Deliverable D8.2

Report with industry standards and instructions for molecular tools in paternity assignment, traceability and marker assisted selection

Date: 1st September 2018
HORIZON 2020 - INFRADEV
Implementation and operation of cross-cutting services and solutions for clusters of ESFRI

Grant Agreement number: 654008
Project acronym: EMBRIC
Contract start date: 01/06/2015
Project website address: www.embric.eu

Due date of deliverable: 30/09/2018 / month 39
Dissemination level: Public
Deliverable 8.2 investigated different workflows in fish and shellfish and proposes standards for traceability and genetic marker discovery. Genomic resources generated in task 8.1 and 8.2 were used. Genomic analysis as well as parentage assignment pipelines were evaluated. The studies present indications and recommendations for development of industry standards and a minimum instruction set with sample data for the generation and application of molecular tools in paternity assignment, traceability and marker assisted selection. Genomic resources were generated for fish and shellfish i.e. Atlantic salmon (*Salmo salar*), greater amberjack (*Seriola dumerili*), Atlantic Bluefin tuna (*Thunnus thynnus*) and for the king scallop (*Pecten maximus*). In addition, for the King scallop two sample trait datasets were collected and used to establish new data reporting standards for EBI databases. Trait standards were modelled after existing (M2B3) standards and are now currently implemented for public use in EBI databases via the Webin system [https://www.ebi.ac.uk/ena/data/view/ERC000038](https://www.ebi.ac.uk/ena/data/view/ERC000038). For the European sea bass (*Dicentrachus labrax*) a novel phenotyping method combined with genomic information is suggested to improve selection of feed conversion ratio.
Table of Contents

1. Table of Contents ................................................................. 4
2. Introduction ........................................................................... 5
   2.2 References ........................................................................ 8
3. Bioinformatics Pipelines ........................................................... 9
   3.2 Workflow for whole genome sequencing approaches and evaluation (Partner 10, 12, 25) ..... 9
   3.3 Workflow for parentage assignment using micro-haplotypes using King scallop as an example (Partners 10, 12, 23) ................................................................. 13
   3.4 Publications ........................................................................ 16
   3.5 Bioinformatics tools .......................................................... 17
   3.6 References ........................................................................ 17
4. Genomic resources for generating molecular tools of key species ........ 18
   4.2 Atlantic Bluefin tuna (Thunnus thynnus) (Partners 11 and 25) ........................................ 18
   4.3 Greater amberjack genome sequencing (Partner 25) ...................................................... 21
   4.4 Atlantic salmon (Salmo salar L.) Paternity assignment and traceability tools and genetic marker discovery pipelines (Partners 10 and 12) ......................................................... 22
   4.5 King scallop (Pecten maximus): Trait datasets for the King scallop (Pecten maximus)—an example of a highly polymorphic molluscan genome (Partners 6, 10 and 12) ............. 23
   4.6 References ........................................................................ 25
5. Measuring complex traits in fin and shellfish .................................... 26
   5.1 European sea bass (Dicentrarchus labrax): Using a novel phenotyping method combined with genomic information to improve selection of feed conversion ratio in sea bass (Partner 13) 26
   5.2 Atlantic salmon: Developing experimental disease models underpinning investigation of the genetic basis disease resistance (Partner 6) .................................................. 40
   5.3 Blue mussels (Mytilus edulis): Developing experimental disease models underpinning investigation of susceptibility of blue mussels to bacterial infections (Partner 6) ............. 51
   5.4 King scallop (Pecten maximus): Characterization of baseline microbiome of healthy king scallop (Partner 6) ................................................................. 69
2 Introduction

Only 10% of aquaculture production worldwide is based on genetically improved stocks (Gjedrem et al 2012). One of the main factors that hinders the development of optimized selective breeding programs in aquaculture is the technical weakness of obtaining pedigree information with an easy and cost-effective way. Pedigree information is essential for the implementation of selective breeding strategies that manage to avoid inbreeding depression and loss of performance (Kincaid 1983). In livestock or poultry, pedigrees can be easily maintained by tagging of offspring with physical tags. In most farmed aquatic species, however, physical tagging of young animals at hatching is very difficult and time-consuming due to their great numbers and small size.

Two ways have been traditionally adopted in order to overcome the pedigree issue in aquaculture (see Figure 1):

1. **Single-family rearing in individual tanks**
   In this case offspring from various families remain separated in different tanks until they are large enough to be physically tagged, as in the Norwegian salmon breeding program, the first family-based selective breeding program in aquaculture, started in 1972 (Gjedrem 2010). This approach secures the availability of pedigree information but requires high operational costs that in some cases makes the implementation unfeasible.

2. **Molecular parentage analysis of mixed families reared together**
   Molecular parentage removes the need for single-family rearing allowing the reconstruction of pedigrees using genotypic data in cases that individual crosses are prohibitive due to biological (e.g mass spawning species) or economical constraints. Pedigrees can be traced in groups of mixed families with any type of family structure. In addition, this approach allows physical tagging at a much later age, and offers the opportunity to link phenotypes from individuals grown in commercial ponds to
broodstock grown in specialized tank systems. Molecular parentage analysis can therefore be used to estimate genetic parameters for selective breeding and to estimate reproductive success for hatchery management.

The use of molecular techniques to determine parentage utilizes the principle of Mendelian inheritance to assign offspring to potential parents. According to Mendelian inheritance, a diploid offspring inherits one of two alleles at each locus from each of its parents, therefore carries one allele from the mother and the other from the father for all loci. Paternity can be determined after comparison of genotypes at co-dominant molecular markers of progenies with those of potential parents. There are six different methodological approaches to parentage analysis (see Box 1 below, reproduced from Jones et al 2010), but all are built on the basic exclusion-based method. The exclusion method determines parent-offspring relationships by screening offspring against all potential parents to exclude potential parents that fail to share at least one allele at all loci. If a potential parent does not share an allele with the offspring in question, then that individual can be excluded from consideration as a true parent. In some cases, however, difficulties or errors in scoring alleles are encountered (e.g. null alleles, preferential amplification of some alleles at some marker loci) and thus some genotypes are wrongly scored either in potential parents or offspring, rendering complete exclusion impossible. In such cases, categorical parentage allocation approach can be applied to analyze the remaining non-excluded candidate parents and offspring. In this approach, the entire offspring is assigned to the candidate parent with the highest likelihood or posterior probability of being the true parent. In aquaculture species, categorical allocation is the most commonly used method of parentage analysis, since it is appropriate for handling scoring errors or mutations, absence of all parents and can include methods for determining confidence in parentage assignment.
While the very first trials of parentage assignment in fish were done in the 1970s in Israel using allozymes, the real start of parentage assignment studies was in the 1990s with the availability of microsatellite markers (Herbinger et al 1995; Estoup et al 1998). Microsatellites have become the main marker of choice for parentage analysis owing to their properties: they are abundant, easily amplified in multiplexes and genotyped, co-dominant, and highly polymorphic markers (Goldstein and Schlotterer 1999). During the last 20 years microsatellites have been used for the parentage analysis of numerous aquaculture species, such as salmon, tilapia, Asian sea bass, common carp, rainbow trout, European sea bass and sea bream (Yue and Xia, 2014 and ref. therein). In general 8–15 polymorphic microsatellite markers provide adequate

Box 1: Six approaches to parentage analysis

**Exclusion** – The exclusion method takes advantage of the fact that in diploid, sexually reproducing organisms, each parent shares at least one allele per locus with each of its offspring. In this approach, the genotypes of candidate parents are compared with that of a focal offspring. Any candidate parent who fails to share at least one allele with the offspring at any locus is eliminated from consideration. In practice, most exclusion studies actually require at least two mismatching loci between the candidate and the offspring to account for typing errors or mutations.

**Categorical Allocation** – If complete exclusion is impossible, then a parentage allocation approach (also known as parentage assignment) can be used to choose among the remaining non-excluded candidate parents. In categorical assignment, the entire offspring is assigned to the candidate parent with the highest likelihood or posterior probability of being the true parent. Categorical assignment approaches can handle scoring errors or mutations and can include methods for determining confidence in parentage assignment.

**Fractional Allocation** – In the fractional allocation approach, likelihoods or posterior probabilities are determined in the same way as in the categorical assignment methods. Each offspring is then assigned partially to each of the nonexcluded candidate parents on the basis of their relative likelihoods of parentage. Even though a fractional assignment has no biological meaning, from a statistical standpoint, this approach may have better properties than categorical allocation.

**Full Probability Parentage Analysis** – The full probability approach estimates patterns of parentage in a modelling framework. Many different models are possible, but this approach has the potential to estimate simultaneously patterns of parentage and other population-level variables of interest. This approach makes better use of the data by incorporating any uncertainty in the parentage analysis into the estimation of the variables of interest.

**Parental Reconstruction** – The parental reconstruction technique uses the genotypes of offspring in full- or half-sib families to reconstruct parental genotypes. For full- or half-sib progeny arrays, all of the offspring will share at least one parent. The genotype of the shared parent may be available from the sampling scheme or can be reconstructed by identifying a pair of alleles, for which every offspring inherited at least one of the members of the pair. The genotypes of the unknown parents can be determined by examining associations of alleles originating from the unknown parents across loci. Available techniques are based on parsimony (i.e. assuming the minimum number of parents), maximum likelihood or Bayesian approaches. Once the genotypes are reconstructed, they can be compared with the genotypes of candidate parents to assign parentage.

**Sibship Reconstruction** – If no parents are available and known groups of full- or half-sibs cannot be sampled, then sibship reconstruction is the last resort in the realm of parentage analysis. This technique requires a sample of individuals, some of which are full- or half-sibs. The algorithms use patterns of relatedness or maximum likelihood techniques to group individuals into different classes of relationship, often full-siblings, half-siblings and unrelated individuals. Once half-sib or full-sib groups are identified by these approaches, the parental genotypes can be reconstructed and used for parentage analysis.

(Jones et al 2010)
assignment power for crosses involving a few tens or hundreds of parents. The main drawbacks of microsatellites are considered the genotyping errors (particularly in automated multiplex systems) and lack of standardization between labs. Among microsatellites, tri- and tetranucleotide microsatellites are relatively easier to genotype and less prone to genotype errors due to less stuttering effects during PCR. However, they are usually less abundant in genomes and less polymorphic in comparison to dinucleotide microsatellites (Goldstein and Schlotterer 1999), thus a greater investment in time and money is usually required for developing a multiplex panel with adequate assignment power.

The aim of task 8.2 was to investigate different workflows to assess genetic variability in fish and shellfish and to suggest standards for traceability and genetic marker discovery. Genomic resources generated in task 8.1 and 8.2 were used as test data sets (Partner 6, 10, 11, 12, 25). Furthermore, genomic information was combined with a novel phenotyping method to improve the feed conversion ratio in sea bass (Partner 13). In addition, for the King scallop two stocks were identified using the generated tools (Partner 6, 10, 12). Furthermore, workflow for the detection of structural variations through resequencing the Atlantic salmon genome is presented (Partner 11 and collaborators) and genomic data generated for the greater amberjack and the Bluefin tuna significantly enriches the molecular toolbox for marker discovery and usage (Partner 25).

2.2 References


3 Bioinformatics Pipelines

3.2 Workflow for whole genome sequencing approaches and evaluation (Partner 10, 12, 25)

3.2.1 Introduction
Genome assemblies can be categorized into draft, high quality and complete assembly, with no clear boundaries amongst the three. For large genomes, even ‘complete’ assemblies have many gaps and unsequenced regions. The human genome, despite the amount of sequencing effort put into it, problematic regions remain, such as very low or high GC-regions, and highly repetitive regions such as the centromeres and telomeres. Assemblies will in addition contain various errors, such as substitutions, indels, and chimeric parts. The types of errors occurring in the raw sequence will be affected by the sequencing technology, while assembly errors will additionally depend on factors such choice of assembly tool, parameter choices in the tool, as well as the complexity of the specific genome one is assembling.

3.2.2 Sequencing technologies
As of 2018, the dominant NGS sequencing technology is Illumina’s Sequencing By Synthesis (SBS), with read lengths up to 2*300 bases, but most commonly 2*150 bases in the HiSeq family of instruments. The technology is mature and delivers high accuracy and low cost per base. The main drawback of the technology is the short read lengths, which makes it difficult to resolve repetitive regions in genomes. There is also a coverage bias against both low and high GC-content; the technique works best with intermediate GC-content. Some methods exist to alleviate this (e.g. Tilak et al 2018). Regions with high GC-content have also been shown to have a slightly higher incidence of errors (Allhoff et al 2013), which could lead to false inferences of SNPs.

Two technologies exist that can give much longer reads than with SBS, namely Single Molecule RealTime sequencing (SMRT) (Pacific Biosciences), and Nanopore sequencing (Oxford Nanopore Technologies). Both technologies currently have error rates around 10-15 %; the Pacbio reads have random errors, while the Nanopore reads have a problem with long homopolymer stretches and consequently have a lower consensus accuracy than other methods. Nanopore basecallers are frequently updated, thus it is possible to re-basecall old sequence data and possibly get lower error rates. Other methods exist that give ‘synthetic’ long reads with clever barcoding methods, such as Illumina TruSeq synthetic long reads, but these methods cannot resolve long simple tandem repeats.

3.2.3 DNA extraction
One of the first steps in a genome sequencing project is to extract high-quality DNA. Getting a good amount of high molecular weight, pure DNA is of paramount importance and will make all subsequent steps easier. In general, recommended protocols are: Qiagen Genomic Tip kits, or phenol-chloroform manual extraction, optionally with Qiagen MaxTract for easier separation of the phases. Other protocols may work equally well or better; this must be evaluated on a case-by-case basis and may depend on the species, tissue type, preservation method, and other factors. For evaluation of the DNA quality and quantity it is highly recommended to use all three of...
the following methods: Nanodrop for purity estimation, agarose gel (and/or Pulsed Field Gel Electrophoresis or Femtopulse) for DNA size and integrity, and a fluorometric assay such as Qubit or Picogreen for accurate quantification. Under no circumstances quantify with Nanodrop, as it may overestimate the DNA concentrations, sometimes by orders of magnitude.

3.2.4 Illumina library construction
In order to achieve the most even genome coverage it is important to try to avoid sources of bias. Usually the first step in library preparation consists of fragmenting the DNA into a suitable size range. Mechanical DNA shearing (by focused ultrasonication) is almost random, but GC-content does influence to a small degree the likelihood of fragmentation occurring at a specific site (Poptsova et al 2014). A small amount of base substitutions may occur during sonication (Costello et al 2013), probably as a result of oxidation happening in samples which contain some carryover contaminants from the DNA extraction. This may be partially alleviated by adding certain antioxidants to the solution being sonicated, and should not be a problem unless attempting to detect extremely low frequency variants. It is also possible to shear DNA with pressurized nitrogen gas (nebulization), but this leads to some sample loss through evaporation. Enzymatic shearing gives less even coverage than mechanical methods. GC-rich regions will have a lower coverage, and this can have a negative effect on variant calling (Lan et al 2015).

If possible, avoid library protocols involving a PCR step. PCR by its nature introduces bias and therefore less even genome coverage (Kozarewa et al 2009, Aird et al 2011). If PCR can not be avoided, attempt to minimize the number of PCR cycles. PCR-free protocols require larger amounts of DNA. If only small amounts of DNA are available, prefer protocols with PCR enrichment over protocols with whole genome amplification.

If libraries are to be indexed and pooled, take care to choose index combinations that generate sufficient base diversity. Confer ‘What is nucleotide diversity and why is it important?’, https://support.illumina.com/bulletins/2016/07/what-is-nucleotide-diversity-and-why-is-it-important.html for details about base diversity in Illumina sequencing.

After Illumina’s introduction of patterned flow cells there have been some concerns over index mis-assignments or index “hopping” (Sinha et al 2017, but see also van der Valk et al 2017). This is potentially a problem if several samples are multiplexed in the same lane. To minimize this issue, it is important to do thorough library cleanups. If an adapter peak is seen in a Bioanalyzer trace, perform additional cleanups until it disappears. To further reduce this problem it is possible to use dual indexing with unique indexes at both ends (Kircher et al 2011). It may be also possible to bioinformatically filter out much of the index “hops” (Wright and Vetsigian 2016, Bartram et al 2016), since misassigned indexed reads tend to have lower quality scores than correctly assigned indexes. See also Illumina’s “Best practices for multiplexing samples and minimizing index mis-assignment” (https://support.illumina.com/bulletins/2017/05/best-practices-for-multiplexing-samples-and-minimizing-index-mis.html) and “Insights into Index Hopping Illumina Webinar” (https://www.youtube.com/watch?v=DR_8KbGGiAh).

3.2.5 Nanopore library construction
As for Illumina libraries, PCR-free methods are preferred for less coverage bias. On the other hand, the inclusion of a PCR-step will produce DNA free from base
modifications, which may in some cases work better for Nanopore sequencing. Community experience has shown that fresh DNA extracts tend to give longer read lengths than older extractions, even if no visible breakdown of the DNA can be seen. It is recommended not to freeze extracted DNA since this tends to reduce read lengths. There may be significant variations in sequence yields and sequence quality from species to species, irrespective of the purity of the template DNA. Generally the 1D ligation method (LSK108) has shown itself to be a good method for high yields and read lengths. Although DNA fragmentation is an optional step in this protocol, experience has shown that without fragmenting the DNA, sequence yields tend to be extremely low. DNA may be mechanically fragmented by centrifugation through a narrow orifice in a Covaris G-tube. It is important that the solution containing the DNA is completely free of particulate matter which may otherwise enter the orifice and interfere with the shearing. Very high DNA concentrations may also lead to inconsistent DNA shearing as the liquid becomes more viscous. To enrich for long reads in the finished library, small DNA fragments can be removed in a BluePippin instrument. A recently released Nanopore kit (LSK109) promises to eliminate the need for fragmentation and thus deliver longer read lengths. At the time of writing there is little experience with this kit in the Nanopore community. Note that at present two pore versions are available from ONT - R9.4 and R9.5 respectively. It is recommended to use R9.4, since experience has shown R9.5 to give lower yields and poorer sequence qualities. In fact the inclusion of the R9.5 sequences (Run4) in the Hybrid assemblies worsened the N50 and other statistics. Oxford Nanopore has acknowledged to us that R9.5 may have a problem.

3.2.6 Coverage recommendations
For de novo assembly from Illumina short reads it is suggested to have around 100x genome coverage. For protocols with whole genome amplification, the coverage should be much higher. For hybrid assembly with Nanopore and Illumina reads it is recommended to have at least 10x Nanopore coverage, although a higher coverage will be better. For Nanopore only assemblies it is recommended to have at least 30x coverage of reads above 10Kb. The longer the reads, the lower the needed coverage, and vice versa. An efficient strategy to get larger contigs is to include a run with a library specially prepared for ultra-long reads, at a low coverage, along with the regularly prepared Nanopore libraries; see http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/

3.2.7 Data processing
For Illumina reads, adapter trimming is usually done automatically by the instrument, and remaining small amounts of adapter will normally be removed automatically as part of the assembly process (table 1). In case there is a need for more stringent adapter trimming one can use e.g. tools like Cutadapt and Scythe. To quality trim or not is somewhat controversial. If starting with high genome coverage dataset it is probably better to do some moderate quality trimming, with e.g. Sickle. A metric that may help guide decisions about trimming is the base proportion along the read length (i.e. the base proportions at each sequencing cycle); in a randomly fragmented genome this should be constant from start to end. If base proportions changes towards the start or end of reads it is an indication that the sequence is no longer random, and one should consider carefully what could be the cause and if further trimming is necessary. Typically there is some deviation from randomness in the first 10 or so cycles and in some cases in the very last cycle. Reasons for this are unclear, but may
be related to the mentioned fragmentation bias (Poptsova et al 2014) or bias in the ligation of adapters (Seguin-Orlando et al 2013).

