheless, it has been shown by several studies that the eDNA datasets have semi-quantitative character and that the relative frequency of sequences corresponds relatively well to the abundance of the taxon. At present, species represented by >1% of relative abundance of macroinvertebrates are considered as abundant. In the future, this indicator could be replaced by the proportion of DNA sequences belonging to opportunistic vs non-opportunistic species, which would be much easier to infer from metabarcoding data.

6. Conclusions and further developments

The project led to following conclusions	**The eDNA metabarcoding approach** developed in this project **is sufficiently mature to be implemented** for all soft-sediment sites situated in the areas, where reference datasets are available.
	Meiofaunal taxa, in particular nematodes, have potential to replace the macro-invertebrates as alternative bioindicators of organic enrichment in metabarcoding data.
	Biotic indices predicted from metabarcoding data **using machine learning method show similar ecological status** as the indices inferred from macro-invertebrates surveys.

Further developments

Based on the conclusions of this study, we recommend:

1. **To increase the number of sites** used for training data **to account for the impact of spatiotemporal dynamics** on the accuracy of biotic indices predicted using machine learning approach.
2. **To establish a new nematodes-based index of organic enrichment**, which could replace the current macrofauna-based index and complement machine learning predictions with a taxonomic framework adapted to eDNA datasets, thus maximizing the use of metabarcoding data.
3. **To conduct an interlaboratory validation** on a random set of samples **to ensure the reproducibility of the results** obtained by different practitioners, as recommended by the Scientific Panel.

Moreover, we recommend developing an **eDNA test for sand-shelly and hard bottom** stations. While not solving the abundance issue, such test based on newly published method based on isolating eDNA from ethanol preservative could increase the likelihood of detecting pollution or non-pollution indicators in macrofauna samples.

7. Deliverables

Papers published in peer-reviewed scientific journals

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. Environ Sci Technol. 51(16):9118-9126.

Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Mol Ecol Resour. 2018 Jul 17. doi: 10.1111/1755-0998.12926.

Cordier T, Lanzén A, Apothéloz-Perret-Gentil L, Stoeck T, Pawlowski J. Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. Trends Microbiol. 2019 May;27(5):387-397.

Cordier T, Pawlowski J (2018) BBI: an R package for the computation of Benthic Biotic Indices from composition data. Metabarcoding and Metagenomics 2, e25649

Frühe L, Cordier T, Dully V, Breiner H-W, 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 (submitted to Mol Ecol)

Pawlowski J, Esling P, Lejzerowicz F, Cordier T, Visco JA, Martins CIM, Kvalvik A, Staven K, Cedhagen T (2016) Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. Aquacult Environ Interact. 8:371-386

Stoeck T, Frühe L, Forster D, Cordier T, Martins CIM, Pawlowski J. (2018) Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. Marine Pollution Bulletin 127:139-149

Papers in preparation

Cordier et al. Comparing SML algorithms for benthic monitoring

Cordier et al. Extracting novel metrics from metabarcoding data for biomonitoring

Pawlowski et al. Expanding the range of bioindicator taxa through assigning indicator values to selected nematode species

Pawlowski et al. Benthic foraminifera as indicators of organic enrichment associated with marine aquaculture activities: molecular perspective

Stoeck et al. Impact of biogeographic patterns on bacterial and ciliate eDNA indicators for salmon farm monitoring

Presentations at international conferences (talks, posters)

