COMPLETED PROJECT CASE STUDY

DEVELOPING DNA-BASED METHODOLOGIES FOR SEABED MONITORING

PARTNERS

Scottish Association for Marine Science (SAMS) | Rivers and Lochs Institute - UHI Inverness College | Mowi Scotland | Scottish Environment Protection Agency (SEPA) | University of Kaiserslautern | Scottish Sea Farms | Salmon Scotland

PROJECT LEADS

Dr Tom Wilding, Dr Mark Coulson, Stephen MacIntyre, Willie Duncan, Connor McKnight, Thorsten Stoeck, Richard Fyfe, Malcom Baptie, Ryan Eustace.

BACKGROUND

Regulatory framework is being implemented by the Scottish Environment Protection Agency (SEPA) that will require salmon farmers to increase the level of monitoring needed to effectively demonstrate that environmental standards are being met underneath and around fish net-pens as part of their consent conditions. Traditionally, compliance-monitoring around salmon farming has been based on macrobenthic analysis, i.e., counts of organisms present in benthic grab samples, which are then used to generate diversity indices including the 'infaunal quality index' (IQI). However, this approach, which relies on manual sorting and identification of specimens, is time consuming and expensive, costing the Scottish sector approximately £1m per year. In order to meet the increased monitoring requirements set out by SEPA at a reasonable cost and timeframe, the sector needs a rapid method of assessing benthic status.

The overarching aim of this project was to provide a revolutionary approach to this process by extracting DNA from the sediment and using a next-generation sequencer (NGS) to identify the organisms present. Next-generation sequencers can sequence marker-DNA regions and, by comparison to sequence-databases, species can be identified. This process, when applied to a mixture of species and multiple samples, is called metabarcoding, and enables thousands of taxa to be identified simultaneously in about 400 samples. Species identification via metabarcoding offers an alternative to morphology-based identification and is potentially more cost-effective, less subjective, and much faster.

This six-year project was broken down into two phases: MeioMetBar (MMB) and MeioMetBar 1.5 (MMB1.5), both with the ultimate goal of implementing the NGS approach to benthic monitoring in Scotland. The first phase of this project (MMB), driven by leading experts at the Scottish Association of Marine Science (SAMS), Mowi Scotland, the University of Highlands and Islands (UHI), and Scottish Environment Protection Agency (SEPA), was delivered across two work packages. These set out to:

1. optimise NGS methodology in terms of sample collection protocols for the generation of consistent molecular (genomic) indices of benthic status;
2. identify key indicator species and link the genomic indices and the microbenthic indices to generate comparable management decisions.

Running in parallel with MMB was a project run by Thorsten Stoeck at the University of Kaiserslautern, Germany. Thorsten's project, DNAmon, had similar objectives, methodologies, and data outputs to MMB, offering an opportunity for collaboration and the collation of two complementary datasets for further exploration. This led to the development of phase 2 (MMB1.5), bringing in additional project collaborators including the University of Kaiserslautern, Scottish Sea Farms and Salmon Scotland, to gain further knowledge by combining the two datasets (MMB and DNAmon). Phase 2 of this project used both data sets, which characterised the bacterial composition of sediment samples taken from around fish farms, to train a machine learning (ML) algorithm to predict the IQI of samples based on their bacterial characterisation. Furthermore, the accuracy of the algorithm was evaluated by comparing IQI predictions from independent samples. Overall, this work set out to optimise the protocols and tools necessary to implement this technology across the sector.

It should be noted that this case study represents a very broad overview of an extensive project that was delivered in multiple phases. For a more comprehensive review of the terms, technology, and methodology used throughout, we encourage you to consult the published works detailed below.

AIMS

The overall goal of this work was to develop and optimise protocols for sampling, sample preparation, NGS and data analysis, and to deliver a 'scoping tool', or intellectual process, for the cost-effective, reproducible identification and classification of benthic indices to ensure that salmon farms continue to meet the regulatory consenting conditions set out by SEPA.

> *The MeioMetBar project has developed the capacity, in Scotland, to apply next-generation sequencer (NGS) technology to aquaculture compliance assessment, representing a major step forward. This toolkit has the potential to be used across the sector to upload and test new data sets, and moreover, it has been developed as open-source, so its use is not restricted.*

MEIOMETBAR

Phase one of this project, MeioMetBar, was initiated with the following key objectives:

1. To optimise sample collection for NGS analysis and the data processing 'pipeline' to generate consistent molecular indices of benthic status.
2. To identify indicator molecular species and cross references with traditional indices to deliver a reliable alternative method of benthic quality assessment.