Usually a small amount of PhiX genomic DNA library is added to Illumina sequencing runs in order to give quality metrics of the run. Although most of the PhiX reads are automatically filtered out by the onboard instrument software, significant amounts of PhiX sequence typically remains and must be filtered out during subsequent data processing.

For Nanopore reads, adapters will be present in most reads. Due to the high error rate of the technology the adapters will appear to vary quite a lot in length and exact sequence. The tool PoreChop is designed to trim the adapters, but for genome assemblies it may not be necessary to trim, as the assembly process itself removes the adapters. Nevertheless it is good practice to make sure there is no leftover adapter sequence in any contigs after the assembly. Quality trimming may or may not improve an assembly; it is suggested to do only mild quality trimming, e.g. ≥ Q7. As with Illumina reads, a known viral genome may be added during the library prep; in Nanopore sequencing this is a 3.6 Kb piece of the lambda phage, and the sequence, if found in an assembly, must be filtered out.

3.2.8 Assembly
Hybrid assembly of Illumina short reads and nanopore reads (Fig. 2) was performed with MaSuRCA v3.2.2. Numerous options exist for optimization of an assembly, and it was not feasible to fully explore the whole parameter space. The table shows the settings used for the final assembly. Other assemblers were also used (SPAdes, Canu), but they did not perform as well as MaSuRCA with our dataset. BUSCO was used to evaluate the quality and completeness of the assemblies.

Table 1. Settings for the assembly producing the best N50 (=265095 bp).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Read type</th>
<th>Comment</th>
<th>Viet Hybrid123 assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina quality trim</td>
<td>Illumina</td>
<td>Trim low-quality 3' end</td>
<td>no</td>
</tr>
<tr>
<td>Illumina normalize (Trinity)</td>
<td>Illumina</td>
<td>Coverage normalization w/ Trinity</td>
<td>no</td>
</tr>
<tr>
<td>Minlength</td>
<td>Nanopore</td>
<td>Minimum read length to be included, bp</td>
<td>1000</td>
</tr>
<tr>
<td>Allowed error rate</td>
<td>Nanopore</td>
<td>Maximum error rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Porechop</td>
<td>Nanopore</td>
<td>adapter trim</td>
<td>no</td>
</tr>
<tr>
<td>Basecaller version</td>
<td>Nanopore</td>
<td></td>
<td>Albacore v1.2.4</td>
</tr>
<tr>
<td>Quality filter</td>
<td>Nanopore</td>
<td>Exclude low quality reads</td>
<td>no</td>
</tr>
<tr>
<td>Canu/ed</td>
<td>Nanopore</td>
<td>Canu &quot;self-correction&quot; of reads</td>
<td>no</td>
</tr>
<tr>
<td>Nanopolish</td>
<td>Nanopore</td>
<td>reference-based error correction</td>
<td>no</td>
</tr>
<tr>
<td>Nanopore version</td>
<td>Nanopore</td>
<td>R9.4, R9.5</td>
<td>R9.4</td>
</tr>
<tr>
<td>Assembler/MaSuRCA version</td>
<td></td>
<td></td>
<td>v3.2.2</td>
</tr>
<tr>
<td>Runtime</td>
<td></td>
<td>IMBBC &quot;Zorba&quot; computer cluster</td>
<td>6d 5h</td>
</tr>
</tbody>
</table>

3.2.9 Nanopore consensus accuracy improvement
Nanopolish is currently regarded as the standard for improving Nanopore consensus accuracy. Several other polishers exist and the landscape is changing rapidly. It has been shown that using more than one polisher may give better results. For instance, using Pilon followed by Racon gives a better accuracy than using either alone. The
polishers may be used iteratively; the accuracy will tend to improve with each successive round, up to about 4-5 rounds. Whenever a PCR-free method is used, the sequenced DNA will contain some modified bases which affect the signal as the bases pass through the pore. Nanopolish can optionally use methylation-aware polishing, which may further increase consensus accuracy. When Illumina reads are available, they are used for error corrections instead of a polisher.

3.2.10 Mapping, visualization
For mapping of Illumina reads to a reference, BWA-mem and SAMtools were used. For mapping Nanopore reads onto a reference GraphMap, Minimap2 or Miniasm can be used. For general visualization and assessment of Nanopore data, Nanoplot is a convenient tool. For visualizing mapped reads, Tablet is a good tool.

3.2.11 Variant calling
The GATK pipeline will be used for SNP discovery and variant calling.

![Diagram of genome hybrid assembly]

**Figure 2** Example pipeline for genome hybrid assembly.

3.3 Workflow for parentage assignment using microhaplotypes using King scallop as an example (Partners 10, 12, 23).

The SNP resource generated for King scallop in D8.1 was further filtered to identify a number of microhaplotype loci suitable for parentage assignment using an amplicon sequencing approach. A panel of 24 microhaplotype loci was developed based on potential informativeness for parentage assignment (number of haplotype alleles and expected heterozygosity) and conformance to Hardy-Weinberg equilibrium expectations in wild populations. In addition, a bioinformatic pipeline was developed for processing raw amplicon sequencing data into microhaplotype-based genotypes suitable for parentage assignment ([https://github.com/stan-fmrg/Snakemake_haps](https://github.com/stan-fmrg/Snakemake_haps)).
The bioinformatic pipeline developed is illustrated in Figure 3 and can be used with both paired-end and single-end reads. It was designed using the workflow manager Snakemake. The workflow is executable from a command line environment in an easy and straightforward way with no prior knowledge of scripting being necessary. The pipeline is comprised of four steps. The read pre-processing (1), the mapping of the pre-processed reads to a reference genome/sequence (2), the variant calling and filtering (3) and the haplotype calling based on the high-quality and high-confidence SNPs identified by the workflow (4). The first step includes the quality control and pre-processing of raw FASTQ files. Raw FASTQ files need to be quality trimmed to minimize possible sequencing errors being included in downstream analysis. A second source of error can be contamination from adapter sequences included in the reads. If the DNA fragments in the library are below the target size, then the sequencing can continue on into the adapter sequence. Trimmomatic can deal with both of these issues in a single step. The next step is the mapping of the pre-processed reads to the available reference genome/sequence, using BWA-mem, where the reads are mapped to the best possible location and resulted files used for the next step which is the variant calling. The variant calling step uses the existing mapping of the read to iterate over all locations and checking each of the variants. Coverage (the number of times a given region has been sequenced) and QUAL scores (a phred-like score that expresses the likelihood that the variant is real) are important parameters for reliable variant discovery and reduce the possibility of False-Positive variants called as real ones. The produced results are stored in a Variant Call Format (VCF) file, which can be used for the calling of microhaplotype-based genotypes suitable for parentage assignment.

Figure 3 Genetic marker discovery workflow illustrating a detailed analysis of RAD-seq including read pre-processing, variant calling and filtering, and finally haplotype calling using the high-confidence SNPs identified from the workflow. (Tools used for every step of the workflow and their parameters are given adjacent to every box). Parameters illustrated in the workflow are the proposed parameters, however may be easily adjusted depending on the sequencing data that are analyzed.

The development of a bioinformatic workflow for a Whole Genome Association Study (GWAS) (Partner 10).
The workflow developed is shown in Figure 4. It makes use of GAPIT - Genome Association and Prediction Integrated Tool – which is an R package that can use a number of methods, such as the unified mixed model, EMMA, the compressed mixed linear model, and P3D/EMMAx. GAPIT can handle large data sets in a fast and easy way for the user. Required input files for the analysis include phenotypic and genotypic data, as well as a kinship matrix, which can either be provided by the user or can be computed by GAPIT using the VanRaden (VanRaden, 2008) method. Multiple phenotypes can also be used in GAPIT when all observed phenotypic data for each individual are included in the phenotypic text file. Multiple analysis parameters are available and explained in the GAPIT manual (http://www.zzlab.net/GAPIT/gapit_help_document.pdf), which based on the input data provided, should be adjusted.

Figure 4 Workflow for Genome Wide Association Study (GWAS) and genome prediction (or selection) using high-quality SNPs as genetic markers for the identification of associations with available phenotypes.

Workflow for the detection of structural variations through resequencing the Atlantic salmon genome (Partner 11 and collaborators)

Finding traits impacted by structural variants (SVs) using next generation sequencing (NGS) requires the accurate and consistent identification of SVs in each new genome sequenced. When SVs are being catalogued for the first time in a species, multiple bioinformatic steps are necessary to translate a whole genome sequence output to a set of high confidence calls.
We present a comprehensive and replicable workflow to successfully identify variants in the Atlantic salmon (Salmo salar) genome from raw sequencing reads to a squared-off file containing genotypes and trait associations. This pipeline can be adapted to any species with a reference genome, with section three (Figure 5) being crucial for novel SV characterization.

The workflow presents a single framework that can be divided into five main sections. Aside from section 1, which includes laboratory work, it is designed to be performed sequentially in one script.

Section 1: Genomic DNA is extracted and sequenced on a standard Next Generation Sequencer (NGS).

Section 2: Raw read data is checked for sequencing errors, aligned to a reference genome and transformed into a binary format necessary for downstream analyses.

Section 3: Alignment files are checked for aberrant results, genome-wide depth of coverage and overall alignment quality. All complex regions including assembly gaps and regions of extremely high coverage are identified and grouped into bed files. This section is essential when identifying SVs for the first time to reduce the number of false positive calls in section 4.

Section 4: Structural variant programmes are used to identify SVs outside of complex regions determined in section 3 and unanchored contigs.

Section 5: SVs present in all samples are removed. Then all SVs are manually curated using a cloud-based, multi-user framework. The curated SVs can then be annotated a queried for trait association to produce a final file in the variant calling format (VCF) that can be used as standard input for any downstream analyses.

![Flowchart](https://www.frontiersin.org/articles/10.3389/fgene.2018.00253)

**Figure 5** Flowchart detailing steps for SV calling in Atlantic salmon genome. In every section, the boxes detail the number of bioinformatic stages that need to be ran and completed before the next section is started. Above each arrow is the name of the programming tool required for the completion.

### 3.4 Publications


EMBRIC – D8.2: Report with industry standards and instructions for molecular tools in paternity assignment, traceability and marker assisted selection, page 16 of 76.
3.5 Bioinformatics tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUSCO</td>
<td><a href="https://busco.ezlab.org/">https://busco.ezlab.org/</a></td>
</tr>
<tr>
<td>BWA</td>
<td><a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a></td>
</tr>
<tr>
<td>Canu</td>
<td><a href="https://github.com/marbl/canu">https://github.com/marbl/canu</a></td>
</tr>
<tr>
<td>GATK</td>
<td><a href="https://software.broadinstitute.org/gatk/">https://software.broadinstitute.org/gatk/</a></td>
</tr>
<tr>
<td>GraphMap</td>
<td><a href="https://github.com/isovic/graphmap">https://github.com/isovic/graphmap</a></td>
</tr>
<tr>
<td>MaSuRCA</td>
<td><a href="https://github.com/alekseyzimin/masurca">https://github.com/alekseyzimin/masurca</a></td>
</tr>
<tr>
<td>Miniasm</td>
<td><a href="https://github.com/lh3/miniasm">https://github.com/lh3/miniasm</a></td>
</tr>
<tr>
<td>MiniMap2</td>
<td><a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a></td>
</tr>
<tr>
<td>NanoPlot</td>
<td><a href="https://github.com/wdecoester/NanoPlot">https://github.com/wdecoester/NanoPlot</a></td>
</tr>
<tr>
<td>Nanopolish</td>
<td><a href="https://github.com/jts/nanopolish">https://github.com/jts/nanopolish</a></td>
</tr>
<tr>
<td>Pilon</td>
<td><a href="https://github.com/broadinstitute/pilon/wiki">https://github.com/broadinstitute/pilon/wiki</a></td>
</tr>
<tr>
<td>PoreChop</td>
<td><a href="https://github.com/rrwick/Porechop">https://github.com/rrwick/Porechop</a></td>
</tr>
<tr>
<td>Racon</td>
<td><a href="https://github.com/isovic/racon">https://github.com/isovic/racon</a></td>
</tr>
<tr>
<td>SAMtools</td>
<td><a href="http://samtools.sourceforge.net/">http://samtools.sourceforge.net/</a></td>
</tr>
<tr>
<td>Scythe</td>
<td><a href="https://github.com/vsbuffalo/scythe">https://github.com/vsbuffalo/scythe</a></td>
</tr>
<tr>
<td>Sickle</td>
<td><a href="https://github.com/najoshi/sickle">https://github.com/najoshi/sickle</a></td>
</tr>
<tr>
<td>SMARTdenovo</td>
<td><a href="https://github.com/ruanjue/smartdenovo">https://github.com/ruanjue/smartdenovo</a></td>
</tr>
<tr>
<td>SPAdes</td>
<td><a href="http://cab.spbu.ru/software/spades/">http://cab.spbu.ru/software/spades/</a></td>
</tr>
<tr>
<td>Tablet</td>
<td><a href="https://ics.hutton.ac.uk/tablet/">https://ics.hutton.ac.uk/tablet/</a></td>
</tr>
<tr>
<td>Trinity</td>
<td><a href="https://github.com/trinityrnaseq/trinityrnaseq/wiki">https://github.com/trinityrnaseq/trinityrnaseq/wiki</a></td>
</tr>
</tbody>
</table>

3.6 References


4 Genomic resources for generating molecular tools of key species

4.2 Atlantic Bluefin tuna (*Thunnus thynnus*) (Partners 11 and 25)

4.2.1 Genomic data

4.2.1.1 Tuna genome sequencing

In total eight tuna samples obtained from Malta (4) and from Croatia (4) were sequenced using Illumina NGS technology (Table 4.2.1) to obtain putative variations between the two population resources. For one sample (BFT4) insert size and coverage were calculated (Fig. 4.2.1). To generate a high quality reference genome for the Atlantic Bluefin tuna from one sample, BFT 4, three Nanopore libraries were made (figure 4.2.2) using the now discontinued 2D ligation kit (LSK208) and R9.4 flow cells, and one library with a 1D2 (LSK308) ligation kit and an R9.5 version flow cell. For each library, two to three micrograms of DNA was sheared by centrifugation in a G-tube to a nominal size of 8 Kb. Library construction followed the standard protocols, with the optional FFPE repair step. Each library was run for 48 hours in a MinION sequencer. The raw (fast5) reads were basecalled and output as fastq-files with the Albacore basecaller. The first three runs (LSK208) produced three types of reads: ‘template’, ‘complement’, and ‘2D’, which corresponds to the first DNA strand, its complementary strand, and the consensus of the two reads, respectively. Around 62 % of the reads were 2D; the remained had either only the first strand or only the complementary strand. A custom script was written to keep only unique reads, i.e. the template and complement were not kept for reads that also had a 2D read. The reads from the 1D2 run produced two types of reads; template and consensus. In this library type both strands are considered templates and have equal qualities. Only about 20 % of the reads formed consensus reads. A similar custom script as above was employed to keep only unique reads. A small proportion of reads formed false consensuses, which could be identified by the same template reads occurring in more than one consensus. The false consensus could be identified and filtered out from its lower quality score. Reads from the R9.5 pore/1D2 kit did not assemble well, and in the end were not used for the best assembly. When mapping these reads onto our preliminary reference genome, the R9.5 reads had equal overall error rates, but more secondary alignments than R9.4 reads, and were more frequently split into very short pieces (< 500 bp) of mapped reads. Reasons for this are as yet unclear; furthermore it is unclear to what degree this was caused by the pore version (R9.5 vs. R9.4) vs. the library kit version (1D2 vs. other).
Table 4.2.1 Atlantic bluefin tuna samples and sequence yields from Illumina and Nanopore sequencing. The genome coverage is estimated based on an expected genome size of 800 Mbases. The “Nanopore Gbases” does not include Run4.

<table>
<thead>
<tr>
<th>ID</th>
<th>Area</th>
<th>sex</th>
<th>Raw reads</th>
<th>Trimmed reads</th>
<th>Trimmed Gbases</th>
<th>~ coverage</th>
<th>Nanopore Gbases</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFT-1</td>
<td>Malta</td>
<td>m</td>
<td>546585362</td>
<td>522564550</td>
<td>76.73</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>BFT-2</td>
<td>Malta</td>
<td>f</td>
<td>540546178</td>
<td>515259238</td>
<td>75.65</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>BFT-3</td>
<td>Malta</td>
<td>f</td>
<td>557973596</td>
<td>529689686</td>
<td>77.79</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>BFT-4</td>
<td>Malta</td>
<td>m</td>
<td>566176332</td>
<td>537748554</td>
<td>78.95</td>
<td>99</td>
<td>3.85</td>
</tr>
<tr>
<td>BFT-5</td>
<td>Croatia</td>
<td></td>
<td>567432320</td>
<td>542550082</td>
<td>79.67</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>BFT-6</td>
<td>Croatia</td>
<td></td>
<td>536864432</td>
<td>506717596</td>
<td>74.51</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>BFT-7</td>
<td>Croatia</td>
<td></td>
<td>689645378</td>
<td>655465506</td>
<td>96.23</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>BFT-8</td>
<td>Croatia</td>
<td></td>
<td>596267888</td>
<td>566346554</td>
<td>83.23</td>
<td>104</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.2.1 Calculated insert size (a) and calculated genome coverage (b) from Illumina reads in sample BFT4 in Sample BFT4.

Figure 4.2.2. Nanopore read lengths (BFT4).
Comparison of the three NGS technologies used as well as the combined approach (Illumina technology and MinION technology) revealed that although the expected genome size has been retrieved by Illumina only sequencing the contig number is high. On the other hand low number of contigs are obtained when applying only the MinION technology but the required genome size is not reached. Looking at the hybrid assembly approach the desired genome size as well as coverage is given, with a lower number of contigs than when using Illumina only (Figure 4.2.3 and Table 4.2.2)

![Cumulative genome lengths](image)

**Figure 4.2.3.** Cumulative genome lengths.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>N50</th>
<th>Largest contig</th>
<th>Assembler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina trimmed</td>
<td>26317</td>
<td>1000448</td>
<td>Masurca</td>
</tr>
<tr>
<td>Illumina untrimmed</td>
<td>19000</td>
<td>631785</td>
<td>Masurca</td>
</tr>
<tr>
<td>Nanopore Run1-3</td>
<td>27158</td>
<td>241949</td>
<td>Canu</td>
</tr>
<tr>
<td>Nanopore Run1-4</td>
<td>60687</td>
<td>389535</td>
<td>Canu</td>
</tr>
<tr>
<td>Hybrid, MinION Run1-3</td>
<td>265095</td>
<td>2006256</td>
<td>Masurca</td>
</tr>
<tr>
<td>Hybrid, MinION Run1-4</td>
<td>162310</td>
<td>1256260</td>
<td>Masurca</td>
</tr>
</tbody>
</table>

For assembly approaches BUSCO analysis was carried out and Table 4.2.3 shows the results of the best assembly for each technology applied in the present work. Clearly the hybrid approach had the best outcome.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Complete</th>
<th>Single</th>
<th>Duplicated</th>
<th>Fragmented</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina trimmed</td>
<td>83</td>
<td>76</td>
<td>7</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Hybrid Run1-3</td>
<td>93.3</td>
<td>89.9</td>
<td>3.5</td>
<td>2.1</td>
<td>4.6</td>
</tr>
<tr>
<td>Nanopore Run1-4</td>
<td>15.2</td>
<td>15.2</td>
<td>0</td>
<td>10.6</td>
<td>74.2</td>
</tr>
</tbody>
</table>

Annotation was carried out with AUGUSTUS (version 3.2.3) against zebrafish.gff3 resulting in 63,538 putative genes demonstrating that most of a general fish transcriptome was covered by the assembly.
4.3 Greater amberjack genome sequencing (Partner 25)

The present work was performed in collaboration with the national project KRHPIS and valuable molecular resources have been generated for the Greater Amberjack (Seriola dumerili) in form of a first draft genome applying MinION as well as Illumina sequencing technologies. Therefore whole genome sequencing as well as 8 transcriptomes of the greater amberjack were sequenced using Illumina sequencing technology. Complementary to our work the male genome of the greater amberjack was published in NCBI. In this context additional Minlon run using the to-date latest Minlon chemistry was carried out in order to obtain the female genome sequence applying Minlon technology with the latest chemistry available.