- EAS (European aquaculture society) meeting, Rotterdam, Netherlands, October 2015
  - organization of special session on environmental monitoring
  - oral presentation: *"Environmental DNA surveys for benthic monitoring of salmon farms"*
- ECOP (European Congress of Protistology), Seville, Spain, September 2015 – keynote lecture *"Protist Metabarcoding and its Applications"*
- Workshop on application of genomic tools for biomonitoring of marine environment: from technology to legal and socio-economic aspects, Geneva, Switzerland, May 2016 – Philippe Esling oral presentation *"Equations of species, communities of numbers: towards next generation biomonitoring"*
- EAS meeting, Edinburgh, UK, September 2016 – poster *"Benthic monitoring of salmon farms in Norway using environmental DNA metabarcoding"*
- Workshop Live Foraminifera as a new model system for monitoring and reconstructing marine environments, Eilat, Israel, September 2016 – oral presentation: *"Foraminiferal DNA metabarcoding applied to benthic monitoring: promises and challenges"*
- Environmental Genomics Workshop, St. Johns', Canada, June 2017 – talk *"Predicting benthic ecological status from eDNA metabarcoding data using supervised machine learning"*
- International Barcode of Life Conference, Kruger, South Africa – November 2017 – talk *"Ecobarcoding: taxonomy-free approach for high-throughput environmental DNA-based biomonitoring"*
- Seminar Stazione Zoologica Anton Dohrn, Naples, Italy, October 2018 – invited talk *"Next-generation biomonitoring or how DNA metabarcording can help assessing human impacts on marine environments"*
- Workshop Mainstreaming molecular approaches in national environmental monitoring programs – Porto, Portugal, December 2018 – talk: *"Biotic indices. How to change from conventional to molecular approaches?"*
- International Workshop on Foraminiferal Biomonitoring, Sao Paolo, Brazil, May 2019 – talk *"Foraminiferal DNA metabarcoding – a new promising tool for benthic monitoring"*
- International Barcode of Life Conference, Trondheim, Norway – June 2019 – oral presentation *"Expanding the range of bioindicator taxa through assigning indicator values to selected metabarcodes*"
- DNAqua-net meeting, Limassol, Cyprus, September 2019 – oral presentation "*Using DNA metabarcoding for environmental impact assessment in marine industry*"
- EAS meeting, Berlin, Germany, October 2019 – oral presentation "*Monitor the environmental impacts through artificial intelligence applied to metagenomic data*"

Guidelines for using eDNA metabarcoding to assess the impact of salmon farms on benthic communities (Annex)

Report of Scientific Panel (Annex)

8. 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. Mol Ecol Resour. 17: 1231-1242.

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. Environ Sci Technol. 51(16):9118-9126.

Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J. (2018) Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Mol Ecol Resour.18(6):1381-1391

Cordier T, Pawlowski J (2018) BBI: an R package for the computation of Benthic Biotic Indices from composition data. Metabarcoding and Metagenomics 2, e25649

Cordier T, Lanzén A, Apothéloz-Perret-Gentil L, Stoeck T, Pawlowski J. (2019) Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. Trends Microbiol. 27(5):387-397

Dowle E, Pochon X, Keeley N, Wood SA (2015) Assessing the effects of salmon farming seabed enrichment using bacterial community diversity and high-throughput sequencing. FEMS Microbiology Ecology, 91:fiv089

Forster D, Filker S, Kochems R, Breiner H-W, Cordier T, Pawlowski J, Stoeck T. (2019) A comparison of different ciliate DNA metabarcode genes as bioindicators for environmental impact assessments of salmon aquaculture. J Eukaryot Microbiol. 66(2):294-308.

He X, Sutherland TF, Pawlowski J, Abbott CL. (2019) Responses of foraminifera communities to aquaculture-derived organic enrichment as revealed by environmental DNA metabarcoding. Mol Ecol. 28(5):1138-1153.

Keeley N, Wood SA, Pochon X (2018) Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. Ecol. Indicators 85:1044-1057

Leese F, Bouchez A, Abarenkov K, Altermatt F, Borja A, Bruce K, Ekrem T, Čiampor F, Čiamporová-Zaťovičová Z, Costa FO, Duarte S, Elbrecht V, Fontaneto D, Franc A, Geiger MF, Hering D, Kahlert M, Kalamujić Stroil B, Kelly M, Keskin E, Liska I, Mergen P, Meissner K, Pawlowski J, Penev L, Reyjol Y, Rotter A, Steinke D, van der Wal B, Vitecek S, Zimmermann J, Weigand AM (2018) Why We Need Sustainable Networks Bridging Countries, Disciplines, Cultures and Generations for Aquatic Biomonitoring 2.0: A Perspective Derived From the DNAqua-Net COST Action. In: Bohan et al. Next-Generation Biomonitoring. Advances in Ecological Research 58:63-99

Lejzerowicz F, Esling P, Pillet L, Wilding TA, Black KD, Pawlowski J. (2015) High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. Sci Rep. 5:13932.