The infaunal quality index (IQI) is used to assess the ecological status of the macrobenthic invertebrate infaunal groups (i.e. organisms that live in the sediment) within sediment samples and to monitor habitats in UK coastal and transitional water bodies. For SEPA, is their main tool for seabed compliance monitoring at commercial aquaculture net-pen sites.

Within the term 'metabarcoding', 'meta' relates to the sequencing of numerous (millions) of DNA fragments simultaneously, and 'barcoding' relates to the identification of those sequences via comparison to databases. This project set out to develop a machine learning model capable of predicting IQI based on identifying species – via their DNA – from benthic sediment samples and, moreover, to operate under the assumption that if the model-predicted IQI is high enough for compliance (>0.64) then no further investigation is needed, but if the prediction falls below the given threshold, then a direct assessment can be carried out before declaring non-compliance.

Sampling methodology and bioinformatic pipeline optimisation

All organisms contain DNA, which consists of a long sequence of various permutations of four types of bases. Certain parts of the DNA sequence contain base-permutations (markers) that are particular to a given specie. In this context, species can be identified through the use of NGS to sequence marker-DNA regions for comparison to sequence databases. For this project, three makers – bacteria (16S), general eukaryotes (18S) and the cytochrome oxidase subunit I (COI) marker – were chosen to cover the species groups mostly likely to show meaningful change over farm-related environmental gradients and assessed on their accuracy for species prediction.

Grab samples for benthic analysis are typically 0.02–0.1m^2 and contain up to 20kg of sediment compared to the ~10g of sediment required for analysis via metabarcoding. The first step was to optimise sample collection and preparation methodology that limits 'within grab' variability, ahead of NGS and analysis to ensure reliable and repeatable results.

An assessment of sample standardisation was carried out on the ~10 g sediment samples by taking series of replicates per grab, at a range of distances from three net-pen sites, and then comparing similarities 'within grab' to 'between grab' as well as establishing the degree of variability in results that could be attributed directly to the NGS process itself. This analysis was conducted using the three different markers selected, and following the NGS protocol described by Lejzerowicz et al. (2015), which roughly included DNA extraction, amplification, library preparation, sequencing, identification, and subsequent bioinformatics and statistical analysis.

Results from this analysis demonstrated that the 16S bacterial marker showed the highest degree of precision and clearly discriminated between cage-edge and samples taken at set distances from the farm and reference stations. In terms of precision, and based on the available databases, the 18S marker showed the least precision (greatest variability between replicates) with the COI marker showing intermediate precision.

Linking molecular and macrobenthic indices to generate comparable management decisions

The second part of this work was based around the identification of indicator molecular species and then linking these indicator species to the indices derived through traditional methodology, in order to deliver a reliable and comparable alternative method of benthic quality assessment that allows for classification of seabed quality against set standards. The identification of indicator molecular species is achieved through taxonomic annotation of sequences, which is only possible where corresponding sequences are available on comparative databases. Where annotation is achieved, statistical algorithms can be used to identify those taxa that most clearly link metabarcode-identified taxa with environmental gradients. In this case, the environmental status of samples was based on the IQI as determined using traditional morphology-based methods.

Using the standardised sampling approach developed earlier in the project, sub-samples were obtained from sediment grabs that were taken as part of ongoing standard SEPA compliance monitoring routinely conducted by Mowi, SEPA, and SAMS from several independent locations.

Prior to NGS, each sample was processed and measured for IQI (derived from traditional macrobenthic analyses and generated through UK Environmental Agency standard), distance to pen-edge, particle size and organic carbon content. These samples were then metabarcoded and a supervised machine learning technique or algorithm for pattern matching, called ‘Random Forests’ (RF) was used to predict the IQI based on the metabarcodes.

Next, taxa, identified via their metabarcode, were highlighted based on their accuracy in matching predicted vs actual IQI. Through this, it was found that predicting IQI based on bacterial-class level identification was optimal in terms of consistency across a number of data transformations. The key bacterial classes identified through this process included Acidobacteria Subgroup22, Cytophagia, Nitrospira and Phycisphaerae and Bacteriodes but did not include typical indicator taxa such as sulphate reducing bacteria or Beggiatoa.