In total three MinION runs were carried out. The first two with R9.5 pore and 1D2 sequencing kit (LSK308); the last with R9.4 pore, Rev D flow cell with 96 hours runtime, and 1D sequencing kit (LSK108). While the former showed to have serious issues while assembling, the later resulted in raw yield of 18.1 Gbases and with a mean read length of about 9K (Figure 4.3.1 a). With this kit version the basecaller automatically filters reads below Q7, which means that 16.9 Gbases of "passed" reads were obtained and the mean read length was about 10K.

![Figure 4.3.1](image.png)

Figure 4.3.1 (a) Histogram of read lengths after log transformation (b) Outcome of assembly strategies

The genome size of the greater amberjack is estimated at 0.74pg ~724 Mbp. The greater amberjack male/female genome published by Sarropoulou et al. 2017 comprises 45,909 contigs and assembled 669Mbp the while greater amberjack male genome published by Araki et al 2018 comprises 41,188contigs, 34,656 scaffolds and assembled 678Mbp (Figure 4.3.1 b).

References


4.4 Atlantic salmon (*Salmo salar* L.) Paternity assignment and traceability tools and genetic marker discovery pipelines (Partners 10 and 12).

Atlantic salmon – an example of a complex recently duplicated genome. A total of 86,468 SNPs were identified from Restriction Site Associated DNA Sequencing (RAD-seq) libraries, and reduced to 1517 following the application of quality control filters and stringent selection criteria. A subsample of SNPs was chosen for the design of high-throughput SNP assays using the Fluidigm EP1 genotyping platform and a training set of known parents and offspring was then used to achieve further filtering. The workflow to identify the panel SNPs is illustrated in Figure 12.

![Figure 12](image)

A panel comprising 94 SNPs balanced across the salmon genome were identified, providing 100% assignment accuracy in known pedigrees. Additionally, the panel was able to assign individuals to one of three farmed salmon populations used in this study with near 100% accuracy (Figure 13).

We conclude that the workflow described is suitable for the design of cost effective parentage assignment and traceability tools for aquaculture species. Full details of the methodology and results have been published in Holman et al (2017).
4.5 King scallop (*Pecten maximus*): Trait datasets for the King scallop (*Pecten maximus*)—an example of a highly polymorphic molluscan genome (Partners 6, 10 and 12)

In contrast to Atlantic salmon and blue mussel, the value of the king scallop, *Pecten maximus*, within the Scottish economy is relatively small, with an annual farmed production for the table of 30-80,000 tonnes since 2009 (Scottish Shellfish Production Survey 2016).

Partners 10 and 12 created a set of database standards for shellfish phenotypic traits in coordination with the European Bioinformatics Institute (EBI). Two sample trait datasets for the King scallop (*Pecten maximus*), consisting of a wide range of trait measurements, were collected and used to establish new data reporting standards for EBI databases. The traits from 500 wild caught scallops from the Western Highlands of Scotland included growth rate, measured from seasonal growth rings, shell dimensions and yield of meat and gonads (Partners 10, 12, 23). Trait standards were modelled after existing (M2B3) standards and are now currently implemented for public use in EBI databases via the Webin system [https://www.ebi.ac.uk/ena/data/view/ERC000038](https://www.ebi.ac.uk/ena/data/view/ERC000038).

The available resources and tools available for selective breeding in molluscs were reviewed prior to prioritizing workflows for parentage assignment and genetic marker discovery in the King scallop as a model species (Hollenbeck and Johnston 2018). To provide a data set for the development of workflows for parentage assignment panels and genetic marker discovery we sequenced double-digest restriction-site associated (ddRAD) libraries from 229 individuals sampled from eight geographic localities in Europe. After stringent filtering for quality, coverage, and minor allele frequency, 2194 high-quality SNP loci were identified for downstream applications such as population genetics, sample traceability, and parentage assignment. High-resolution population genetic analysis identified the existence of a southern and northern stock of King scallop in European waters (Figure 14), which is critical information for founding diverse, domesticated breeding populations in the future. Environmental analysis using the SNP allele frequencies also identified 24
SNPs potentially associated with sea temperature, which will be candidates for functional markers for thermal tolerance in selective breeding programs.

![Principal components analysis of wild King scallops showing a northern stock (SNO, southern Norway; NNO, northern Norway) and southern stock (ESP: Spain; SW: southwest Scotland; SE: southeast Scotland; NW: northwest Scotland; NE: northeast Scotland; SLD: Shetland Islands).](image)

**Figure 14** Principal components analysis of wild King scallops showing a northern stock (SNO, southern Norway; NNO, northern Norway) and southern stock (ESP: Spain; SW: southwest Scotland; SE: southeast Scotland; NW: northwest Scotland; NE: northeast Scotland; SLD: Shetland Islands).

We also made use of the trait data for wild scallops set collected in Task 8.1. Scallops selected on the basis of favourable trait values of growth rate and meat yield were shipped to Scalpro A/S for spawning, larval rearing and on-growing to the spat stage by partner 23. In total 7 single pair crosses were successfully made and spat from each family were on-grown for 15 months at the Scalpro facility in Bergen. Trait values were measured in the offspring for the 7 families produced by partner 23. The results for meat yield are illustrated in Figure 15.

![Distribution of meat yield for offspring from 7 families of King scallop.](image)

**Figure 15** Distribution of meat yield for offspring from 7 families of King scallop.
4.6 References


Lan, J.H., Yin, Y., Reed, E.F., Moua, K., Thomas, K., Zhang, Q., 2015. Impact of Three Illumina Library Construction Methods on GC Bias and HLA Genotype Calling. Hum Immunology 76, 166–175.


Wright, E.S., Vetsigian, K.H., 2016. Quality filtering of Illumina index reads mitigates sample cross-talk. BMC Genomics 17, 876.
5 Measuring complex traits in fin and shellfish

5.1 European sea bass (*Dicentrarchus labrax*): Using a novel phenotyping method combined with genomic information to improve selection of feed conversion ratio in sea bass (Partner 13)

5.1.3 Background
Feed conversion ratio (FCR) is a trait characterizing fish for their ability to convert feed into biomass (individual FCR = individual feed intake / individual weight gain). Improving FCR reduces the amount of feed needed to produce 1 kg of fish. Thus, this trait is very important because feed is the main economic cost of farmers and also the highest contributor of environmental impacts of fish farming (Besson et al. 2017). The issue with FCR is that it cannot be measured on individual fish held in groups (the usual situation in fish farming) because there is no practical way to measure individual feed intake (FI) in such conditions. Thus, without individual phenotypes, we cannot get proper estimate of genetic parameters of FCR (heritability and genetic correlation with other traits). Still, several studies investigated the genetic background of feed efficiency related traits with different approaches (see review by de Verdal et al. 2016).

A first method consists of measuring FCR in tanks exclusively containing fish of the same family. Using separately reared full-sib families, this allows to estimate genetic variability of FCR, and then to proceed to between-family selection. With this method, FI is measured knowing precisely the amount of feed distributed in each tank, and by collecting uneaten pellets (e.g. Kolstad et al. 2005). Family measurements has the main advantage of estimating the genetic variation between families but it has the main issue of being unable to estimate intra-family variation, and to have a certain level of confounding of tank effects and genetic effects, thus resulting in and overestimation of genetic variance.

Estimating intra-family variation can only be done by measuring FI at individual fish scale. Individual FI can be estimated using X-radiography of eaten meals, using X-ray dense markers (generally radio-opaque glass beads) included in pellets (Kause et al. 2016). Just after feeding, the number of radio-opaque beads in the gastrointestinal tract is counted on an individual radiography of each fish, allowing the estimate of FI of each fish. This technique is highly accurate for one meal but FI in fish is known to vary across meals, which could bias the measure of FI. Furthermore, this method does not allow estimating FCR of fish over a significant period of time, because it is impossible to perform it on each meal without seriously perturbing fish welfare and feed intake.

Thus, instead of trying to measure feed intake directly, other methods have been developed to phenotype fish for indirect criteria that could be related to feed intake and easy to measure. One of these indirect criteria is weight loss during fasting. During fasting, fish use the energy stored in their body to cover maintenance costs, so weight loss during fasting is an indicator of maintenance metabolic rate. In theory, therefore, selecting for fish showing lower weight loss during fasting would reduce FCR due to lower maintenance needs. This trait has two main advantages, it is easy to measure and it can be measured on a large number of fish at the same time. (Grima et al. 2010) showed that weight loss during fasting and weight gain during re-feeding were correlated to FCR in sea bass. However, Daulé et al. (2014) did not show a
correlated response on feed efficiency after a divergent selection for fasting tolerance in sea bass, but this result could be explained by technical issues during the measurement of feed intake of fish in groups. Thus, the link between weights loss at fasting and FCR remains uncertain. Another indirect trait is fat content measured on selection candidates with non-invasive ultrasound measurements. In the pig industry, Knap and Wang (2012) reported positive correlations between back fat depth and FCR, meaning that selection for leaner pigs led to an improvement of FCR. This is because the deposition of fat is less efficient in terms of energy used per unit of wet weight gain than the deposition of protein. In fish, fat is mostly deposited as visceral fat, and intramuscular/subcutaneous fat and several studies already showed correlations between fat related traits and FCR (Quinton et al 2007; Kause et al 2016). In 2007, Quillet et al (2007) showed that a trout line selected for low muscle lipid content was more efficient than a line selected for high muscle lipid content.

Although these indirect criteria seem promising for the genetic improvement of individual FCR, breeders are still lacking of accurate genetic parameters for FCR and the real relationship between indirect criteria and individual FCR is still missing. Consequently, FCR is not included in breeding programs and its improvement relies on the implicit assumption that there is a negative genetic correlation between FCR and growth (i.e., that faster growing fish also have a lower – better -- FCR). Measuring individual feed intake directly on individual fish would be the best solution in order to estimate accurate genetic parameters and to develop efficient breeding programs. Thus, we developed a novel method to estimate feed intake of individual fish in aquaria over a reasonably long period of time. The estimation of feed intake in these conditions would enable the calculation of an individual FCR that could then be related directly to FCR of fish in groups. In addition, we chose to feed the fish with a restricted ration because, following the results of a study in rabbit (Drouilhet et al. 2016), selecting faster growing animal under restricted feeding improves FCR as a correlated response. This is because in restricted feeding condition, the animals that grow faster are de facto the most efficient.

However, this measurement and can only be made on a (relatively) small number of animals and thus cannot be made on all selection candidates (only on sibs). We genotyped the individuals in order to test whether genomic information could enhance the prediction of the genetic value of selection candidates. Genomic selection (GS) based on SNP marker can be used to estimate genomic estimated breeding values (GEBVs). In GS, sibs of the selection candidates with both phenotype and genotype are used to estimate marker effects, and the prediction equations are later used to predict GEBVs for the selection candidates that are only genotyped. In aquaculture, several studies have shown the higher performances of GS in terms of genetic gain, accuracy of selection for traits such as, growth (Tsai et al. 2017) or disease resistance (Bangera et al 2017; Vallejo et al 2017).

Finally, the second part of the project consisted of measuring FCR of groups of fish made depending on individual FCR in aquarium to test whether individual FCR could predict performance in group.

5.1.4 Material and Method

*Origins of the fish*

This study was performed on the European sea bass (*Dicentrachus labrax* L.), a fish of major economic importance for aquaculture in the Mediterranean area. *Generation 0*: The animals of G0 were caught from the wild in the Gulf of Lions (West Mediterranean).
Generation 1: 41 sires and 8 dams randomly chosen from G0 were mated in a full factorial mating design to create the G1 generation (Grima et al. 2010).

Generation 2: 20 sires and 5 dams selected from G1 were selected to create the G2 generation (Daulé et al. 2014).

Generation 3: Two females from G1 were mated with 30 males from G2 in a full factorial mating design. Thus, the pedigree of these G3 fish is known back to their wild great grand-parents. The G3 individuals were phenotyped for their individual FCR in aquaria. The figure 16 shows the pedigree of these fish.

![Pedigree of fish generations](image)

**Figure 16** Origin of the fish.

Phenotyping fish for feed conversion ratio in aquariums

Two hundred 10 l aquaria were provided 5 l/h of 20-21 °C seawater by a recirculating aquaculture system (Figure 17).

![Aquarium facilities](image)

**Figure 17** Aquarium facilities at the aquaculture research station of Ifremer in Palavas-les-Flots.

European sea bass (25g on average) were first reared in groups of five in each aquaria, to enable acclimation of the fish to the new environment. After 14 days, they were randomly distributed to individual aquaria. After 14 days of acclimation to rearing in isolation, the fish were weighed (Body weight, BW) in a “go, no go” biometrics. The fish that lost weight during this first period were removed considering they did not adapt
well to this new environment. The remaining ones were kept in the aquaria for two more periods of 14 days. In total, a “successful” fish stayed 56 days in the system and was weighed 5 times (Figure 18).

![Experimental scheme.](image)

**Figure 18** Experimental scheme.

Individual BW at each biometrics was used to estimate individual feeding ration for the following period. This ration (1.3% BW/day) was half the standard ration (2.6% BW/day for this size of fish) given by the feed manufacturer. We choose to feed fish with restricted ration because, in this condition, the fish with the highest relative weight gain would also be the fish with the best FCR. Fish were fed automatically once a day in the morning (feeders are on pictures in Figure 17). Every afternoon, the number of uneaten pellets was counted in each aquarium and converted to grams (1 pellet ≈ 0.00925 g). Thus, for each period of individual phenotyping in aquariums, data were available for individual weight gain and individual feed intake. From weight gain and feed intake, feed conversion ratio (FCR) and daily growth coefficient (DGC) over period 2 and 3 were calculated as:

\[
FCR = \frac{\text{final} \_ \text{BW}_3 - \text{initial} \_ \text{BW}_2}{\text{feed} \_ \text{intake}_2 + \text{feed} \_ \text{intake}_3}
\]

\[
DGC = \frac{\text{final} \_ \text{BW}_3^{1/3} - \text{initial} \_ \text{BW}_2^{1/3}}{D_{2-3}}
\]

Where \text{initial} \_ \text{BW}_2 is the weight of the fish at the beginning of period 2 and \text{final} \_ \text{BW}_3 is the weight of the fish at the end of period 3. Where \text{feed} \_ \text{intake}_2 and \text{feed} \_ \text{intake}_3 are the feed intake of the fish in periods 2 and 3. \text{D}_{2-3} is the cumulated number of days of period 2 and 3 (28 days in total). We did not consider period 1 because this period was an acclimation period and not all data of feed intake were available.

**Feed conversion ratio of groups**

To test a potential link between individual feed efficiency and group feed efficiency, the 588 fish phenotyped in aquariums were split in groups according to their individual performances in aquariums as follow:

- First, 6 groups of 98 fish were constituted based on relative feed intake (cumulated feed intake divided by initial body weight at period 2) (see Figure 19).
- Second, in each group, the 98 individuals were split in 2 sub-groups of 49 individuals based on their relative weight gain (cumulated weight gain divided by initial body weight at period 2) (see Figure 19).
Thus, 12 groups of individually tested fish were made. We chose this method to make groups because we wanted to test if fish that grow faster in aquariums at a given relative feed intake were the most efficient fish in groups. In addition, we made 4 more groups of 44 fish with the fish that lost weight during the first period alone in aquariums. These groups were made to test if losing weight in aquariums could be linked to group feed efficiency performances.

Figure 19 Division and subdivision of groups of fish individually phenotyped for relative feed intake and relative weight gain in aquariums.

Fish were stocked in 16 tanks of 2m² covered by plastic sheeting to avoid disturbance. They were fed once a day ad libitum using an automatic feeder delivering 20 portions over 6h and 20 minutes with decreasing frequency. The feeders were filled with a known amount of pellets. Uneaten pellets were collected in the faecal trap of each tank. Every day, at the end of the feeding period, each faecal trap was checked. If pellets were found, it meant that fish of the tank reached ad libitum. If a faecal trap was empty or that only few pellets were present, an additional portion was given to the tank by activating the feeder manually. Additional portions were then given every 30 minutes until pellets would go through to the faecal trap meaning that ad libitum was reached. Uneaten pellets of all tanks were then collected, photographed and counted using ImageJ. Daily group feed intake was then estimated for 3 periods of 3 weeks. In addition, all fish were weighed at the end of each period in order to estimate the weight gain of the groups. Thus, with the weight gain of the fish and the feed intake of the tanks we estimated the FCR of each tank for the 3 periods. For data analysis, however, we used only data of periods 2 and 3 because during period 1 we did not distribute additional rations to the tanks that ate the entire ration. Consequently, during period 1, fish were not fed completely ad libitum for several days.

Genotyping
The 48 grand parents from G1 and the 49 grand grand parents from G0 were genotyped with an iSelect Custom Infinium Illumina® European sea bass array of 2,722 SNP. Then, 400 out of 588 individually phenotyped fish and their 32 parents (2 females and 30 males) were genotyped a similar array of 3,987 SNP. This second
array uses the same markers as the original 2,722 SNP array plus 1,265 duplicated markers that were ineffective on the original array.

These 3,987 SNP were chosen from a genome-wide variation map containing 2,628,725 SNPs phased into chromosome-wide haplotypes. These SNPs were discovered from 14 wild individuals using whole-genome sequencing as described by (Duranton et al. 2017). The selection of the 3,987 SNPs from these 2,628,725 initial SNPs was first done by removing variants closer than 80bp from another known variant. Then, all A/T and C/G variants were removed following the recommendation of Illumina. Finally, the SNPs were chosen to cover all the chromosomes but with a variable SNP sampling effort depending on the local nucleotide diversity (π), as reported by Tine et al (2014). Since nucleotide diversity is negatively correlated to the local recombination rate in sea bass, this local adjustment in the density of SNPs aimed at homogenizing the density of markers along the recombination map instead of the physical map.

Once the animals were genotyped, the first step to create the SNP dataset used in our genomic analysis was to apply classic quality control ignoring all SNPs with a MAF inferior to 5% and a call rate inferior to 90% in the G3 animals. A MAF inferior to 5% and a call rate inferior to 90% could indicate problems during the genotyping. This quality control resulted in keeping 2,100 SNPs for G3 individuals and their parents. From this 2,100 SNPs, we discarded the original version of all duplicated markers, which resulted in keeping 1,923 SNPs markers for G3 individuals and their parents. Then, we kept only the markers that were in common between both chips, representing 1,110 SNPs.

Finally, we discarded all animals for which the call rate (number of SNPs genotyped over the number of SNPs on the array) was lower than 90% indicating potential quality issue of the DNA sample. This resulted in keeping 5 individuals of G0 (out of 49), 52 individuals of G1 (out of 52), 30 fish of G2 (out of 30) and 398 fish of G3 (out of 400). The genotyping of individuals had 2 purposes:

1) Retrieving pedigree. SNP markers obey to Mendelian transmission rules so they can be used to find the parents of individuals. We used VITASSIGN, exclusion-based parentage assignment software described by (Vandeputte et al 2006) and adapted to SNP markers, to retrieve the pedigree of the 400 fish. 398 animals of G3 were successfully assigned to their parents. Individual of previous generations were assign in earlier projects using microsatellite markers.