Pawlowski J, Esling P, Lejzerowicz F, Cedhagen T, Wilding TA. (2014) Environmental monitoring through protist NGS metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. Molecular Ecology Resources 14(6):1129-1140.

Pawlowski J, Esling P, Lejzerowicz F, Cordier T, Visco JA, Martins CIM, Kvalvik A, Staven K, Cedhagen T (2016) Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. Aquacult Environ Interact. 8:371-386

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

Pawlowski J, Lejzerowicz F, Apotheloz-Perret-Gentil L, Visco J, Esling P (2016) Protist Metabarcoding and Environmental Biomonitoring: Time for Change. Eur J Protistol. S0932-4739(16)30003-7.

Pochon X, Wood SA, Keeley NB, Lejzerowicz F, Esling P, Drew J, Pawlowski J. (2015) Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. Mar Pollut Bull. 100(1):370-82.

Rygg, B, Norling, K. (2013). REPORT SNO 6475-2013 Norwegian Sensitivity Index (NSI) for marine macroinvertebrates, and an update of Indicator Species Index (ISI).

Stoeck T, Frühe L, Forster D, Cordier T, Martins CIM, Pawlowski J. (2018) Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. Marine Pollution Bulletin 127:139-149

Stoeck T, Kochems R, Forster D, Lejzerowicz F, Pawlowski J. (2018) Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. Ecol. Indicators 85:153-164

Valentini A, Taberlet P, Miaud C, et al. (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Mol Ecol. 25(4):929-942.

Veileder 02:2018: Klassifisering av miljøtilstand i vann, www.vannportalen.no

Word, JQ (1980) Classification of benthic invertebrates into Infaunal Trophic Index feeding groups. Southern California Coastal Water Research Project Biennial Report 1979–1980. pp. 103–121

9. Supplementary tables and figures

Table S1. List of sampled sites, with sampling date, number of stations and samples and coordinates.

N°	Site	Date	Samples	Stations	GPS_lat	GPS_long
FHF 1	Bjørnsvik	03.06.15	24	4	67.5325	15.38731667
FHF 2	Nedre Kvarv	04.06.15	30	5	67.46036667	15.50891667
FHF 3	Beitveit	12.10.15	30	5	62.1398	5.32765
FHF 4	Storvika	14.10.15	30	5	62.80316667	6.9794
FHF 5	Aukrasanden	15.10.15	30	5	62.78171667	6.923916667
FHF 6	Rundreimstranda	15.03.16	30	5	62.01298333	5.3302
FHF 7	Flåtegrunnen	16.03.16	30	5	61.57618333	4.8085
FHF 8	Kvalvika	17.08.16	36	6	66.66843333	13.39335
FHF 9	Brattholmen	07.08.16	36	6	65.90968333	12.22151667
FHF 10	Bukkholmen	01.07.16	36	6	65.9199	12.24838333
FHF 11	Digermulen	27.10.15	30	5	66.67661667	13.31881667
FHF 12	Napp	10.01.17	24	5	58.246767	6.530067
FHF 13	Skipningsdalen	11.01.17	15	4	58.240283	6.607
FHF 14	Støytland	11.01.17	20	5	58.22928333	6.6302
FHF 15	Salvågvika	12.01.17	24	6	58.22236667	6.63865
FHF 16	Hjellberget	25.04.17	24	6	69.91165	21.59515
FHF 17	Karvika	26.04.17	20	5	69.8552	21.8282
FHF 18	Tendalsvik	24.05.18	24	4	59.73605	5.884
FHF 19	Utåker	23.05.18	24	4	59.778333	5.9024
FHF 20	Mefaldskjæret	22.05.18	36	6	65.85145	12.2872
FHF 21	Olderøy	26.06.18	36	6	61.103283	4.712233
FHF 22	Rakkenes	19.07.18	30	5	69.885067	21.729983
FHF 23	Alsåkervik	29.01.19	18	3	59.7601	6.074183
FHF 24	Slåttenes	29.01.19	18	3	59.832117	5.98365
FHF 25	Tveit	04.03.19	30	5	61.110083	5.340483
FHF 26	Bekksneset	05.03.19	36	6	61.076617	5.38595
FHF 27	Røysa	07.03.19	28	7	62.54865	6.1417