Based on this model and level of data included, it was determined that the 16S bacterial marker appears to offer the most consistency in terms of predicting IQI and that one could be reasonably certain (95% prediction interval) that an unknown sample with a predicted IQI of 0.6 would have an actual IQI of between 0.4 and 0.8.

Project partners concluded that further research and data collation was needed to improve the precision and accuracy of the 16S bacterial marker in order to narrow the IQI prediction interval. Moreover, that a more detailed understanding was required of how key-bacterial (and other) taxa are related to the IQI. Knowing more about how that varies between locations and over time would be essential in the development of metabarcoding technology as a replacement for macrobenthic-based environmental assessments. These conclusions, along with the prospect of collaboration and combined learnings from the paralleled DNAmon project, led to the establishment of a second phase of this work: MeioMetBar 1.5.

MeioMetBar 1.5

Following on from the above and based on using data from DNAmon, MMB1.5 set out three key objectives:

1. To collate two complimentary datasets that characterise the bacterial composition and associated metadata of sediment samples taken around fish farms.
2. To use the combined data set to train a ‘random forest’ machine learning algorithm to predict the IQI of samples based on their bacterial characterisation.
3. To use the trained machine to predict IQI from independent samples provided by industry and evaluate the accuracy of the algorithm.

Before the two datasets (DNAmon and MMB) could be compared and combined, a series of steps were taken to convert the raw sequence data to a common format.

Using the combined datasets, RFs was used to make predictions of data that was produced independently at three sites, more specifically, to find associations in predictor data (bacterial family abundances) and the IQI. ML algorithms seek to generalise from the training data to new samples that it has not ‘seen’ before, and more training data improves this generalisation.

The MMB and DNAmon datasets were joined to a single database for RF model development. The RF model developed accurately predicted the IQI in a ‘leave-one-out’ cross validation (Figure 1). The predictions were precise but conservative, predominantly underestimating the actual IQI. In one pen-edge sample, the IQI was overestimated by the machine compared to the measured macrobenthic IQI.

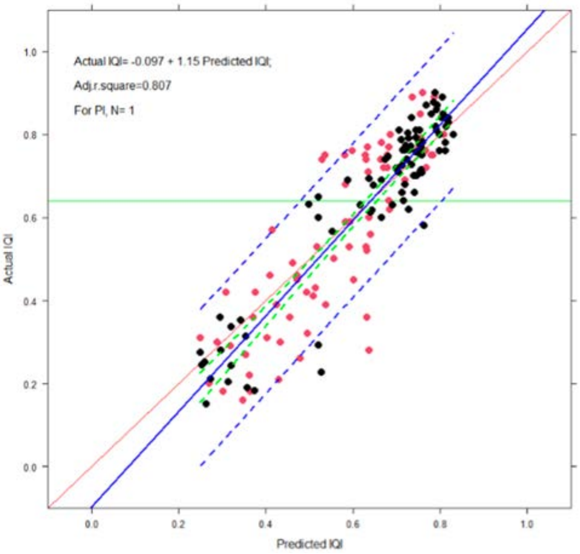

Figure 1 – Results from ‘leave-one-out’ cross-validation. The RF predicted IQI (X axis) and corresponding actual IQI (Y axis) are plotted. The linear regression model output is shown (blue line, with model detail included in the upper left). 95% confidence intervals (green dashes) and 95% single-sample prediction intervals (blue dashes) are also shown. Black and red dots indicate MeioMetBar and DNAmon sources respectively. The ideal 1:1 relationship is indicated by the red diagonal line. The 0.64 IQI regulatory threshold is shown in green on Y axis.

This data provides compelling evidence that NGS-bacterial characterisation has considerable potential as a replacement for traditional macrobenthic analysis. Furthermore, the main output from MMB1.5 consists of the two sets of code necessary to bring the data sets together: 1) The BASH code that enables the reproducible and stable denoising, annotation and analysing of raw sequence data to create files formatted for input for the R code; and 2) The R code, which enables the uploading of test data (following the protocols in the BASH code) and the generation of RF-based predictions, together with accompanying graphical and tabular outputs. This code has been made available to SEPA and partners and is currently being implemented.