2) Genomic evaluation. Genetic progress is based on the calculation of estimated breeding value (EBV) of individuals which is a prediction of the ability of the parents to breed the most efficient offspring. EBVs can be predicted using the realized genomic relationship matrix (drawn based on the animals’ genotypes) combined with best linear unbiased prediction (BLUP) in a method called GBLUP. The genomic relationship matrix establishes relationship between individuals based on the SNPs they are sharing. In theory, the relationship estimated by the genomic relationship matrix is more accurate than the one estimated by genealogy. Thus, the accuracy of GEBV is, in theory, higher than EBV using the classic BLUP method without genotypes.

Breeding value estimation
Variance components and estimated breeding values for both traits (log(FCR) and log(DGC)) were computed based on multivariate linear mixed animal model. All the fish from G0, G1, G2 and G3 were included in the pedigree to integrate the selection
process realized on these fish. The models were then fitted by restricted maximum likelihood in AIREMLF90 with 1) a classic pedigree based BLUP or 2) with a single-step GBLUP using the genomic relationship matrix:

1) Pedigree based BLUP - PBLUP
The conventional pedigree-based variance components and EBV were estimated using PBLUP:

$$y_{ijkl} = \mu + \text{Batch}_j + \text{Rack}_k + \text{Column}_l + \text{animal}_i + \epsilon_{ijkl}$$

Where $y_{ijkl}$ is the observed trait (log(FCR) or log(DGC)) of individual $i$ from Batch $j$, phenotyped in Rack $k$ in column $l$. We used log transformed FCR and DGC in order to linearize the relationships between the traits. $\mu$ is the overall mean, Batch $j$ is the fixed effect of the batch in which the fish has been phenotyped in aquariums (1 to 10). Rack $k$ and Column $l$ are the fixed effect of the physical position of the aquarium in which the individual fish have been phenotyped. There were 4 racks of 50 tanks in 5 lines and 10 columns. Accounting for fixed effects is important when estimating genetic parameters. We used boot.stepAIC function in R to find out which fixed effects had to be included in our statistical model for our phenotypes observed. The boot.stepAIC function looks for the model with the lowest Akaike information criterion (AIC). This AIC is an estimator of the quality of each model relative to each of the other models by estimating the trade-off between the goodness of fit of the model and the simplicity of the model. Finally, animal $i$ is the additive genetic effect of animal $i$ and $\epsilon_{ijkl}$ is the random residual.

2) Single step genomic BLUP - SS-GBLUP
The SNP based variance components and GEBV were estimated using single-step GBLUP, a procedure calculating GEBVs based on combined pedigree, genomic and phenotypic information using BLUPF90 programs. In a single step GBLUP, the EBVs of fish not genotyped are calculated based on the pedigree relationship matrix (A matrix) while the EBVs of fish with genotypes are calculated based the genomic relationship matrix described by VanRaden (2008) (G matrix). All other parameters and details of the analysis for log(FCR) and log(DGC) are the same as PBLUP.

Cross validation scheme (CSV) to test predictive abilities
We tested the abilities of the different models described above (PBLUP and SS-GBLUP) to predict the (G) EBV of log(DGC) depending on the number of fish phenotyped. This was assessed using a cross validation scheme. The fish phenotyped and genotyped (400 fish) were randomly split in two groups; a given number of fish was set as the training group and the remaining fish were set as validation group. In order to test the predictive ability of our models depending on the number of fish phenotyped and genotyped, we tested a decreasing number of individual in the training group from 360 to 280, 200,120 and 40. For each number tested, the EBV or the GEBV for the log(DGC) of all fish were predicted while masking the phenotype of the validation group (phenotype set missing). The predictive ability of the model (PBLUP or SS-GBLUP) was calculated as the squared correlation between the averaged (G)EBV and the phenotypes corrected for fixed effects ($r_{EBV,y}^2$). We calculated the slope of this regression to estimate the bias of the models (Echeverri et al 2014) and we calculated the Spearman rank correlation to estimate the degree of similarity between the rankings of fish based their phenotypes or their (G)EBV. To reduce to stochastic effects we replicated the cross validation scheme 30 times with different individuals in training and validation group at each time. Then, at each
repetition, the predictive ability of the model (PBLUP or SS-GBLUP) was calculated for each of the five sizes of training population (40, 120, 280 and 360 fish). The average $r^2$ was used to estimate the reliability of PBLUP and SS-GBLUP models using the following formula as in Bangera et al. (2017):

$$R_{EBV,BV}^2 = \frac{r_{EBV,y}^2}{h^2}$$

Where, $r_{EBV,y}^2$ is the squared correlation between the predicted (G)EBV for log(DGC) for all the fish in the validation group and the recorded log(DGC), corrected for fixed effects, and $h^2$ is the heritability of log(DGC) estimated using PBLUP including all fish with phenotypes.

5.1.5 Results

*Individual phenotypes*

Over a period of 6 months, 831 fish were tested, 185 fish did not pass the “go, no go” biometrics, which means that 646 fish (77.73 %) gained weight after the first period in isolation and were then evaluated for individual FCR for 3 periods. For those 646 fish, we calculated their cumulated FCR using their cumulated weight gain and cumulated feed intake over periods 2 and 3. We excluded fish with aberrant performances: 6 fish with negative cumulated FCR and 52 fish with cumulated FCR higher than 2.6. Applying this threshold, we could keep 588 fish with an average FCR of 1.45 and a coefficient of variation of 0.21 (Figure 20). This coefficient of variation was close to what has been observed in Nile tilapia in a different setting for evaluating individual FCR (de Verdal et al 2017).

![Figure 20](image)

**Figure 20** Histogram of the distribution of cumulated feed conversion ratio for 588 fish.

Furthermore, Figure 21 shows the regression of log(FCR) on log(DGC). The phenotypic correlation between the two traits was 0.87.
Figure 21 Regression of log(FCR) on log(DGC) for the 588 fish with valid phenotypes of individual FCR in aquaria.

Group FCR
The results of the experience in group showed that the tanks composed of fish with lower individual FCR (more efficient fish) were more efficient than the paired tanks composed of fish with higher individual FCR (Table 5). The result was shown significant using a paired t-test ($p = 0.012$).

Table 5. FCR over period 2 and 3 of paired tanks, one composed of fish with low individual FCR (more efficient fish) and one composed of fish with high individual FCR.

<table>
<thead>
<tr>
<th>Group based on similar relative individual feed intake</th>
<th>Low individual FCR</th>
<th>High individual FCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.23</td>
<td>1.24</td>
</tr>
<tr>
<td>2</td>
<td>1.19</td>
<td>1.22</td>
</tr>
<tr>
<td>3</td>
<td>1.20</td>
<td>1.25</td>
</tr>
<tr>
<td>4</td>
<td>1.23</td>
<td>1.24</td>
</tr>
<tr>
<td>5</td>
<td>1.24</td>
<td>1.26</td>
</tr>
<tr>
<td>6</td>
<td>1.19</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Genetic parameters
The two traits showed moderate heritability and strong genetic correlations when using PBLUP (Table 6). With SS-GBLUP, the heritability of both traits increased (Table 7). There was also a strong negative genetic correlation between log(FCR) and log(DGC), meaning that the fish with the best feed efficiency were the fish that grew faster in aquariums. This correlation was similar between PBLUP and SS-GBLUP (0.96 and 0.98) but the standard errors were smaller using SS-GBLUP.

Table 6. Heritability of log(FCR) and log(DGC) on the diagonal and genetic correlations above the diagonal calculated using PBLUP. Standard errors are between brackets.

<table>
<thead>
<tr>
<th></th>
<th>log(FCR)</th>
<th>log(DGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(FCR)</td>
<td>0.22 (0.07)</td>
<td>-0.96 (0.20)</td>
</tr>
<tr>
<td>log(DGC)</td>
<td>0.41 (0.12)</td>
<td></td>
</tr>
</tbody>
</table>
Table 7. Heritability of log(FCR) and log(DGC) on the diagonal and genetic correlations above the diagonal calculated using single-step GBLUP. Standard errors are between brackets.

<table>
<thead>
<tr>
<th></th>
<th>log(FCR)</th>
<th>log(DGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(FCR)</td>
<td>0.40 (0.07)</td>
<td>-0.98 (0.01)</td>
</tr>
<tr>
<td>log(DGC)</td>
<td>0.73 (0.06)</td>
<td></td>
</tr>
</tbody>
</table>

**Predictive abilities: PBLUP vs SS-GBLUP**

The predictive ability, based on $r^2$ and spearman rank correlation, of the PBLUP model increased with an increasing number of animal used as training population (Table 8). Similarly, the predictive ability of the SS-GBLUP model also increased with an increasing number of animal used as training (Table 9). Nevertheless, the slope of the regression is steeper with SS-GBLUP than with PBLUP, meaning that including more individuals in the training group increased the predictive ability of SS-GBLUP more than the predictive ability of PBLUP (Figure 22). Furthermore, the $r^2$ and the spearman rank correlation are both higher for SS-GBLUP compared to PBLUP at the same number of animals in the training group. These results show the higher predictive ability of SS-GBLUP compared to PBLUP.

Table 8. Results of the cross validation scheme of PBLUP model for different numbers of fish in the training group.

<table>
<thead>
<tr>
<th>Number of animals in training group</th>
<th>$r^2$</th>
<th>Bias</th>
<th>Spearman rank correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>360</td>
<td>0.073 (± 0.004)</td>
<td>0.086 (± 0.003)</td>
<td>0.20 (± 0.008)</td>
</tr>
<tr>
<td>280</td>
<td>0.054 (± 0.002)</td>
<td>0.081 (± 0.001)</td>
<td>0.20 (± 0.004)</td>
</tr>
<tr>
<td>200</td>
<td>0.047 (± 0.001)</td>
<td>0.076 (± 0.001)</td>
<td>0.20 (± 0.003)</td>
</tr>
<tr>
<td>120</td>
<td>0.039 (± 0.001)</td>
<td>0.034 (± 0.001)</td>
<td>0.19 (± 0.002)</td>
</tr>
<tr>
<td>40</td>
<td>0.032 (± 0.001)</td>
<td>0.026 (± 0.001)</td>
<td>0.19 (± 0.005)</td>
</tr>
</tbody>
</table>

Table 9. Results of the cross validation scheme of SS-GBLUP model for different numbers of fish in the training group.

<table>
<thead>
<tr>
<th>Number of animals in training group</th>
<th>$r^2$</th>
<th>Bias</th>
<th>Spearman rank correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>360</td>
<td>0.14 (± 0.005)</td>
<td>0.27 (± 0.005)</td>
<td>0.36 (± 0.007)</td>
</tr>
<tr>
<td>280</td>
<td>0.12 (± 0.003)</td>
<td>0.23 (± 0.003)</td>
<td>0.33 (± 0.004)</td>
</tr>
<tr>
<td>200</td>
<td>0.10 (± 0.002)</td>
<td>0.20 (± 0.002)</td>
<td>0.31 (± 0.003)</td>
</tr>
<tr>
<td>120</td>
<td>0.07 (± 0.001)</td>
<td>0.14 (± 0.002)</td>
<td>0.26 (± 0.002)</td>
</tr>
<tr>
<td>40</td>
<td>0.03 (± 0.001)</td>
<td>0.04 (± 0.001)</td>
<td>0.20 (± 0.002)</td>
</tr>
</tbody>
</table>
Figure 22: Predictive ability of PBLUP and SS-GBLUP models calculated from cross validation scheme as a function of the number of animals in the training group.

5.1.6 Conclusions and recommendations
This study is the first to present a method to evaluate individual feed intake and FCR of a large number of European sea bass. With this new method, we found phenotypic variability in FCR and also in DGC. The phenotypic coefficient of variation for FCR was close to that observed in Nile tilapia, also phenotyped individually, but by video observation over a period of 10 days (de Verdal et al. 2017). Thus, we have made an important step towards a better understanding of the individual variation of feed efficiency in European sea bass. However, our phenotyping method requires a special procedure in order to obtain meaningful estimates of FCR:

- First, it is important to phenotype juveniles of about 25g, first because we could observe in preliminary trials that that bigger fish will not acclimatize properly and will not feed to the expected level. When fish are larger, the ration is also larger, and any suboptimal feed intake results in a larger number of pellets wasted compared to a smaller fish at the same relative intake, and this could induce errors in the daily counting the number of uneaten pellets.

- Second, acclimation must be done step by step. Fish must be initially placed by groups of 5 to “discover” the new environment in groups for a period of 2 weeks. Then, fish can be split alone for another period of 2 weeks of acclimation to individual rearing. Despite this acclimation procedure, some fish might still not be acclimatized to the aquaria and they will not eat properly. These fish that are losing weight and must be removed from the aquarium.

- Another important point is to set up an automatic feeding to avoid any disturbance during feeding, and to enable feeding 200 animals in 200 tanks at the same time of the day.
Feeding must also be restricted, for 2 main reasons. First, with 50% restricted feeding, about 30% of the fish eat the whole ration, which greatly simplifies the work of estimating feed intake, as for these fish no uneaten pellets need to be counted. Second, when feed is restricted the faster growing fish are the most efficient fish.

We could confirm both the phenotypic and the genetic link of growth (DGC) under restricted feeding with individual FCR. Indeed, the genetic analysis showed that log(FCR) and log(DGC) measured in aquaria are both heritable and also strongly genetically correlated (genetic correlation close to unity). This result was true for both PBLUP using pedigree information and SS-GBLUP using pedigree information combined with genomic information. A practical consequence of this is that the counting of uneaten pellets could be avoided and FCR could be estimated on based on DGC under restricted feeding in individual aquarium conditions. This has a very important impact on the workload necessary to perform a proper estimate of a selection criterion for FCR. Our results suggest also that the response to selection in FCR would be larger when selecting on DGC rather than selecting on FCR due to the higher heritability of DGC (0.73 > 0.40) and the close to unity genetic correlation between them. This is an encouraging result towards possible genetic improvement of FCR in fish because, compared to X-ray analysis, phenotyping individuals in aquaria allows to estimate FCR over a continuous period of several weeks with a very precise estimate of feed intake.

Nevertheless, even though the results are promising, one could argue that individual FCR in aquarium remains an indirect criterion for FCR of a fish in a group. Indeed, when measuring feed intake of fish alone in aquaria, social interactions are ignored which could bias the feed intake of the fish. For instance, fish that are dominated in a group could benefit of not having social competition in aquaria to eat more and potentially affect their FCR. This would result in bias in the estimation of FCR. This is the reason why we tested if groups of individually more efficient fish would perform better than groups of individually less efficient fish. The results of this group testing are positive and fish that are individually more efficient are also more efficient in groups. Nevertheless, the differences observed between groups are small, which confirms that individual FCR measured in aquaria is not a perfect predictor of FCR in groups. Because the differences between groups were small, we needed a reliable protocol to be able to detect those differences:

- An important point of group measurement of FCR is to feed the groups ad libitum. With ad libitum feeding, fish are placed in conditions close to those in commercial aquaculture. During the first period of group measurements we did not feed ad libitum, which resulted in large differences between groups. One reason explaining the effect of not feeding ad libitum could be that social competitions between fish are enhanced causing some fish to eat less than their needs, which could affect their FCR and the FCR of their group.
- One consequence of feeding ad libitum is that some pellets are wasted. Thus, in order to not underestimate FCR, these wasted feed pellets need to be collected and counted precisely. With the number of tanks and the number of fish in tanks in our experiment, the number of pellets wasted per day was substantial, which could lead to errors when counting manually. To overpass this issue, we developed a special procedure where wasted pellets were photographed and where pellets were counted using ImageJ. This procedure has 2 benefits: first, the counting was more precise and second, pictures were saved and stored which allowed us to re-analyse them if needed.
We successfully used the genomic information to trace the pedigree of our 400 fish. Having the correct pedigree is crucial to estimate correct genetic parameters with PBLUP and single-step GBLUP models. Using SNP markers to retrieve the pedigree of fish is easier and very efficient even with only a small number of SNPs (Holman et al 2017). In our case, we used the 1,100 SNPs obtained after filtering on polymorphism, on minor allele frequency lower than 5% and on call rate lower than 90%. In our project, we could assign 398 fish out of 400 of G3 to their parents. Only two fish could not be assigned to any couple of parents, which was probably due to a poor quality of the DNA samples.

Furthermore, as expected, the genomic information reduced the standard error of the genetic parameters with SS - GBLUP compared to PBLUP. SNP data also improved the predictive abilities of our model compared to a traditional PBLUP model. These results are important because phenotyping fish in individual aquariums is long and tedious, thus pushing towards a low number of fish phenotyped. With a low number of fish phenotyped, the accuracy of genetic parameters is reduced and the predictive ability of the model is lowered (Figure 22). This means that, in a commercial breeding program, estimating the EBV of selection candidates would not be fully accurate. However, the consequences of this low number of fish phenotyped can be mitigated by genotyping the fish, which would increase the accuracy of genetic parameters and increase the predictive ability of the model. Thus, the EBVs of selection candidates will be more accurate using genomic information with a SS - GBLUP model. In addition, the results showed that the predictive ability of SS-GBLUP model could be even higher if more fish were phenotyped and genotyped. However, we do not know the ideal number of individuals needed to be phenotyped and genotyped to reach the highest predictive ability possible, as with the numbers tested no plateau was reached for $r^2$. We also do not know if increasing the number of SNPs used in the SS-GBLUP model would increase the predictive ability. In our case, we restricted the number of SNP used for the genomic analysis to 1,110. This was because we chose to work with SNP that were in common between two arrays that were used to genotype our fish. Therefore, we would recommend optimizing the design of the arrays to obtain a maximum number of informative SNPs. Nevertheless, the positive point is that even with this limited number of SNPs, we could show that genomic information significantly increased the predictive ability of the models used (+90% when using 360 animals in the training group). Hence, these results confirm that genotyping fish is interesting to enhance the estimation of breeding values, especially for traits that are difficult to measure.

5.1.7 References


5.2 Atlantic salmon: Developing experimental disease models underpinning investigation of the genetic basis disease resistance (Partner 6)

5.2.1 Background
Selective breeding for particular performance trait is fast developing field supporting sustainability and profitability of finfish aquaculture worldwide. Family based selection for improved growth performance, condition index are the most commonly applied schemes however individual based selection or selection for other desirable trait such as flesh quality or disease resistance becoming more popular in the European aquaculture (Chavanne et al 2016). Diseases leading to production losses due to poor quality or increased mortality represent one of the major concerns for sustainability of Atlantic salmon aquaculture. Recent advances in genomics facilitated discovery of genetic basis for fish resistance and susceptibility to economically important viral, bacterial and parasitic diseases (e.g Gilbey et al 2006; Houston et al 2008; Ozaki et al 2013; Rodríguez-Ramilo et al 2011). To date majority of disease resistance is under polygenic control however single major quantitative trait loci (QTLs) are known to explain over 80% of genetic variation for a such complex trait as disease resistance (Moen et al 2009; Baerwald et al 2011). The best described example is the major QTL for Atlantic salmon resistance to infectious pancreatic necrosis (IPNV) and subsequent commercially successful marker-assisted breeding programmes which indicate 75% reduction in IPN outbreaks in Norway since the introduction of QTL resistant fish (Houston et al 2012; Moen et al 2015).