Table S2. Selected bioindicator species of Gastrotriches assigned to AMBI and NSI ecological groups using splines analysis

SPECIES (Gastrotricha)	**OTU**	**AMBI group**	**AMBI value**	**NSI group**	**NSI value**
Aspidiophorus tentaculatus	OTU259	4	4.98	4	10.74
Aspidiophorus tentaculatus	OTU912	NA	NA	3	18.92
Chaetonotida	OTU1435	NA	NA	3	16.29
Chaetonotida	OTU252	5	6.00	5	4.68
Chaetonotida	OTU292	2	3.02	3	19.02
Chaetonotida	OTU523	3	3.43	3	18.24
Chaetonotida	OTU866	2	2.08	3	19.32
Chaetonotus cf. dispar TK146	OTU1892	NA	NA	3	19.48
Chaetonotus cf. dispar TK146	OTU4459	4	5.14	NA	NA
Chaetonotus neptuni	OTU152	2	2.76	3	22.00
Chaetonotus schultzei	OTU360	2	2.25	NA	NA
Halichaetonotus euromarinus	OTU1382	2	2.91	3	17.95
Heterolepidoderma loricatum	OTU1127	NA	NA	2	22.52
Heterolepidoderma loricatum	OTU1128	2	2.26	2	22.52
Heterolepidoderma loricatum	OTU149	2	3.07	2	22.52
Heterolepidoderma loricatum	OTU1648	2	2.25	2	22.70
Heterolepidoderma loricatum	OTU305	2	2.82	3	22.31
Heterolepidoderma loricatum	OTU45	2	3.07	3	19.41
Macrodasys sp. 2 KG-2011	OTU665	4	5.00	4	11.58
Polymerurus nodicaudus	OTU343	2	2.91	3	19.02
Polymerurus rhomboides	OTU645	NA	NA	3	19.03
Thaumastoderma ramuliferum	OTU1621	2	2.91	3	19.03
Thaumastoderma ramuliferum	OTU3140	2	2.86	NA	NA

Table S3. Selected bioindicator species of Platyhelminthes assigned to AMBI and NSI ecological groups using splines analysis

SPECIES (Platyhelminthes)	OTU	AMBI group	AMBI value	NSI group	NSI value
Bresslauilla relicta	OTU567	5	5.85	5	7.56
Byrsophlebs delamarei	OTU1842	NA	NA	1	25.41
Coronhelmis lutheri	OTU1274	3	3.29	3	18.24
Ethmorhynchus anophthalmus	OTU1059	NA	NA	1	24.63
Eubothrium crassum	OTU2077	4	5.23	5	6.87
Kalyptorhynchia	OTU14	4	4.79	4	11.58
Kalyptorhynchia	OTU3287	4	5.00	4	11.06
Kalyptorhynchia	OTU3409	5	5.81	5	10.74
Maehrenthalia agilis	OTU890	5	6.00	5	6.87
Mesorhynchus terminostylis	OTU303	3	3.35	3	18.57
Microstomum sp. C TJ-2015	OTU1146	1	1.14	NA	NA
Microstomum sp. C TJ-2015	OTU92	4	5.45	4	11.97
Odontorhynchus aculeatus	OTU1260	3	4.17	NA	NA
Odontorhynchus aculeatus	OTU2204	NA	NA	4	14.29
Odontorhynchus aculeatus	OTU2620	3	4.35	NA	NA
Odontorhynchus aculeatus	OTU39	3	4.14	3	15.40
Odontorhynchus aculeatus	OTU4199	NA	NA	3	16.29
Placorhynchus octaculeatus	OTU2027	5	6.00	NA	NA
Placorhynchus octaculeatus	OTU2726	5	6.00	5	6.99
Placorhynchus octaculeatus	OTU3846	4	5.00	4	11.06
Placorhynchus octaculeatus	OTU5365	4	4.98	NA	NA
Proseriata	OTU3275	1	0.43	1	26.69
Provortex karlingi	OTU6403	5	6.00	5	5.26
Provortex karlingi	OTU91	4	6.00	4	4.68
Psammomacrostomum sp. 4 TJ-2015	OTU925	NA	NA	1	24.41
Scanorhynchus forcipatus	OTU165	4	4.35	4	15.48
Styloplanella strongylostomoides	OTU2529	1	0.92	NA	NA
Styloplanella strongylostomoides	OTU900	2	1.99	NA	NA