IMPACT

The regulation requirements regarding the maximum allowable biomass at any given salmon net-pen site represents a major limiting factor in the growth of the Scottish salmon-farming sector and is reducing its international competitiveness. Current changes in regulations have resulted in a substantial increase in the level of benthic monitoring required to demonstrate compliance with environmental standards underneath and around salmon net-pens. The overarching goal of this project was to demonstrate the use of modern DNA sequencing techniques to identify the bacterial composition in sediments around fish farms with a view to providing a quicker, more cost-effective measure of the 'infaunal quality index' (IQI) for use by regulatory bodies to assess and manage the environmental impact of sea pen fish farming.

This project successfully achieved its original objectives in relation to the application of metabarcoding to aquaculture compliance monitoring by highlighting the varying precisions of NGS and the different genetic markers and identifying key indicator taxa, which demonstrated that the 16s bacterial marker offers the greatest precision. Moreover, that metabarcoding data, particularly where based around 16S data, can be used to predict macrobenthic IQI. The second phase of the project went on to further validate these findings by developing a ML algorithm and subsequent code, which has demonstrated the ability to combine the two key training data sets (MeioMetBar and DNAmon). The project team has successfully shown the proof of concept, where use of bacterial communities for compliance-based monitoring can be used instead of traditional macrobenthic analysis.

The project is highly technical with respect to the DNA sequencing, data analysis and statistics, and development of the ML algorithm to predict the IQI of samples based on their bacterial characterisation. The code and approach were independently reviewed by Biostatistics Scotland (BIOSS) under the direction of SEPA, which initiated a discussion about the approaches adopted and the limitations and opportunities offered. BIOSS, in discussion with the project partners, has made several recommendations that should improve the system going forward, and that link to the ongoing BactMetBar project.

The MeioMetBar project has developed the capacity, in Scotland, to apply NGS technology to aquaculture compliance assessment, representing a major step forward. This toolkit has the potential to be used across the sector to upload and test new data sets, and moreover, it has been developed as open-source, so its use is not restricted.

The unique collaboration between the academic community and the aquaculture regulatory bodies underpinning this work provides a clear path, in terms of uptake and implementation of the technology, which will enable environmental consultancies to apply this technology, in the near future, to facilitate the most cost-effective compliance demonstration.

In addition to the proof of concept, project partners have produced documentation surrounding the standard operating procedures (SOPs) for this technology, which details field sampling, sample storage and transportation, laboratory analysis, and an interim solution for the submission of data. Following these SOPs will ensure consistency among future users and keep work streamlined to maintain a standard of quality across the aquaculture sector as a whole.

The project addressed a clear sector requirement; achieved its technical objectives by providing compelling evidence that NGS-16S bacterial characterisation has considerable potential as a replacement for traditional macrobenthic analysis; and developed a useful new method for assessing the environmental impact of sea pen fish farming. Results from this work have the potential to provide significant and widespread benefits to both the aquaculture sector and the regulatory authorities. Furthermore, collaborative projects of this nature, bringing together both national and international academic institutions with sector partners and regulatory bodies are not only important, but a key factor in maintaining and protecting the aquatic environment while furthering the sustainable development of the aquaculture sector.

ADDITIONAL INFORMATION

Forster, D., G. Lentendu, S. Filker, E. Dubois, **T. A. Wilding** and T. Stoeck (2019). "Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants." Environ Microbiol. DOI: . 10.1111/1462-2920.14764

THE FOLLOWING PAPERS ARE IN PREPARATION:

Larissa F, Cordier T, Dully V, Breiner H-W, Pawlowski J, Catarina IM Martins, **Wilding T A**, Stoeck T* (TBC) "Supervised machine learning is superior to indicator value inference to monitor environmental impact of salmon aquaculture using eDNA metabarcodes".

Wilding, TA, Coulson, M, Stoeck, T. (TBC) "Taxonomic redundancy also applies in metabarcoding".

Wilding, TA, Coulson, M, Stoeck, T. (TBC) "COI v 18S v 16S – a comparison of markers in mapping environmental gradients around fish-cages".

REFERENCES:

Lejzerowicz, F., Esling, P., Pillet, L.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, 13932.