Fish/family disease resistance/susceptibility can be measured either directly as a rate of survival, where surviving fish/family are utilized for genetic marker discovery or indirectly by selecting individuals which exhibit desirable levels of immunological parameters such as cortisol levels (Weber et al 2008) or cytokine in response to selected pathogen (Wang and Secombes 2013). Both approaches for selection of hosts directly from fish farms undergoing mortality event comes with added complications as disease outbreaks are often complex events involving multiple pathogen, additional stress due to adverse abiotic conditions such as storm and/or negative impact of treatment procedures. Therefore selection of survivals from disease infectious trials carried out in the highly controlled aquarium environment are the preferred route to obtain appropriate genetic material for genetic marker discovery (Gonen et al 2015; Houston et al 2008; Moen et al 2009; Vallejo et al 2014). The ultimate goal for successful selective breeding is to ensure that the selected trait is heritable and high heritability estimates for resistance to a range of bacterial and viral diseases were found for several salmonids (Kjøglum et al 2008; Olesen et al 2007; Perry et al 2004; Vallejo et al 2010).

Experimental disease models underpinning investigation of the genetic basis for susceptibility/resistance to the amoebic gill disease, economically significant disease of Atlantic salmon, are described in multiple studies (Collins et al 2017; Morrison et al 2004; Taylor et al 2007). Three different approaches, gross gill score, image gill score and histopathological changes to fish gill tissue, can be taken to describe level of amoeba infection or fish susceptibility to AGD, Positive correlation between gross and histology scores are often reported (Collins et al 2017; Downes et al 2015) however some discrepancies can occur when DNA based methods such as real time PCR are used as diagnostic tools to detect presence of the amoebas
especially for fish with low gill scores (Collins et al 2017). In terms of the administration of infectious dose of *P. perurans* to Atlantic salmon, immersion bath in suspension of amoebas ranging between 10 to 5000 amoeba L\(^{-1}\) is being frequently used as an infectious dose (Collins et al 2017; Haugland et al 2016; Taylor et al 2007) and more recently a cohabitation of naïve fish with the infected “seeders” in ration of 15% of infected to naïve salmon has been shown effective to transmit the disease (Robledo et al 2018). From the published studied, it appears that there is genetic component in susceptibility/resistance of Atlantic salmon to AGD with the heritability estimates ranging from 0.16 to 0.48 (Taylor et al 2007, 2009). As for other parasitic diseases (Gilbey et al 2006; Correa et al 2017), the architecture of AGD appears to be polygenic and two regions of chromosome 18 showed an association with both gross gill score and amoeba load measured by real time PCR (Robledo et al 2018).

5.2.2 Aims of the study

First aim of the study was to develop an AGD challenge model, to allow estimation of minimum number of amoeba/L to achieve a median infective dose (ED50) for reproduction of the disease. Fish families, exhibiting a various levels of AGD resistance, provided by USTAN/Xelect partners (Partners 10 and 12), will be challenged with *P. perurans* and phenotyping data such as gill scores associated with AGD will recorded to investigate genetic variance of the AGD resistance in Atlantic salmon.

The second aim was to assess genetic variation of *P. perurans* strains collected from different European countries in addition to samples from Australia using well described cytochrome oxidase I (COI) gene marker. Ability to characterize *P. perurans* strains on fish farms will provide vital information for disease management strategies, in terms of persistence of particular strains, and provide an insight into origin of the parasite in European aquaculture.

5.2.3 Material and Methods

Preparation of *P. perurans* culture for in vivo challenge

Polyclonal in vitro cultures of *P. perurans*, originally isolated from gills of infected farmed Atlantic salmon exhibiting signs of AGD in November 2012, were used in the experiment. Presence of the parasite was confirmed originally by sequencing of 18S ribosomal subunit (Young et al 2008) and culture virulence was later established in two published studies (Collins et al 2017; McCarthy et al 2015). Cultures were maintained as described in McCarthy et al (2015) and one day prior to the challenge two separate 'family' lines (H4C* and H4OY, at passage numbers 50 and 46 respectively) were pooled and counted using both a haemocytometer (2.1 ± 0.42 x 10\(^4\) cells ml\(^{-1}\)) and a Sedgwick Rafter counting chamber (4.0 ± 0.36 x 10\(^4\) cells ml\(^{-1}\)). As the Sedgwick Rafter chamber allows counts of the number of amoebae which have settled out of the water column, in a 1 ml volume without constriction, this suggests it is likely to be the more accurate method. *P. perurans* culture was diluted in sterile seawater, chilled to 4°C to reduce cell division, and nine experimental dilutions of cells were made to give final concentration of 10, 20, 40, 80, 160, 320, 640, 1280 and 5000 cells/L in 350L flow through experimental fish tank (1.1m in diameter).

Atlantic salmon in vivo challenge and sample collection

Seawater used to supply experimental tanks was filtered and UV treated. The population of Atlantic salmon (20 fish) selected for the experimental challenge was screened for presence of common viral and bacterial diseases using relevant real time
PCR assays. Two weeks prior to the challenge Atlantic salmon (132 ± 18g) were transferred into 350 L seawater tank kept at 10°C in groups of 10 fish per individual tank. Fish were handled in accordance with the Animals (under Scientific Procedures) Act 1986 and the experimental design was subjected to the internal statistical review to allow for statistical robustness of the challenge results. Nine groups of fish were exposed to the various amoebas concentrations (see above) and fish in one tank were mock-exposed to a sterile culture media and kept as a negative control. Fish were fed daily to 1% body weight with Skretting Atlantic N 500 + 50 PAX diet. Prior to the exposure, the seawater volume in all tanks was lowered to 150 L and kept static with aeration. *P. perurans* cultures were dispersed in tanks and fish exposure was carried out for 4 hours. During this period, fish were observed regularly for signs of distress. Following the exposure period, the water flow was resumed and volume of water was increased to 350 L for the remaining period of the challenge. Fish in negative treatment tank were subjected to the regime as the treatment groups. Experiment was carried out for 28 days following *P. perurans* exposure and fish were observed twice a day to monitor welfare and changes in behaviour attributed to the disease. All fish were euthanized by exposure to 200 mg/L tricaine methanesulfonate MS 222 (Sigma) at the terminal point.

To reduce excessive bleeding into gills which make lesion scoring difficult, fish were bled by a caudal venous puncture, followed by removing heart. All gill arches on right side of the fish were removed and fixed with 10% neutral buffered formalin (NBF) for gill scoring analysis. Third gill arch from left side of the fish were fixed with 100% ethanol for real time PCR analysis. Additional samples of head kidney, liver and spleen were taken into RNAlater (Sigma) and placed overnight at +4°C. NBF samples were kept in room temperature prior to analysis and all tissue samples were stored in -80°C prior further processing.

**Gross scoring of lesions associated with amoebic gill disease**

NBF fixed gill arches from the right side of each fish were placed, facing dorsal surface, into a Petri dish filled with sterile seawater and examined using light microscope (Olympus BX 50) at x 40 and x 100 magnification. A gill score from 0 to 5 was given to each gill arch based on the presence of white patches and proportion of affected area to the total surface area of gill arch according to the system developed by Taylor et al (2009). Gross scoring was performed blind, without previous knowledge of *P. perurans* infection dose.

**DNA-based detection of *P. perurans* using real time PCR (qPCR)**

For each gill tissue sample, two sets of qPCR was performed, first for detection of *P. perurans* using the *P. perurans* 18S rRNA gene (Fringuelli et al. 2012) and second for Atlantic salmon housekeeping gene, elongation factor 1 α (Bruno et al 2007). qPCR reaction contained 2X PerfeCTa qPCR ToughMix UNG (Quanta Biosciences, VWR UK) using LightCycler 480 (Roche). Each sample was run in triplicates and no template controls (NTC) were included for each plate. The upper limit of detection for the *P. perurans* qPCR detection was set to 35 Cp and detection over this threshold were considered as negative. *P. perurans* equivalent target amount (ETA) was normalized to the amount of host ETA detected in gill samples.

**Genomic DNA extraction from *P. perurans* cultures and infected fish gills and PCR amplification**
Total genomic DNA was extracted from a variety of material such as infected fish tissue, swabs, polyclonal and monoclonal cultures of *P. perurans* using Qiagen DNA extraction kits (DNeasy Blood and Tissue kit, QIAamp DNA mini Qiacube kit, Qiagen). Thirty two samples of *P. perurans* were obtained from Ireland, Norway, Scotland and Tasmania and three cultures of *P. pemaquidensis* were obtained from the Culture Collection of Algae and Protozoa (Oban, Scotland) and the Institute of Parasitology, Czech Academy of Sciences (Ceske Budejovice, Czech Republic). For amplification of the cytochrome oxidase I gene (COI) gene, two sets of universal primers LCO1490F and HCO2198R (Folmer et al., 1994) and Eucox 1F and Euglycox 1R (Heger et al. 2011) were used respectively. Partial fragment of the Internal Transcribed Spacer (ITS) rDNA region was amplified using primers designed from the sequence DQ660492 (Caraguel et al 2007). Number of *P. perurans* COI haplotypes were identified using ElimDupes tool from https://hcv.lanl.gov/content/sequence/ELIMDUPES/elimdupes.html.

**Phylogenetic analysis of ITS for *P. perurans* strains**

A final dataset of 30 sequences of *P. perurans*, *P. pemaquidensis* and *P. branchiphila* spanning the partial 475bp fragment of COI gene was aligned using MAFFT v.7 (Katoh & Standley, 2013). For ITS phylogenetic analysis, a data set of 42 sequences originated from *P. perurans* and spanning 410 bp region was aligned. FASTA alignments were uploaded into BEAUTI software v2.4.7 to create the xml file for subsequent analysis. A Bayesian phylogenetic analysis was performed in BEAST software v2.4.7 incorporating the HKY + G substitution model under Coalescent Bayesian Skyline priors. Ancestral location reconstruction was implemented using the method outlined in Lemey et al. (2009) to predict the geographic origins of *P. perurans* clones/strains. The Markov chain Monte Carlo method (MCMC) was run for 100 000 000 generations (every 1 000 generation was logged), with the first 10% of trees being discarded for burn-in using TreeAnnotator (Drummond et al 2012). Tracer software v.1.6 was used to assess convergence and mixing were assessed the convergence of MCMC where all effective sample size statistics (ESS) were >200.

5.2.4 Results

*P. perurans* in vivo experimental exposure in Atlantic salmon

No changes in fish behaviour were observed at the exposure stage and all fish survived until the end of experiment. No moribund fish were observed in the negative tank or any of treatment’s tanks. During tissue dissection, gills from all fish appeared healthy and no gross pathology signs associated with the AGD such as pale gills or extensive white mucosal patches were observed in any of infected or negative control fish. There appeared to be no visible signs of the disease even in the fish infected with the highest dose of 5000 *P. perurans* cells/L.

Gross gill scoring of lesions from the NBF fixed material associated with AGD was performed for all gill arches and the data is summarized in Table 10. No signs of lesions were observed in 96% of infected fish in the entire experimental infection challenge. Only fish infected with 20, 1280 and 5000 cells/L were exhibiting any signs of AGD with very low prevalence of 2/10, 1/10 and 3/10 fish in the individuals tanks respectively. In all cases only one gill arch was affected (data not shown). Taking into account results from the previous experiments (Collins et al 2017; McCarthy 2015) where Atlantic salmon, following exposure to 5000 *P. perurans* cells/L, exhibited gross gill scores of 3 or higher, it was concluded that the experimental challenge failed to
infected fish on this occasion. No further analyses were performed and it was established that material from the disease challenge was unsuitable for further work.

Table 10 Gill scores associated with amoebic gill disease, following experimental immersion challenge in Atlantic salmon

<table>
<thead>
<tr>
<th>Tank No.</th>
<th>P. perurans cells/L</th>
<th>No. fish</th>
<th>Status</th>
<th>Gross gill score (according to Taylor et al 2009)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>survivor</td>
<td>moribund</td>
</tr>
<tr>
<td>1</td>
<td>mock 10</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>160</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>320</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>640</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>1280</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>5000</td>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Totals (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Genetic diversity in the P. perurans COI gene**

A PCR product of approximately 900b was amplified, using Euglycox 1F and Euglycox 1R, from 30 samples of *P. perurans* originated from Ireland, Norway, Scotland and Tasmania, one sample of *P. branchiphila* and two samples of *P. pemaquidensis* obtained from the public culture collections. No product was obtained with the LCO1490F and HCO2198R primers. A blastn search confirmed that the partial sequences from *P. perurans* and *P. branchiphila* represents the amoebas rather than their endosymbiont with the closest hit to *P. pemaquidensis* sequence KX611830 (89% similarity). Sequence of the *P. pemaquidensis* isolate 1560/4 generated in this study (619 bp) and the same isolate sequenced by Tanifuji et al (2011) (KX611830) were identical. On the other hand, there was 10% variation found between the COI sequences generated from the two *P. pemaquidensis* strains, 1560/04 originally isolated from environmental sample in Wales and ATCC50172 isolated from coho salmon gill in USA.

From the multisequence alignment, spanning 475bp, 6 different haplotypes of *P. perurans* were identified using the ElimDupes from the material available. The results suggested that there is no variation between Scottish *P. perurans* strains regardless of the fish host origin. Also *P. perurans* sequences generated from a polyculture isolated farmed Atlantic salmon and two monocultures developed in MSS (data not shown) were identical, regardless of the varying virulence established in a disease challenge experiment (Collins et al 2017). Samples collected in Norway and Ireland were also identical to Scottish *P. perurans*. The greatest diversity of haplotypes was found in Tasmania (5 haplotypes). There was no overlap in the *P. perurans* COI haplotypes between Europe and Tasmania and the variation ranged from 0.84-1.47%.
*P. perurans* genetic diversity in the ITS rDNA

The partial ITS rDNA (1.1kb) was amplified and sequenced from a single *P. perurans* polyculture from Scotland, Norway and Tasmania respectively and two *P. perurans* monocultures from Scotland. A high level of intraspecific variation was observed in ITS generated from different *P. perurans* strains. There was also high variability between clones originated from individual polyculture and surprisingly also from *P. perurans* monocultures (Table 11). Due to multiple deletions and insertions, diversity within the ITS region was calculated from smaller data set of 537 bp generated after removal of the deletions/insertions and ambiguous bases.

A Bayesian inference analysis was run for the *P. perurans* partial ITS region generated from the polycultures only and *P. pemaquidensis* (EU884483) and *P. branchiphila* (EU884491) were included as outgroups. There was no structuring of clades based on the geographic origin of the *P. perurans* cultures. The genetic information generated from the partial ITS was very “noisy” and any of the ITS sequences from clones, originating from individual *P. perurans* strains, form any monophyletic groups (Figure 23).

![Bayesian phylogenetic analysis performed with 537bp partial ITS rDNA](image)

*Figure 23* Bayesian phylogenetic analysis performed with 537bp partial ITS rDNA
Table 11 Diversity in the *P. perurans* partial ITS rDNA fragment

<table>
<thead>
<tr>
<th>Country</th>
<th>N of clones screened/ N of unique ITS sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Polyculture</strong></td>
<td><strong>Scotland</strong> / 16/15</td>
</tr>
<tr>
<td><strong>Polyculture</strong></td>
<td><strong>Norway</strong> / 19/17</td>
</tr>
<tr>
<td><strong>Polyculture</strong></td>
<td><strong>Tasmania</strong> / 8/8</td>
</tr>
<tr>
<td><strong>Monoculture</strong></td>
<td><strong>Scotland</strong> / 9/7</td>
</tr>
<tr>
<td><strong>Monoculture</strong></td>
<td><strong>Scotland</strong> / 5/5</td>
</tr>
</tbody>
</table>

A multisequence alignment spanning the complete 150 bp region of 5.8 rRNA gene was generated from the same data set. Significantly lower intraspecific diversity was recorded from this region and similarly, as for the ITS region, no country or strain specific clustering was observed. In addition, due to the reduction in number of informative sites, the 5.8S sequences belonging to *P. perurans* did not form a monophyletic group and could not be separated from *P. pemaquidensis* and *P. branchiphila* (data not shown).

5.2.5 Discussion

Unfortunately, the analysis of the gill scores, following exposure of Atlantic salmon to *P. perurans* polyculture maintained in the MSS, revealed a very low level of infection and was concluded as unsuccessful. Low gill scores (0-1) were observed even for fish infected with the highest dose of 5000 *P. perurans* cells/L. The most likely explanation of the reported failure was loss of virulence of the *P. perurans* culture used in experimental challenge. Loss of *P. perurans* virulence when maintained in cultures, without being passed back through a host, has been previously reported by another laboratory (Bridle et al 2015; Crosbie et al 2012). The mechanism by which cultures lose virulence has not been identified as genome information of *P. perurans* remains largely unexplored to date. In the experimental conditions, the percentage of fish developing gill lesions, following exposure to a clonal culture, was reported to be reduced by a half as early as 70 days post isolation and total loss of infectivity was observed with cultures maintained for 3 years (Bridle et al 2015; Crosbie et al 2012). However the cultures used in the reported challenge were previously used in multiple published infection trials carried out in 2014 and 2015 (McCarthy et al 2015; Collins et al 2017) and most recently in February 2016 (unpublished). All experiments resulted in Atlantic salmon morbidity and high gill scores in a large proportion of infected fish. Therefore there were no obvious known indications suggesting that the culture used would fail to infect Atlantic salmon prior to executing this trial. There are currently no available techniques to establish infectivity of *P. perurans* cultures which could be performed prior to an experimental challenge. It is a common practise to refer to the published results obtained from past challenges and even if the infectivity of *P. perurans* decreased by 50% as described by Crosbie et al (2012) during similar period of time, more infected fish with higher gross gill scores were expected to be reported.
here. The failure to reproduce the AGD was unfortunate and unforeseen and alternative objective was agreed with the project co-ordinator and is reported below.

There are some sequence data available for the P. perurans ribosomal DNA, either of the small ribosomal subunit (18S) or internal transcribed spacer (ITS) which describe mainly diversity among Paramoeba/Neoparamoeba species such as P. perurans, P. pemaquidensis and P. branchiphila however the changes in nomenclature among these species/genera, over the period of last 10 years, making sequence comparisons incredible hard. Majority of published papers are reporting on diversity within the 18S ribosomal gene, suggesting a high level of similarity between both Paramoeba species (>90% similarity) (Dyкова et al 2005; Steinum et al 2008; Wong et al 2004). A low level of intraspecific variability was also reported for both P. perurans and P. pemaquidensis collected from different geographic region (1.18% and 0.85% diversity for both species respectively) however the data set were very limited for both geographic region and number of samples analysed (Wong et al 2004).

Only few studies attempted to investigate P. perurans diversity within the ITS region considering it is one of the most common regions studied for finfish parasites (Hansen et al 2007; Hayward et al 2001; Yazawa et al 2008; Matejusova et al 2001, 2004). Paramoeba ITS region, especially ITS1 is highly variable and often excluded from phylogenetic analysis due to lack of reliable homology resulted from presence of numerous indels. On the other hand the 5.8S gene region is highly conservative and lacks interspecific variation (Young et al 2014) which was also observed in the present study. Young et al (2004) utilized the ITS2 region for phylogenetic analysis and demonstrated monophyletic clades for numerous Paramoeba species such as P. pemaquidensis, P. perurans, P. branchiphila and P. aestuarina. Intraspecific variability of ITS2 was reported for P. perurans (97.4-99.4% similarity) however Young et al (2004) only studied Tasmanian strains and therefore it was not possible to assess if this region can be utilized for tracking of pathogen spread. High level of microheterogeneity was also observed in P. pemaquidensis ITS and it was concluded that variation did not resulted from PCR errors or mixed strains of amoeba in a single samples as they were observed in sequences generated from individual clones obtained from monoclonal cultures (Caraguel et al 2007). The present study also describes variation with the ITS region however our analysis concluded that this region is not phylogenetically informative. This is because of the fact that sequences obtained from numerous clones of a same sample failed to form monophyletic groups and/or show fixed nucleotide differences which can explored for strain differentiation.