Table S4. Results of biotic index value predictions with the three different approaches, and two different supervised learning algorithms. Best predictive models are in bold. Significance is as follow: *:P < 0.05; **:P < 0.01; ***:P < 0.001

Biotic Index	Data used for predictions	Supervised learning algorithm	R^2	Kappa
AMBI	Correlation screening	-	0.568***	0.624***
	Diversity metrics	Random Forest	0.641***	0.69***
		Self-Organizing Map	0.492***	0.618***
	Composition data	Random Forest	0.662***	0.555***
		Self-Organizing Map	**0.669*** **	**0.711*** **
ISI	**Correlation screening**	-	0.65***	0.53***
	Diversity metrics	Random Forest	0.505***	0.626***
		Self-Organizing Map	0.449***	0.61***
	Composition data	Random Forest	0.56***	0.631***
		Self-Organizing Map	**0.615*** **	**0.774*** **
NSI	Correlation screening	-	0.508***	0.607***
	Diversity metrics	**Random Forest**	**0.83*** **	**0.907*** **
		Self-Organizing Map	0.83***	0.88***
	Composition data	Random Forest	0.827***	0.832***
		Self-Organizing Map	0.794***	0.871***
NQI1	Correlation screening	-	0.76***	0.8***
	Diversity metrics	**Random Forest**	**0.834*** **	**0.88*** **
		Self-Organizing Map	0.805***	0.846***
	Composition data	Random Forest	0.81***	0.856***
		Self-Organizing Map	**0.803*** **	**0.873*** **

Table S5. Distribution of the five ecological inferences with morphology and eDNA for NSI, Shannon, AMBI, NQI1 and ISI indices, according to previous and new class limits.

Class limits system	Methodology	Index	Very bad	Bad	Moderate	Good	Very good
Previous class limits	Morphology	NSI	33	14	21	52	12
	eDNA	NSI	9	36	32	55	0
	Morphology	Shannon	28	13	18	63	10
	eDNA	Shannon	2	33	28	69	0
	Morphology	AMBI	32	16	15	66	3
	eDNA	AMBI	4	39	25	64	0
	Morphology	NQI1	23	20	22	61	6
	eDNA	NQI1	5	38	28	61	0
	Morphology	ISI	8	22	26	51	25
	eDNA	ISI	1	11	46	66	8
	Morphology	total	124	85	102	293	56
	eDNA	total	21	157	159	315	8
New class limits	Morphology	NSI	33	13	20	52	14
	eDNA	NSI	9	35	29	50	9
	Morphology	Shannon	28	13	19	26	46
	eDNA	Shannon	3	30	29	26	44
	Morphology	AMBI	32	16	15	66	3
	eDNA	AMBI	4	39	25	64	0
	Morphology	NQI1	23	20	24	23	42
	eDNA	NQI1	5	38	30	28	31
	Morphology	ISI	9	25	27	22	49
	eDNA	ISI	1	22	39	27	43
	Morphology	total	125	87	105	189	154
	eDNA	total	22	164	152	195	127

Figure S1. Linear models between eukaryotic diversity metrics (OTUs richness, Shannon diversity and Chao index), with biotic indices variations (AMBI, NSI and Shannon) obtained from macrofauna survey. R^2 and their significance is indicated on the plots. The blue line indicates the fitted linear model and the dashed green lines indicate the 95% confidence intervals.