This report provides to our knowledge a first insight into mitochondrial DNA variability of an important pathogen P. perurans responsible for significant economic losses in farmed Atlantic salmon worldwide. Clear separation between P. perurans haplotypes originated from Tasmania and Europe was observed. There were five unique Tasmanian P. perurans haplotypes exhibiting low variability. On the other hand, P. perurans haplotypes circulating in the European aquaculture appeared to be mixing and there were no unique haplotypes associated with either Ireland, Norway or Scotland. Genome wide analysis of P. perurans strains spanning wide spatial and temporal aspects is essential to shed more light on spread of this pathogen in aquaculture and to discover potential genetic markers associated with P. perurans virulence.
5.2.6 References


EMBRIC – D8.2: Report with industry standards and instructions for molecular tools in paternity assignment, traceability and marker assisted selection, page 48 of 76


Moën, T., Baranski, M., Sonesson, A.K., Kjøglum, S., 2009. Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait. BMC Genomics 10, 368


5.3 Blue mussels (*Mytilus edulis*): Developing experimental disease models underpinning investigation of susceptibility of blue mussels to bacterial infections (Partner 6)

5.2.1 Background

Aquaculture production of blue mussels (*Mytilus edulis*) in Scotland has increased from 7,374 tonnes to 12,669 tonnes between 2010 and 2017, with anticipated potential to reach 21,000 tonnes per annum by 2030 (Anon). In the aquatic environment, bivalves are exposed to a range of bacterial species, many of which are non-cultivable and are only now being better characterised by using culture-independent genetic methods, such as next-generation sequencing. Through filter feeding, bivalves are capable of concentrating bacteria from the water and from sediment and it appears that, except in certain circumstances, the shellfish are not negatively impacted. Since 2010, MSS performed microbial testing of mussel haemolymph or tissue homogenates as part of the routine diagnostic investigations into unexplained shellfish mortalities. In the aquatic environment, mussels are exposed to a range of bacterial species, both as colonisers of their external surfaces and through filter feeding, so that the significance of the isolated bacteria in the context of the mortalities was uncertain. Although it was not possible to characterise all of the bacterial isolates recovered, traditional bacteriological testing and sequencing of the more predominant isolates identified the majority as members of the genus *Vibrio*.

Not all species of *Vibrio* are pathogens, most species are ubiquitous in aquatic environments but some species are known to be associated with Pacific oyster mortality (Garnier et al 2007). Taxonomy within the genus *Vibrio* is not completely resolved and with the increased availability of DNA sequencing *Vibrio* species are often re-classified, for example *Vibrio splendidus* LGP32 associated with disease in shellfish, has been renamed to *Vibrio tasmaniensis* LGP32 based on multilocus sequence analysis (Gay et al 2004; Sawabe et al 2013). In addition, different *Vibrio* strains belonging to the same “species” may exhibit different virulence properties and two proteins, *OmpU* and *Vsm* have been identified as virulence markers in *V. tasmaniensis* LGP32 (Binesse et al 2008; Duperthy et al 2010, 2011). Sequencing the virulence genes as well as the genes necessary to ascertain *Vibrio* species, such as *RecA*, *RpoB* and *AtpA*, is essential to understand potential pathogenicity of *Vibrio* species in shellfish aquaculture. Stagg et al (unpublished) demonstrated that four Scottish *Vibrio splendidus* strains, isolated from moribund cultured blue mussel in Scotland, belong to the *Vibrio tasmaniensis* LGP32 clade however later unpublished study (McCarthy and Chapuis 2015) indicated that, based on the *OmpU* and *Vsm* sequences, these Scottish strains differed from the LGP32 reference strain.

Immune responses of blue mussels to bacterial pathogens are robust and reasonably well described, including an array of defences from production of pallial mucus to phagocytic haemocytes which produce a range of defensive compounds such as the anti-microbial peptides, mytilin and defensin (Allam et al 2016; Parisi et al 2008). Immune signalling pathways involving a range of pattern recognition receptors such as the transmembrane Toll-like receptors (TLRs), which are capable of recognising microbial derived pathogen-associated molecular patterns (PAMPs), have been identified in bivalves (Allam and Raftos 2015; Toubiana et al 2013). In addition, the soluble interleukin IL-17 receptors and related signalling components has been detected in bivalve gill tissue and demonstrated to play an role in defence against bacteria, while peptidoglycan recognition protein (PGRP) or LPS-binding protein (LBP) were more highly expressed in mussel haemocytes (Philipp et al 2012). Therefore
there is an urgent need to establish pathogenicity of the *Vibrio* isolates circulating in Scottish aquaculture and establish robustness of shellfish stock to ensure sustainable growth of shellfish aquaculture for future generations.

5.3.2 Aims of the study
Aim was to further characterize Scottish *Vibrio* strains isolated in 2010 from moribund blue mussels and develop a disease challenge model which allows us to study bacteria persistence in mussels and determine virulence of these isolates. In addition, a various mussel tissues will be collected at subsequent time points to measure the expression of mussel immune-related genes, as indicators of a mussel response to the bacteria.

5.3.3 Material and Methods

*Primer design and genetic characterization of Vibrio strains*

Forty-five Scottish bacterial strains, isolated from shellfish and fish farms, *Vibrio* *tasmaniensis* LGP32 strain (obtained from F. Le Roux, Roscoff, France) and 10 strains *V. splendidus* isolated from marine fish (provided by S. Gulla and D. Colquhoun, Norwegian Veterinary Institute, Oslo, Norway) were cultured on Marine Agar (MA, Difco) or tryptone soya agar (Oxoid) with salt (TSAS, 2% final NaCl concentration) at room temperature. Genomic DNA was extracted from single colonies of pure bacterial culture using a DNeasy Blood and Tissue kit (Qiagen, UK), following the manufacturer’s instructions for Gram-negative bacteria. PCR amplification was carried out using GoTaq® G2 Green Master mix (Promega, UK) and primers summarized in Table 12. No template controls were included throughout the entire process.

The *OmpU* and *Vsm* sequences from Scottish *Vibrio* isolates obtained during the earlier study (McCarthy & Chapuis 2015) were aligned with *V. splendidus* sequences published by Nasfi et al (2015) and with sequences from other vibrio species, including *V. tasmaniensis* LGP32, to highlight regions of homology or variation. For *OmpU*, sequences from 12 Scottish isolates obtained with primer pair ompU215F/ompU645R were used in BLASTN searches to retrieve related sequences from 39 *V. splendidus* strains, LGP32 and 7 other vibrio species. For *Vsm*, sequences from 12 Scottish isolates obtained with primers vsm1618F/vsm1919R were used to retrieve sequences from an additional 32 *V. splendidus* strains, *V. tasmaniensis* LGP32 and 36 other vibrio species. A reverse primer for PCR, ompU442R, was designed and used with ompU215F to target the 4 Scottish isolates and LGP32, in order to amplify *OmpU* DNA from samples collected during the challenge experiments. Primers against conserved regions of the *Vsm* gene were used (OligoAnalyzer, http://www.idtdna.com/calc/analyze) for detecting other *Vibrio* species, particularly *V. splendidus*.

A phylogenetic analysis was carried out to establish relationships between the Scottish isolates and other *V. splendidus* using sequences of the two virulence genes. Analysis of *OmpU* sequences was conducted on 356 positions from 59 sequences in MEGA. Analysis of *vsm* sequences was conducted on 227 positions from 53 sequences. Evolutionary history was inferred using the Neighbor-Joining method and distances were computed using the Maximum Composite Likelihood method. Bootstrap percentages (1000 replicates) are shown next to the branches.
Table 12 Primers used for amplification of Vibrio *ompU* and *vsm* genes, and mussel nuclear DNA

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Target</th>
<th>Sequence 5' to3'</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>OmpU215F</td>
<td>Vibrio</td>
<td>TCAGAGYGGCCTATACGTTGT</td>
<td>This study</td>
</tr>
<tr>
<td>OmpU442R</td>
<td>Vibrio</td>
<td>GGAATTGACCTTTGTAAGAAACATGTTGT</td>
<td>This study</td>
</tr>
<tr>
<td>OmpU654R</td>
<td>Vibrio</td>
<td>TATGTTAAGAGCCCGAAGC</td>
<td>This study</td>
</tr>
<tr>
<td>S-METVS2F</td>
<td>Vibrio</td>
<td>ACCTAAGGGAACCAATCT</td>
<td>De Decker et al (2011)</td>
</tr>
<tr>
<td>S-METVS2R</td>
<td>Vibrio</td>
<td>GTGAAAGCGACAGTGTA</td>
<td>De Decker et al (2011)</td>
</tr>
<tr>
<td>vsm1466F</td>
<td>Vibrio</td>
<td>TATYTTGGTAAWGGYYTGGATATGTA</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1553F</td>
<td>Vibrio</td>
<td>RAYGTGAAAAYGCGTYYGTGAA</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1618F</td>
<td>Vibrio</td>
<td>ACCCGATTGCGATACACGTGA</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1652F</td>
<td>Vibrio</td>
<td>AGYCAAGGGNTTYACYGAGCA</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1724F</td>
<td>Vibrio</td>
<td>CNGAYATYGCDGGYGAAGC</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1822R</td>
<td>Vibrio</td>
<td>TGYTCTCRATNACDGCACCAC</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1913R</td>
<td>Vibrio</td>
<td>ATCCAAACCGATTTGCTGCTA</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1919R</td>
<td>Vibrio</td>
<td>CTCTTAGAATTTCAACCGATTTGT</td>
<td>This study</td>
</tr>
<tr>
<td>vsm1969R</td>
<td>Vibrio</td>
<td>GSYGTCCARTANARYTGTTG</td>
<td>This study</td>
</tr>
<tr>
<td>vsm2188R</td>
<td>Vibrio</td>
<td>GAYGCGAATCATCAGGGTAAGAAGGTA</td>
<td>This study</td>
</tr>
<tr>
<td>ME15F</td>
<td>M. edulis</td>
<td>CCAGTATACAAAACCGGAAAGAAGGTA</td>
<td>Inoue et al (1997)</td>
</tr>
</tbody>
</table>

**Blue mussel in vivo challenge and sample collection – 1st challenge**

Bacterial extracellular products (ECPs) were obtained from *V. splendidus* M1, M31 and *V. tasmaniensis* LGP32 strains using a cellophane overlay method (Liu et al 2013). A 500 μl aliquot of an overnight culture of bacteria was spread onto a TSAS plate covered by a cellophane film. After incubation at room temperature for 24 h, bacteria were harvested using 1 ml of cold 1 × phosphate-buffered saline (PBS). After centrifugation for 10 min at 4,000 × g, the supernatant was passed through a 0.22 μm pore filter to provide a crude ECP preparation. Protease activity was determined using azocasein (Sigma, UK) as a substrate. Crude ECPs (250 μl) were added to 250 μl of azocasein (5 mg ml⁻¹ in 50 mM Tris-HCl buffer, pH 8.0). PBS was used in place of crude ECP as a negative control. The mixture was incubated at 20°C for 1 h after which the undigested substrate was precipitated by adding 500 μl of 10% trichloroacetic acid to the reaction mixture and centrifuging at 12,000 × g and 4°C for 5 min. The supernatant (500 μl) was neutralized by adding an equal volume of 1 N NaOH. After mixing, the absorbance at 405 nm was measured using a 96-well plate reader. To produce sufficient bacteria for the challenge experiment, 2 x 1.5 L volumes of Marine Broth (MB, Difco, UK) were inoculated from starter cultures of M31 and LGP32. The bacteria were grown for 36 h at 18°C with shaking at 180 rpm. The final bacterial concentration was 5.7 x 10⁹ colony forming units (CFU) ml⁻¹. The bacterial counts were made by producing a series of tenfold dilutions to 10⁻⁸ in seawater and spreading 100 μl of each dilution onto Brain Heart Infusion (Oxoid) agar made with seawater.

EMBRIC – D8.2: Report with industry standards and instructions for molecular tools in paternity assignment, traceability and marker assisted selection, page 53 of 76
Juvenile mussels (1-3 cm shell length) were acclimatised at 11°C and a salinity of 35 ppt over a three week period in flow-through natural seawater (300 L/h) with supplementary aeration, and fed Instant Algae Shellfish diet (Varikon Aqua Solutions Ltd., Malvern, UK) daily. The mussels were then distributed by weight between 6 baskets (approximately 150 mussels per basket). Two baskets per treatment group were placed in individual tanks with seawater flow-through of 40 L h⁻¹ at 11°C with supplementary aeration. There were three different treatment groups, *Vibrio* strain M31 as a putative pathogenic Scottish strain, LGP32 as a known pathogen and a mock treatment group exposed to bacterial growth medium. Mussels were exposed to 2 x 10⁸ CFU ml⁻¹ with water flow stopped and after 2 h exposure the water flow was resumed. Immediately after the challenge and at 1, 2, 5, 7 and 9 d post-challenge (dpc), 10 mussels from each group were dissected from their shell and placed in RNAlater (Sigma, UK), held at 4°C overnight and stored at -80°C. For DNA extraction, 10 mussels per group were placed in 95% ethanol and stored at -80°C. For histological analysis, 5 mussels were opened by severing the adductor muscle and placed in Davidson’s fixative for 24-48 h before processing and wax embedding using routine histological procedures. Prior to embedding, mussel soft bodies were dissected into 3 layers, producing cross-sections of the mussel organs, with particular focus on the mantle and digestive gland. 5 μm sections were stained with haematoxylin and eosin and Gram stain. 

**Blue mussel in vivo challenge and sample collection – 2nd challenge**

Bacterial suspensions for bath challenges (4.4 x 10⁹ CFU ml⁻¹) were produced in 5 L MB by incubation for 48 h at room temperature, with shaking at 130 rpm. Additional bacterial suspensions for challenges of 10⁶ cfu ml⁻¹ were made by adding 5 ml of M1 or M40 cultures to 5 L MB. Confirmation of bacterial concentrations was done by colony counts on TSAS and MA of 50 μl aliquots taken from triplicate tenfold dilutions to 10⁻⁸ in seawater. 

Farmed juvenile and pre-adult mussels (mean shell length 4.01 ± 0.8 cm) were initially divided into groups of 5-20 individuals within mesh bags for acclimatising at 11°C as described above. During the two week acclimatisation period, mussels were fed twice daily using Phytogreen-S (Brightwell Aquatics) and after 2-3 days, the mesh bags were opened to facilitate water flow through the mussel clusters. The mussels were gently separated and each treatment group was divided between two 5 L beakers containing natural seawater at 15°C with supplementary aeration. First beaker contained 30 individuals which were left “undisturbed” throughout the challenge period. Second beaker contained 50 individuals which were sampled at 3h and 2, 3 and 6 days post challenge. The bacterial challenge was carried out by replacing the seawater with the bacterial culture, holding for 2 h with aeration, decanting the bacterial culture, rinsing the mussels in clean seawater and replacing with 4 L fresh seawater. There were five treatment groups included in this second experimental disease challenge: mussels were exposed to two putatively pathogenic Scottish strains M1 and M40, and a positive control strain (LGP32) at two different doses 4.4 x 10⁹ CFU ml⁻¹ and lower dose at 10⁶ CFU ml⁻¹. Two negative groups of mussels were exposed to seawater (SW) and a heat-killed strain M1 (HK, 3.54 x 10⁹ CFU ml⁻¹) respectively. Inactivation was achieved by autoclaving the bacterial culture at 120°C for 40 min and verified by inoculation of aliquots on MA. 

Pre-exposure samples of mussel tissue were collected 20 h prior to the challenge. Other samples were collected throughout the challenge period as outlined in Table 13. Gill tissue was dissected aseptically, placed in RNAlater, held at 4°C
overnight and stored at -80°C. The remaining tissue from the same individual was macerated using a sterile sample and a subsample was placed in 95% ethanol and stored at -80°C. For histological analysis, mussels were carefully dissected from shells and placed in Davidson's fixative for 24 h before transfer to 10% neutral buffered formalin (NBF) until processing and wax-embedding as described above. Mussels were considered dead or moribund when valves were gaping and failed to close following stimulation.

**Table 13 The different exposure groups and the sampling schedule used in Challenge 2**

<table>
<thead>
<tr>
<th>Treatment (CFU ml⁻¹)</th>
<th>Samples collected (purpose)</th>
<th>Sampling time post challenge and number of samples collected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>pre-exposure 3h 1d 2d 6d</td>
</tr>
<tr>
<td>seawater</td>
<td>Gill (gene expression)</td>
<td>10 10 10 10 10</td>
</tr>
<tr>
<td></td>
<td>Macerated body (Vibrio persistence)</td>
<td>10 5 5 5 5</td>
</tr>
<tr>
<td></td>
<td>Whole mussel (histology)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Undisturbed</td>
<td>30</td>
</tr>
<tr>
<td>MT1 (4.4 x 10⁹)</td>
<td>Gill (gene expression)</td>
<td>10 10 10 10 10 10</td>
</tr>
<tr>
<td></td>
<td>Macerated body (Vibrio persistence)</td>
<td>5 5 5 5 5</td>
</tr>
<tr>
<td></td>
<td>Whole animal (histology)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Undisturbed</td>
<td>30</td>
</tr>
<tr>
<td>MT40 (4.4 x 10⁹)</td>
<td>Gill (gene expression)</td>
<td>10 10 10 10 10</td>
</tr>
<tr>
<td></td>
<td>Macerated body (Vibrio persistence)</td>
<td>5 5 5 5 5</td>
</tr>
<tr>
<td></td>
<td>Whole animal (histology)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Undisturbed</td>
<td>30</td>
</tr>
<tr>
<td>LGP32 (4.4 x 10⁹)</td>
<td>Gill (gene expression)</td>
<td>10 10 10 10 10</td>
</tr>
<tr>
<td></td>
<td>Macerated body (Vibrio persistence)</td>
<td>5 5 5 5 5</td>
</tr>
<tr>
<td></td>
<td>Whole animal (histology)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Undisturbed</td>
<td>30</td>
</tr>
<tr>
<td>MT1 (1x10⁵)</td>
<td>Macerated body (Vibrio persistence)</td>
<td>5 5 5 5 5</td>
</tr>
<tr>
<td></td>
<td>Undisturbed</td>
<td>30</td>
</tr>
<tr>
<td>MT40</td>
<td>Macerated body (Vibrio persistence)</td>
<td>5 5 5 5 5</td>
</tr>
<tr>
<td></td>
<td>Undisturbed</td>
<td>30</td>
</tr>
</tbody>
</table>

*Persistence of Vibrio spp. in experimentally infected blue mussels*

Mussel tissue was homogenised and DNA was extracted from homogenate equivalent to 15 mg tissue using DNeasy Blood and Tissue kit (Qiagen) following the
manufacturer’s protocol. Real time PCR was carried out in duplicate on, in 10 µl reaction volumes using LightCycler® 480 SYBR Green I Master (Roche, UK) mix (Challenge 1) or GoTaq® qPCR Master Mix (Challenge 2). The primers ompU 110F/366R were used to specifically amplify *ompU* DNA from the *Vibrio* strains used in the challenge and the primer pair ME15/16 was used as housekeeping gene. In Challenge 2, an additional PCR amplification with the primers vsmF311a/R612a was used to test for the absence of other vibrio species in the challenged mussels. A log10 dilution series of vibrio DNA diluted in HSDNA, was added in triplicate to each plate to act as a standardiser between amplification runs.

The Vibrio level in the different challenge groups of mussels was derived from the Cp values for vibrio *ompU* DNA (*Vibrio* DNA) and *M. edulis* 18S rRNA DNA (mussel DNA). The maximum cut-off Cp value was set at 40 cycles for both amplifications, and the value for each sample was obtained by subtracting the Cp of each sample from 40. The ratio of the value for Vibrio DNA to mussel DNA was calculated for each sample. The relative Vibrio level for each challenge group at a particular sampling time is the average of the individual ratios, plotted as fold-change (± SD) against the average ratio for seawater-exposed mussels at the same sampling time. Differences between the challenge groups were calculated from the individual ratios for each group at a particular sampling time, using Levene heteroscedasticity with significance level set at 0.05.

**Immune gene expression in blue mussel exposed to Vibrio spp.**

RNA was extracted from gill tissue (equivalent to 10 mg tissue) using an RNeasy Mini Kit (Qiagen) according to the manufacturer’s instructions. A DNase digestion step was incorporated to reduce the amount of genomic DNA contamination. The purity and concentration of extracted RNA was measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific, UK). Reverse transcription was carried out using TaqMan® MultiScribe® Reverse Transcriptase with oligo (dT)16 primers (Thermo Fisher Scientific), following the manufacturer’s protocol except that reactions were carried out in 2x volumes with 2 µg total RNA as template. Extraction controls and ‘no template’ controls were generated to confirm the absence of contamination and the effective reduction of genomic DNA.

Analysis of immune-related genes expression was carried out using the primers listed in Table 14 on the LightCycler® 480 Real-Time PCR System (Roche), in 10 µl reaction volumes using LightCycler® 480 SYBR Green I Master (Roche, UK) mix (Challenge 1) or GoTaq® qPCR Master Mix (Challenge 2). A triplicate set of standards, log 2 dilutions of mixed sample cDNA, was added to each plate to derive the amplification efficiency of each primer. Initial tests indicated inhibition of PCR in neat samples and consequently, all cDNA samples were diluted 1:1 with nuclease-free (NF) water prior to amplification. Extraction controls and no template reverse transcription controls were tested against each primer pair. In addition, amplification controls containing no template were included in each reaction.
### Table 14 Real time PCR assays used in the gene expression analysis of blue mussel tissue exposed to *Vibrio* strains

<table>
<thead>
<tr>
<th>Target gene</th>
<th>Primer name</th>
<th>5’-3’ forward primer</th>
<th>5’-3’ reverse primer</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mytilin(^1)</td>
<td>Myt</td>
<td>GTATTCTGGCTATGCTCTTG</td>
<td>GTATAATGTCAAAAGACGGGTC</td>
<td>Cellura et al (2007)</td>
</tr>
<tr>
<td>Defensin(^2)</td>
<td>Def</td>
<td>GTCGTTGTCTGATGATGCCG</td>
<td>CCTGAGTCTGAAAACACG</td>
<td>Cellura et al (2007)</td>
</tr>
<tr>
<td>ME1, IkB</td>
<td>IkB1</td>
<td>GAGACACACCTTTACATCG</td>
<td>TCCCCACATTTACATCTGGA</td>
<td>Philipp et al (2012)</td>
</tr>
<tr>
<td>ME1-LBP</td>
<td>LBP</td>
<td>CGGTTCAGTTTCAGGGAT</td>
<td>TCTGCTGTCTGAGTGTTGGA</td>
<td>Philipp et al (2012)</td>
</tr>
<tr>
<td>Myd88</td>
<td>MyDU</td>
<td>GAGACGTTGCCGCTCAAATGTT</td>
<td>TTCTGCAATGGAGGCGCTA</td>
<td>Philipp et al (2012)</td>
</tr>
<tr>
<td>Me2-IL17</td>
<td>IL-17</td>
<td>TCGAACACATCTTCCAGA</td>
<td>AACAACAAATACACACTGAAAAG</td>
<td>Philipp et al (2012)</td>
</tr>
<tr>
<td>TLR-i</td>
<td>TLR</td>
<td>AGGATGGCTTGAACCTGGATT</td>
<td>AGTCGAGTACGTTCCTGTA</td>
<td>Toubiana et al (2013)</td>
</tr>
<tr>
<td>Elongation factor 1-alpha</td>
<td>EF-1alpha</td>
<td>GAAAGCCGACAGACAGAGC</td>
<td>GGGAAGGTAAACACCAT</td>
<td>Toubiana et al (2013)</td>
</tr>
</tbody>
</table>

#### 5.3.4 Results

**Genetic characterization of Vibrio strains circulating in Scottish aquaculture**

Amplification of template DNA from M1, M31, M36 and M40 with the primer pair ompU215F/ompU654F produced an amplicon of approximately 450 bp. This region spans area between positions 250 and 626 of *V. tasmaniensis* LGP32 (Accession number FM954972.2; locus tag VS_2494; Protein_ID CAV19651). Comparison with sequences from other vibrios revealed 100% identity of the *ompU* gene in M31, M36, M40 and LGP32 in this region including positions 510 and 540 where some *V. splendidus* strains have an additional 12 and 15 bases. The *ompU* sequence obtained from M1 with these primers (368 bases) segregated this species into a different genetic clade but this sequence was identical to another of the Scottish isolates, M11, associated with mussel mortalities in autumn 2010. The relative relationships between the *ompU* sequences from Scottish vibrio isolates and other vibrio strains is illustrated in Figure 24.

None of the primer pairs used in this study amplified sequence from a *vsm* gene of M1, M32, M36 or M40 strains but products of the expected molecular weight were obtained from LGP32 (not shown). Using vsm1618F/vsm1919R primers, spanning a region in extracellular zinc metalloprotease gene (VS_1267; CAV: 18307; complement 1365498 - 1367321) PCR products identified as the *vsm* gene were obtained from other Scottish vibrio isolates and from Norwegian *V. splendidus* isolates.

The primers vsm1466F/vsm2188R amplified an identical sequence from M1, M36 and M40, which was found to be from a ferrous iron transport protein B (homologous to positions 2397245..2399518, from *V. tasmaniensis* LGP32, Accession number FM954972.2; locus tag VS_2233; Protein_ID CAV19397.1). In contrast, the same primers amplified *vsm* sequence from Scottish isolate M5 and from LGP32. The azocasein digestion test for protease activity in ECPs was positive for LGP32 but negative for M1 and M31. The relative relationships between *vsm* sequences from Scottish vibrio isolates and other vibrio strains is illustrated in Figure 25.
Figure 24 Evolutionary relationships of Scottish vibrio strains, *V. tasmaniensis* LGP32 and selected isolates from GenBank with respect to partial OmpU sequences. The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 2.49941194 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The rate variation among sites was modeled with a gamma distribution (shape parameter = 1). The analysis involved 59 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 356 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.
Figure 25 Evolutionary relationships of Scottish vibrio strains, *V. tasmaniensis* LGP32 and selected isolates from GenBank with respect to partial *Vsm* sequences. The evolutionary history was inferred as for Figure 3. The optimal tree with the sum of branch length = 2.28748522 is shown. The involved a total of 227 positions from 53 nucleotide sequences.

Pathogenicity of Scottish Vibrio isolates in blue mussel

Unfortunately, the mussels were from a population that had been experiencing recent mortalities which continued throughout the acclimatisation period and reached 80% during the challenge. No difference in mortality rate was detected between any of the challenge groups. Consequently, the experiment was terminated and a limited number of samples were used solely to test and optimise PCR and RT-PCR assays, namely mussels exposed to M31 or SW, from 0, 24 and 48 h post-challenge.

The crossing point (Cp) values obtained for *ompU* DNA, in mussels from the M31 (test) and MB (control) groups, are shown in Figure 26. The values indicate a high concentration of *Vibrio* in challenged mussels immediately after exposure, with a considerable reduction by 24 h post-challenge and no difference from that of the control mussels by 48 post-challenge. At all time-points, the Cp values of the unchallenged control mussels were below the minimum value for meaningful amplification (Cp < 36). In endpoint PCR of mussel cDNA, primers EF-1a, IL-17,
MyDU, MyDPp and TLR amplified the target sequence as confirmed by sequencing of products. However, in real-time PCR only the primers EF-1a, IL-17 and MyDU produced a complete and reliable set of Cp values with reproducibility between replicates, Cp values above 35 and efficient amplification across the range of Cp values generated by the test samples. These results were not further analysed but were used to assist optimisation of assays for the second challenge.

![Figure 26](image)

**Figure 26** *Vibrio* level in mussel tissue shown as crossing point (Cp) values (± SD) from real-time PCR amplification of OmpU DNA. Test mussels were exposed to *Vibrio* strain M31 or to bacterial growth medium, Marine Broth (control). Values are from samples taken immediately post-challenge (2 hours) and at 24 and 48 hours post-challenge.

In the second experimental challenge, mussel mortality was observed in all experimental groups with the exception of mussels exposed to seawater (Figure 27). The highest mortalities occurred in groups exposed to the pathogenic *V. tasmaniensis* LGP32 strains (22.9%), and to M40 strain (17.5% and 20.0% in sampled and undisturbed mussels, respectively). However, mortality in the mussels exposed to heat-killed bacteria (5.0% and 6.7% in the sampled and undisturbed beakers, respectively) was similar to the mortality in the LGP32 undisturbed beaker (6.7%).

The level of mortality in the mussels exposed to low dose M1 (11.9%) was considerably greater than in mussels exposed to the higher concentration of M1 (1.4% and 3.3% in the sampled and undisturbed beakers, respectively) and greater than in mussels exposed to low dose M40 (2.0%). It was not possible to detect changes specific to treatment/groups nor any pathology which could specifically be associated with the presence of bacteria.

The persistence of *Vibrio* strains following experimental exposure of blue mussel is illustrated on Figure 28. At 3 hours post-challenge, levels of *Vibrio* DNA in all groups exposed to high concentration of *Vibrio* were significantly higher than in seawater exposed group. At 24 hours post-challenge, *Vibrio* levels in all challenge
Figure 27 Percentage mortality in mussels exposed for 2 h in challenge 2, to: seawater (SW); *V. tasmaniensis* LGP32 (LGP), Scottish isolates M1 or M40 at $4.4 \times 10^9$ colony forming units (CFU) ml$^{-1}$ (M1, M40); heat-killed M1 (HK) at $3.5 \times 10^9$ CFU ml$^{-1}$; or to $10^6$ CFU ml$^{-1}$ of M1 (M1L) or M40 (M40L). Mussel challenge groups were divided between 2 x 5 L beakers, from one of which individuals were sampled (“Sampled”) while mussels in the other beaker (“Undisturbed”) were left undisturbed until the end of the challenge.

Figure 28 *Vibrio* level in mussel tissue calculated as a function of *ompU* DNA (vibrio DNA) relative to 18S rRNA DNA (mussel DNA), both measured by real-time PCR with an upper crossing-point (Cp) limit of 40. Ratios are plotted as fold-change (± SD) compared with the average values for seawater-exposed mussels at each sampling time. Samples are: HK, heat-killed vibrio strain M1 (white bars); M1, vibrio strain M1; M40, strain M40; LGP. *Vibrio tasmaniensis* LGP32; Lo1, $10^6$ colony forming units (CFU) ml$^{-1}$ M1; Lo40, $10^6$ CFU ml$^{-1}$ M40. Samples were collected at 3, 24, 48 and 144 h post-challenge. * indicates a difference at a significant level from SW controls.
groups were higher than in mussels exposed to seawater. By 48 hours post-challenge, the level of *Vibrio* DNA had dropped in the mussels exposed to heat killed bacteria and it was not significantly different from the group exposed to seawater. At 144 hours post-challenge, the levels of *Vibrio* DNA in the M1, M40, LGP32 and M1L groups remained significantly higher than in the seawater exposed group.

Differences of gene expression in gill tissue between the mussel groups exposed to full strength bacterial concentrations or to seawater, were detected for IL-17, MyD88 and IKB1 (Figure 29). At 144 hours post-challenge, IL-17 was down-regulated in mussels exposed to LGP32, heat-killed M1 (HK) and M40, compared with seawater-exposed controls. The expression of MyD88 was down-regulated at 24 hours post-challenge in mussels exposed to M1 and M40, compared with seawater controls while, at 48 hours post-challenge, MyD88 expression was down-regulated in LGP32 and M40 groups compared with seawater controls or with mussels exposed to M1. Also at 48 hours post-challenge IKB1 was downregulated in LGP32 and HK compared with seawater and M40-exposed groups.

The expression of TLR, LBP, mytilin and defensin were examined at 3 and 24 hours post-challenge only (Figure 30). At 3 hours post-challenge, TLR expression was upregulated in mussels exposed to LGP32 compared with seawater controls. However, at 24 hours post-challenge, TLR expression was higher in the heat kill group than in the seawater, LGP32 or M1 groups. No differences in expression between the challenge groups were detected for LBP or mytilin. Only some of the samples were successfully amplified using defensin and, consequently, results for this gene were not further analysed.

5.3.5 Discussion

Our initial screening had found considerable genetic differences between Scottish shellfish *Vibrio* isolates in occurrence and apparent sizes of the two virulence genes, *ompU* and *vsm* (McCarthy and Chapuis 2015). This variation supports the findings of incongruence between the two genes, reported by Nasfi et al (2015). However, unlike Nasfi et al (2015) who found no similarity in the virulence gene sequences between any of their isolates and LGP32, 3 of the 4 Scottish isolates which were the focus of the current study showed 100% identity with the gene from LGP32 across a stretch of 376 bases (36% of the gene). Given this similarity between the Scottish *Vibrio* and pathogenic *V. tasmaniensis* LGP32, it was surprising that we were unable to find any evidence of *Vsm* gene in these Scottish isolates, despite targeting conserved regions. In addition, no protease activity was found using an azocasein digestion assay, which has been used to demonstrate the activity of the *Vsm* protein in *V. tasmaniensis* LGP32 (Binesse et al, 2008). The *Vsm* gene also appeared to be absent in M1 isolate, which shares 100% identity in part of the *ompU* gene with isolate M11 in which *vsm* gene is present. The M1 and M11 isolates were isolated from 2 separate mussel populations which had both experienced elevated mortality during the latter half of 2010. These populations
Figure 29 Expression of a) IL-17, b) MyD88 and c) IKB1 mRNA in *M. edulis* gill tissue at 3, 24, 48 and 144 h following exposure to seawater (SW), *V. tasmaniensis* LGP32 (LGP), Scottish *Vibrio* isolates M1 or M40 or to heat-killed M1 (HK). Expression is shown as the Log10 transformed crossing point (Cp) value for the gene of interest, normalised against the reference gene, elongation factor-1α, for each individual. Points are slightly horizontally jittered for great clarity. Significant difference between groups is shown as: p<0.05,***; p<0.01,****; p<0.001,*****. A flat line indicates similarity between groups. A line with tick marks at each end is used to indicate groups which differ.
Figure 30 Expression of a) TLR, b) LBP and c) mytilin mRNA in *M. edulis* gill tissue at 3 and 24 h following exposure to seawater (SW), *V. tasmaniensis* LGP32 (LGP), Scottish *Vibrio* isolates M1 or M40 or to heat-killed M1 (HK). Expression is shown as the Log10 transformed crossing point (Cp) value for the gene of interest, normalised against the reference gene, elongation factor-1α, for each individual. Points are slightly horizontally jittered for great clarity. Significant difference between groups is shown as: p<0.05,*; p<0.01,**; p<0.001,***. A line with tick marks at each end is used to indicate groups which differ.
were located in 2 enclosed sea lochs, with a 20 km separation between their openings onto the main seawater channel. By comparison, the isolates M31, M36 and M40 were all cultured from mussel populations located 50 km away, at the head of the main channel. *Vibrio* isolates in which *vsm* gene sequence was present were also cultured from these mussels. This raises questions about the processes governing the retention or loss of the *ompU* and *vsm* genes in the *V. splendidus* or *V. tasmaniensis* species, spread of vibrios in marine environment and supports findings of a lateral gene transfer described by Nasfi et al (2015).

Previous *in vivo* challenges of adult shellfish have involved the injection of bacterial preparations directly into the adductor muscle (Costa et al 2009; Rosani et al 2015; Toubiana et al 2013). Although this route allows controlled delivery of a known challenge dose, it bypasses physical and peripheral immune defences and so these challenges may be less reflective of the natural situation (Allam and Raftos 2015; Allam and Pales Espinosa 2016; Bruto et al 2016) in which mussels are immersed in an aquatic environment containing bacteria, and exposed to bacteria through filter-feeding. By using bath challenges, the present study aimed to replicate more typical exposure routes. However, in order to make our challenge comparable to other published studies and much the dose strength achieved by an injection (Balbi et al 2013; Ciacci et al 2009), mussels in the current study were exposed to higher bacterial concentrations than the levels reported in marine environments (Vezzulli et al 2010; Hervio-Heath et al 2002).

In first challenge, the vibrios used for challenge had largely been cleared from the tissues of sampled mussels within 24 hours period and they were no longer detectable after 48 h while in the second challenge 2, the concentration of the live challenge strains increased during the first 24 hours. As the PCR test used to detect vibrio levels was based on extracted genomic DNA rather than RNA, it can be established conclusively if the increased levels resulted from the bacterial replication. However, our observation of the *OmpU* DNA level in the heat kill-treated group of mussels did not increase during the first 24 hours and decreased to the same level as in seawater-treated group by 48 hours post exposure, does suggest that the higher DNA levels observed in the live challenge groups were due to an increase numbers of the viable challenge strain. Despite being held without water exchange and even with an initial increase in numbers, it appears that the mussels were capable of limiting or decreasing levels of the challenge bacteria within 48 hours post challenge. The findings of both challenge experiments suggest that juvenile or pre-adult mussels are capable of efficient clearance of the trialed *Vibrio* isolates from tissues, even when compromised or held under sub-optimal conditions. This timescale of reduction was in line with the previous findings (Parisi et al 2008; Balbi et al 2013).

No meaningful information about the effect of exposure to vibrio on mortality rates was obtained in challenge 1. Although the causative agent responsible for observed mortalities in mussel stock used was unknown, large numbers of the copepod, family *Monstrillidae* which are known endoparasites of marine invertebrates (Suárez-Morales et al 2010), were observed in the water of the acclimatisation tanks. Our analysis indicated their occurrence in 45% of the population (not shown) and histological examination revealed the existence of multiple parasites within mussel tissues. Despite the lack of evidence for associated pathological changes, it was speculated that the presence of these copepods in such large numbers might have played a role in the mortalities (Ben Cheikh et al 2017). While it is accepted practice to use healthy animals in challenge experiments, unfortunately the identity of the copepodits was unknown prior to start of experimental challenge.
In challenge 2, the occurrence of mortalities in the mussels exposed to autoclaved bacterial culture may be a consequence of chemical or pH changes induced by heating the broth for the extended time required to ensure complete inactivation of the bacteria. Alternatively, heating may not fully have inactivated toxic components produced by the bacteria (Lane & Birkbeck 1999; Labreuche et al. 2006) as almost 50% of residual proteolytic activity remained for example following heating of *V. tubiashii* extracellular products to 100°C for 13 min (Mersni-Achour et al 2014). There were some discrepancies observed in the second challenge trial. For example there was also considerable variability in mortality reported between replicates for the LGP32-treated groups, and also mortality in the low-dose M1 challenge group was higher than in the groups exposed to full-strength concentration of the same strain. Overall mussel mortality occurred only in the mussel groups exposed to bacteria and the level of mortality was low (maximum 20-22.9 %). This is significantly lower that observed in other studies (for example Ben Cheik et al 2017) where 45% mortality was observed in a cohabitation challenge with *V. splendidus*. Consequently, it is not possible to make any reliable deductions about the likely virulence of the Scottish vibrio isolates in *M. edulis* and further work is required.