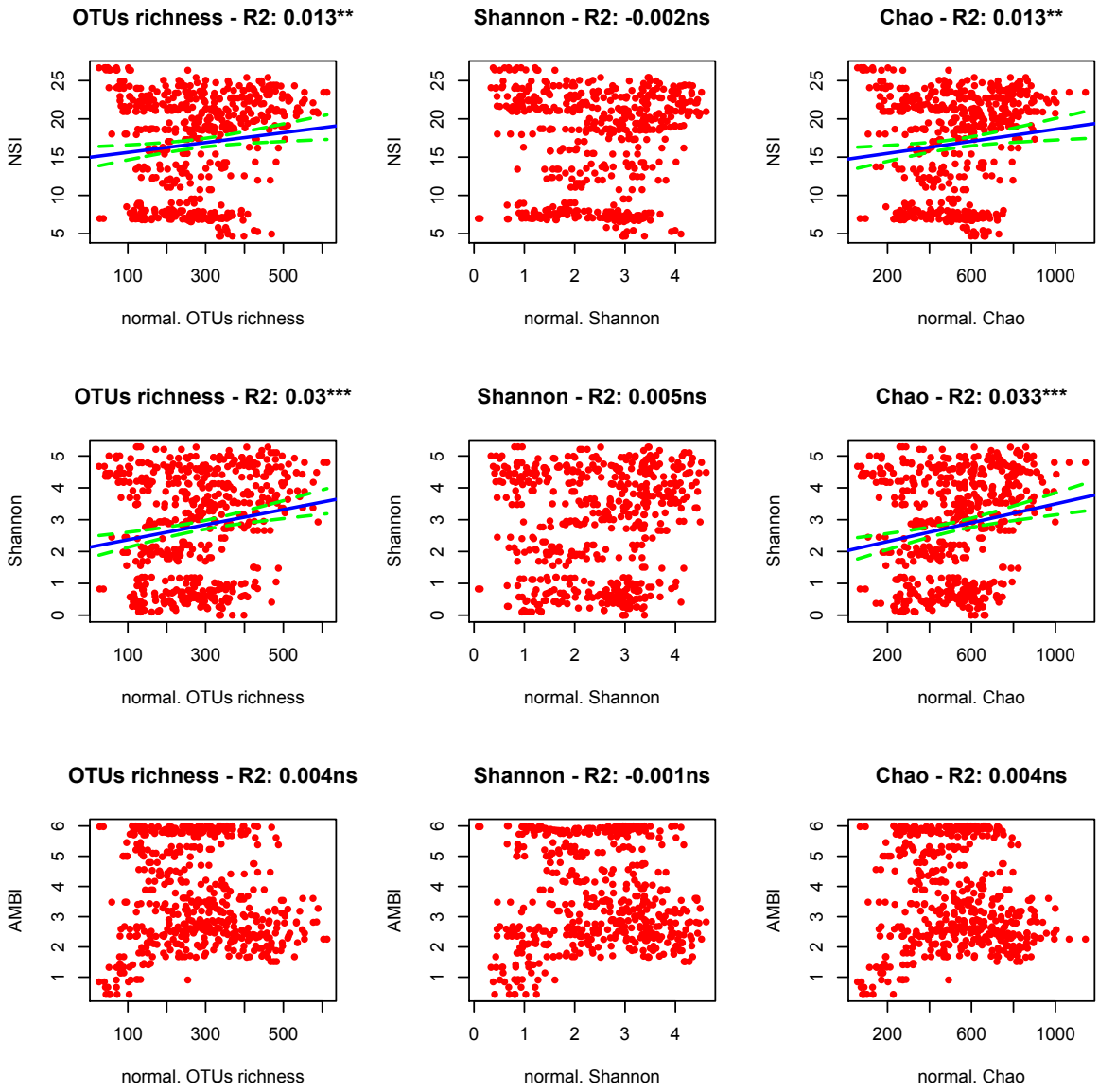


Figure S2. SML predictions of AMBI performance comparison between foraminifera and full eukaryotes dataset. Both R^2 and kappa statistics and significances are indicated as inset in each plot. The eukaryotic predictions were based on the same selection of sites for which the foraminifera dataset was available (n=17). The histogram on the top right of each plot indicates the congruence between the reference discrete status value and the predicted one (number indicates the mismatches, zero being perfect classification). The bottom right boxplots indicate the distribution of errors (the difference between the reference AMBI values and the predicted ones) with medians and quantiles in black and average in red.