As an alternative approach to assessing the effects of the Scottish vibrio isolates on mussels, expression of a number of immune-related genes was examined following the challenge exposure. The results indicated that mussels were responding to the different exposures however the most frequent differences were observed in mussels exposed to *V. tasmaniensis* compared to seawater exposure. Except for TLR, which elicited an upregulation compared with SW controls, other studied genes (MyD88, IKB1 and IL-17) were found to be downregulated in the mussels exposed to bacterial agent. The importance of including a control containing inactivated bacteria, rather than simply seawater, is evident from the fact that exposure to heat-killed M1 was capable of modulating gene expression. For example the inactivated form of M1 (HK) appears to elicit a stronger response than the live bacterium in the same challenge experiment. These discrepancies might be explained by temporal differences in response or by individual variation.

Immune responses and the expression of immune-related genes differ depending on the bacterial species used for challenge, even within the *Vibrio* genus (Cellura et al 2007; Ciacci et al 2009; Mitta et al 2000). In *M. edulis*, one study found that exposure *in vitro* to *V. splendidus* LGP32 elicited stronger or earlier responses in haemocytes than did a non-pathogenic environmental strain of *V. splendidus* (Tanguy et al. 2013) while a later study on the same samples found significant changes induced by the environmental strain instead (Tanguy et al 2018). Other studies (Cellura et al 2007; Ciacci et al 2009; Costa et al 2009; Mitta et al 2000; Philipp et al 2002; Toubianna et al 2013; Tanguy et al 2018) have reported responses in the first hours after exposure, with some genes, such as mytilin, exhibiting peak expression at 2-3 hours post-challenge. As the first samples in the current study were obtained at 3 h, it is expected that early changes in gene expression would have been detected. The lack of differences in expression of mytilin and LBP between the different challenge groups may have been a result of the greater variation in response observed between individuals for these genes in comparison with MyD88 or IKB1. A high degree of variability has been detected in mussel immune responses, with polymorphism in the nucleotide sequence of antimicrobial peptide genes and a considerable number of non-responders, and this variable response can be further affected by environmental conditions as described by Tanguy et al 2018; Cantet et al 2012; Costa et al 2009).
Many of the previous studies have been carried out on isolated haemocytes exposed \textit{in vitro} (Tanguy et al 2013, 2018), \textit{in vivo} (Mitta et al 2000; Toubiana et al 2013) or with heat-killed bacteria (Cellura et al 2006; Ciaccia et al 2009). The complexity of the responses in the present study may reflect the fact that the whole animals were exposed instead, through bath challenge with live bacteria, and measured gene expression in a tissue, rather than in individual cells. While a direct bath challenge is an artificial situation, and the level of bacterial challenge is likely higher than would occur as a singular event in nature, this approach does not bypass the peripheral defences, such as mucus, or the pathogen recognition mechanisms of mucosal tissues (Allam et al 2015, 2016). Inclusion of an inactivated culture as a control allowed us to detect effects which were not due solely to bacterial colonisation and invasion. From the immune responses, it appears that exposure to the Scottish \textit{Vibrio} strains or to their ECPs was capable of exerting a suppressive effect on the mussel immune responses. The outcome of the challenge experiment also suggests that the Scottish isolates are capable of causing limited mortality in \textit{M. edulis}, at a level similar to or less than LGP32. As this strain is not considered pathogenic to mussels (Balbi et al. 2013), the overall conclusion is that the Scottish \textit{Vibrio} isolates M1, M31, M36 and M40 are not, in themselves, a direct threat to \textit{M. edulis} production but have the potential to contribute to mortality under certain environmental conditions or when mussels are stressed.

5.3.6 References


Allam, B., Pales Espinosa, E., 2016. Bivalve immunity and response to infections: Are we looking at the right place? Fish and Shellfish Immunology 53, 4-12.


5.4 King scallop (\textit{Pecten maximus}): Characterization of baseline microbiome of healthy king scallop (Partner 6)

5.4.1 Background
In contrast to Atlantic salmon and blue mussel, the value of the king scallop, \textit{Pecten maximus}, within the Scottish economy is relatively small, with an annual farmed production for the table of 30-80,000 tonnes since 2009 (Scottish Shellfish Production Survey 2016). Understanding the interaction between the environment and microorganisms and their effects on growth and health of farmed shellfish is fundamental to sustainably maximising the production of any farmed animal. It is likely that the microbiota of shellfish will play an important role in their biological processes and, through improved understanding of the roles of different bacterial species, it may prove possible to manipulate these to our advantage in shellfish cultivation. On the other hand, presence of microorganisms in the environment and the ability of shellfish to accumulate them might be leading to disease emergence. Very little is known about scallop diseases and more recently the application of next generation sequencing contributes to description of new scallop pathogens such as \textit{Endozoicomonas} spp originally reported as rickettsia-like organisms (RLOs) world-wide (Cano et al. 2017). In conducting diagnostic investigations during shellfish mortality events, it can be difficult to identify putative pathogens, as little is known about the expected microbial profile of healthy animals. From this perspective, the ability to detect a state of general dysbiosis, as a consequence of pathogen overgrowth or through environmental stress, may be a useful diagnostic tool. For routine screening, it may be less onerous and more cost-effective to detect disease onset as a departure from ‘normality’ rather than by attempting to isolate and identify specific pathogens. Consequently, there is a requirement to collect information about the microbial communities of healthy animals in order to establish what are the baseline community structures.

5.4.2 Aims of the study
Aim was to establish baseline microbiome of healthy king scallop throughout the year as a reference against which to compare community profiles which may arise as a result of disease or increased environmental stress.

5.4.3 Material and Methods
\textit{Collection of samples for king scallop microbiome characterization and DNA extraction}
Scallop samples were collected in July, September and November, when seawater temperatures at 5 m were 12.4, 13.1 and 8.7°C respectively. Scallops were collected by a diver, and removed from the water just before tissue sampling. Scallops were opened by standing the shell on end, hinge uppermost, and making a single pass with a sterile scalpel to sever the adductor muscle where it attached to the flat shell. Samples, approximately 0.5 cm$^3$, were collected aseptically from gill, mantle, digestive gland and adductor muscle from 10 scallop individuals in each sampling season. Tissues were placed into individual sterile, DNA/RNA-free Eppendorf® Safe-Lock microcentrifuge tubes and placed immediately on ice. Numerous attempts were made to collect haemolymph, which were unsuccessful. Samples of seawater (1 L) and of sediment (2 x 50 ml) were collected at the same time as the shellfish. Seawater was filtered using a 1 L bottletop filter unit, and the filter removed aseptically and placed in a sterile 50 ml centrifuge tube. Immediately following collection, all samples were place
into chilled cool boxes for transport back to MSS (approximately 6 h journey time) where they were transferred to storage at -80°C until further processing.

A number of methodological approaches to extract genomic DNA was tested to optimise the yield and quality of extracted DNA. The kits included DNeasy Blood and Tissue Kit, QIAamp Fast DNA Stool Mini Kit (Qiagen), MasterPure™ DNA Purification Kit and E.Z.N.A.® Mollusc DNA Kit (VWR). Tissue was weighed before homogenisation in lysis buffer (20 mM Tris·Cl, pH 8.0, 2 mM sodium EDTA, 1.2% Triton® X-100) with a 5 mm steel bead, using a Tissue Lyser (Qiagen, UK). The homogenate was incubated with lysozyme (20 mg ml⁻¹ final concentration) at 37°C for 60 min to digest the cell wall of Gram-positive bacteria. The homogenate was then incubated overnight at 56°C with Proteinase K, in the respective buffers from the different kits, and the rest of the procedure was conducted following the manufacturers' instructions including an RNase digestion step (5 µg ml⁻¹ for 30 min at 37°C). All procedures were carried out where possible within biosafety cabinets, adopting strict aseptic technique to minimise the introduction of contaminating DNA from personnel or from consumables. The DNA concentration and quality (ratios of absorbance at 260/280 and 260/230 nm) were measured on a NanoDrop™ (ThermoFisher). Because of the difficulty in obtaining reproducible and high quality DNA yields, attempts were made to isolate DNA from several tissues including adductor muscle, mantle and haemolymph. However, the best results were obtained with gill (G) and digestive gland (DG).

Filters were dissected aseptically into quarters and two separate extractions were carried out. Filters were homogenised in 350 μl EZNA kit lysis buffer or 180 ATL buffer (Qiagen) for 2 x 1 min at 25Hz in a Tissue Lyser with a 5 mm steel bead (first extraction) or 1 mm glass beads (second extraction). Proteinase K was added and samples were shaken again for 20 s to mix before incubating overnight at 56°C. Samples were shaken for 1 (first extraction) or 2 min (second extraction) and the rest of the procedure was carried out as described above.

Sediment samples were vortexed vigorously to ensure homogeneity and 2 subsamples collected immediately. These samples were again stored briefly at -80°C until processing. Samples were thawed rapidly using vigorous vortexing and held on ice for 15 min to allow settlement of the larger sediment. Three aliquots of the supernatant were centrifuged at top speed to pellet the finer sediment and bacteria. Pellets were combined to make 2 samples from each month. The pellets were incubated in TET lysis buffer with lysozyme, as above, before extracting with a MoBio PowerBiofilm kit (Qiagen) according to the manufacturer's instructions.

Extraction controls were samples containing only reagents and processed in parallel with tissue samples. Purified DNA was stored at -80°C until it was shipped for analysis.

King scallop microbiome analysis
Once extracted, purified DNA was sent to RTLGenomics, Lubbock, Texas, US for Illumina MiSeq Paired End Sequencing of the bacterial 16S rRNA gene (V3 and V4 regions). Extracted DNA assessed for quality and those from the four individuals from each timepoint with overall highest scores for both tissues (G and DG) were put forward for sequencing (24 samples in total). The resulting forward and reverse reads, provided as paired fastq files, were analysed using mothur v1.40.0 and following the MiSeq SOP (Schloss et al 2009; Kozich et al 2013) with minor modifications: namely sequences were aligned against the SILVA v.132 database version 1.19 (Quast et al 2013), and sequences belonging to Archaea, chloroplasts and mitochondria were
removed. Taxonomic classifications were made using both the Silva and RDP training sets. Sequences from scallop tissues were also analysed using DADA2 version 1.7.8 according to the pipeline tutorial (https://benjjneb.github.io/dada2/tutorial.html), following demultiplexing and removal of barcodes using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). No ‘mock’ group was used but, following taxonomic assignment, operational taxonomic units (OTUs) present only or principally in extraction control samples were removed from the data after manual screening. These denoised data from both Mothur and DADA2 pipelines were imported to the phyloseq package (http://joey711.github.io/phyloseq/) in R, for further analysis.

5.4.4 Results
Yields of king scallop DNA using all extraction methods trialed were low and of poor quality, typically below 20 ng µl⁻¹ and with low 260/280 and 260/230 nm ratios. For scallop tissue, the EZNA kit gave the highest yields with 260/280 ratios close to 1.8, although yields were highly variable: digestive gland 17.6 ± 9.3 ng µl⁻¹ and gill 15.0 ± 9.4 ng µl⁻¹. Successful extraction was achieved only with one of the filter quarters, using the EZNA kit, but DNA yields were low with results for July, September and November of 5.6, 20.1 and 1.4 ng µl⁻¹ respectively. Yields from sediments ranged from 4.4 to 11.7 ng µl⁻¹. The quality of available DNA affected the number of suitable samples available for analysis of microbiome in healthy king scallops. Paired reads were obtained from gill and digestive gland of 4 scallops, 2 sediment samples and 1 seawater sample for July, September and November, and for the respective extraction controls.

A total of 684463 unique sequences were obtained, yielding 233365 sequences after alignment and filtering in mothur software pipeline. Clustering at 97% identity resulted in 605633 sequences. Following manual inspection and removal of OTUs identified as contaminants, the final number of sequences was 378453. Using DADA2, the complete set of fastq files, from scallop, sediment and seawater combined, failed to pass the filtering step (“filterAndTrim”), despite increasing the maxEE parameter to 6 for reverse reads. It was possible to complete the filtering step by processing the sequences from the environmental samples (water and sediment) separately from the scallop tissues, but sequences from the environmental samples had higher error rates. Consequently, DADA2 processing, including chimera removal and classification against the Silva v.128 database was carried out for sequences from scallop tissues alone.

Prior to importing to phyloseq for statistical analysis, the datasets were screened to identify OTUs present in extraction control samples. The number of sequences deemed to be contaminants was 227180, almost 38% of the total and representing 6166 OTUs. The majority were belonging to the Order Burkholderiales, members of the Betaproteobacteria. Following the removal of these sequences, the mean number of sequences per sample type was: digestive gland 10,270 ± 3355; gill 3862 ± 2874; sediment 18312 ± 5326; seawater 18988 ± 4471. Thus, of the total number of sequences obtained, contaminant sequences comprised ca. 5% of sediment, 17% of seawater, 19% of digestive gland and 75% of gill sequences.

Analysis of alpha-diversity on non-rarefied and rarefied data showed a similar pattern, of greater richness (Figure 31A) and diversity (Figure 31B) in sediment and seawater, compared with scallop tissues. However, for both scallop tissues and for seawater, richness and diversity were higher in November than in July or September, while the situation was reversed for sediment. In November, the bacterial community
from the scallop digestive gland was more diverse than from the gill (Figure 31C). The bacterial population of the digestive gland from one individual in September can be observed as low diversity on both Chao1 and Shannon analyses.

Both NMDS and principal component analysis (not shown) ordinations suggest a degree of separation between the scallop bacterial communities and sediment samples, and both illustrate a degree of separation between scallop samples from July and November (Figure 32.) The most abundant classified phyla in scallop tissues are shown in Figure 33 (for digestive gland and gill) and in Figure 34 for sediment and seawater. Across both tissues and the 3 sampling times, Cyanobacteria, Planctomycetes and Proteobacteria are the most consistently abundant. Cyanobacteria are highly abundant in both digestive gland (D) and gill (G) in July (7) but abundance decreases in the subsequent months. Proteobacteria are more abundant in gill than in digestive gland, which is also the case with Actinobacteria. In September (9), Firmicutes are the most abundant in digestive gland and are also abundant in gill, but are at low abundance in July and November (11), while Verrucomicrobia become more abundant in November. The variation between individuals (not shown) is considerable but the patterns for the months are also found in individuals.

In both sediments and seawater, Bacteroidetes and Proteobacteria are the most abundant phyla. However, Cyanobacteria are the next most abundant phylum in seawater while Actinobacteria, Firmicutes and Epsilonbacteraeota are the next abundant in sediment. Epsilonbacteraeota were proposed as a new phylum in 2017 (Waite et al 2017) but more recently, this has been modified to Campylobacterota the phylum of the class Campylobacteria (Waite et al. 2017). Planctomycetes are not appearing among the more abundant phyla in seawater but are being detected at < 1% (not shown) in sediment.
Figure 31. A. Alpha diversity of bacterial communities in the four sample types, scallop digestive gland (Dig), gill (Gill), sediment (Sed) and seawater (SW) using Chao and Shannon diversity measures on data rarefied to the lowest number of sequences (615, Gill I, Nov). B. Sample richness in each sample type, digestive gland (D), gill (G), sediment and seawater in July (month 7), September (9) and November (11). The data points are jittered slightly to aid visualisation. C. Alpha diversity in digestive gland (D) or gill (G) tissues from individual scallops (a – l), in July, September and November. Chao and Shannon diversity were calculated on non-rarefied data.
Figure 32 Non-metric multidimensional scaling (NMDS) ordinations of Bray-Curtis similarity matrix indicating A. differences in July (month 7), September (9) and November (11) for scallop or sediment/seawater samples or B. differences between the sample type, scallop digestive gland (D), gill (G), sediment (Sed) or seawater (SW).

Figure 33 The most abundant classified bacterial phyla detected in healthy king scallop tissue
5.2.5 Discussion
MiSeq analysis of bacterial species present in a sample requires the template DNA to be of a minimum quality, relatively free of substances which would inhibit amplification and sequencing. These inhibitors may come from the reagents used in extraction, eg. phenol or guanidium, or from the matrix in which the bacteria are contained. Molluscs are rich in inhibitory compounds, polyphenolic proteins and polysaccharides, and different approaches have been developed to try and separate sequencing quality DNA from these inhibitors. In order to study the bacterial species associated with *P. maximus*, we sought an extraction method which yielded a minimum DNA concentration of 20 ng/μL with an A260/280 ratio of 1.8, as measured by Nanodrop, and which was suitable for relatively high throughput processing. None of the methods tested in the present study consistently provided DNA of the required quality. The decision to use EZNA E.Z.N.A.® Mollusc DNA Isolation Kit (Omega Biotek) which is based on cetyltrimethylammonium bromide (CTAB) and uses chloroform to remove mucopolysaccharides was because this appeared to yield DNA of satisfactory quality on practice samples; however yields from test samples proved highly variable and 6 of 30 digestive gland samples failed to amplify during pre-sequencing amplification. The best indicator of likely success in pre-sequencing amplification appeared to be an A260/230 ratio greater than 1. Further work is required to develop an optimised extraction method which yield DNA of a consistent quality.

Another area requiring attention was the large number of reads which apparently originated from contaminating bacterial DNA introduced or present during the extraction process. The issue of contaminating DNA in sequencing studies is well-recognised and documented, and various methods have been proposed to address it (Laurence 2014; Salter 2014). To limit the introduction of contamination during sample processing, we conducted all our extractions under laminar flow within Class II biosafety cabinets, using strict aseptic technique and sterile or freshly-opened reagents. In addition, we prepared extraction controls which contained no sample material and which were processed through the extraction procedure in parallel with samples. Outputs from Mothur and DADA2 pipelines were examined to identify the OTUs which were only present or were predominantly abundant in these extraction controls. In the Mothur output, the number of reads in these extraction controls was 6.2% of the total reads following denoising. The proportion of reads in the OTUs present only or principally in the control samples was almost 50% of the total reads of which almost all were classified as Burkholderiales. Results were compared both

---

**Figure 34** The most abundance bacteria phyla detected in sediment and seawater collected in proximity of king scallop collected for the present analysis
including and excluding these OTUs, but more detailed statistical analyses were performed in the absence of these OTUs. Members of this order have been identified as contaminants in other reports, where they are believed to be present in extraction kits (Laurence 2014; Salter 2014).

The principle phyla identified in the scallop tissues are similar to the findings of other groups using high-throughput sequencing to examine bacterial communities in a range of Pacific oyster species (Trabal Fernandez et al, 2013; Wegner et al 2013). Our finding of high abundance during July in both scallop tissues and seawater of Cyanobacteria, principally Synechococcus, mirrors that of Ossai et al (2017) in American oysters from Chesapeake Bay. Contrary to our expectations, based on the prevalence of Vibrio spp. in cultures from shellfish samples, we detected low abundance of Vibrio sp. which has also been reported by both Ossai et al (2017) and Wegner et al (2013). Although the principal bacterial phyla are different from those in the Manila clam, reported by Milan et al (2018), this group observed greater microbial richness in winter months which is also supported by our findings. Further analysis is ongoing, to discriminate the contribution of bacterial classes and orders to the community profile at the different sampling periods, with a view to interpreting their significance.

5.4.2 References

EMBRIC – D8.2: Report with industry standards and instructions for molecular tools in paternity assignment, traceability and marker assisted selection, page 76 of 76