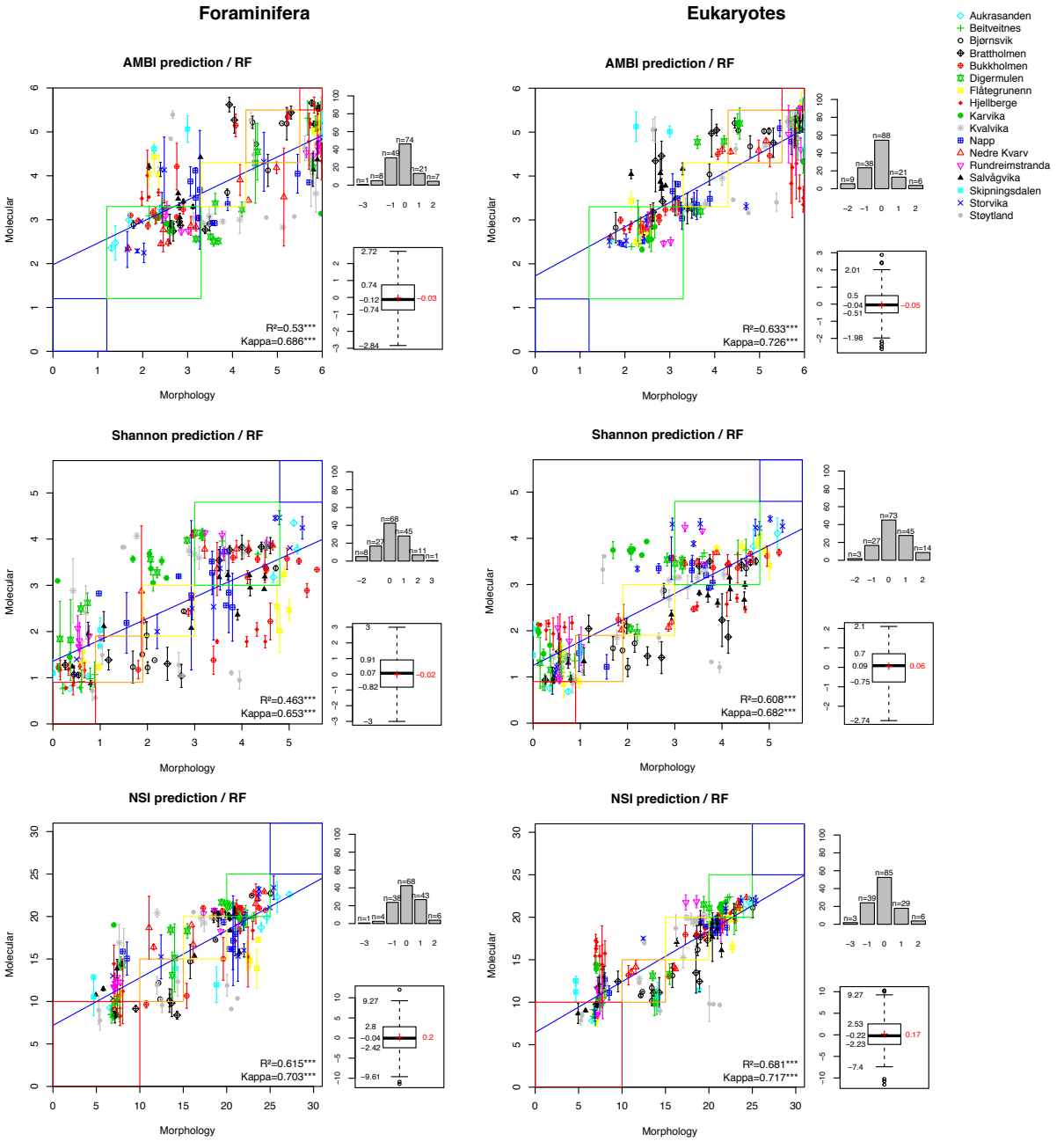


Figure S3. Predictions of four biotic indices (ISI, Shannon, AMBI and NQI1) using the full eukaryotic dataset (27 farms) and supervised machine learning.