Norwegian University
of Life Sciences

Master's Thesis 2023 30 ECTS
Faculty of Bioscience

Machine learning approaches using freshwater gene expression profiles to predict seawater performance in Atlantic Salmon

Sofie Robertsen
Integrated teachers program in Science (Biology)

Machine learning approaches using freshwater gene expression profiles to predict seawater performance in Atlantic Salmon

Bruk av maskinlæring til å forutse tilvekst i sjøvann ut i fra genutrykk i ferskvannsfasen hos laks

Author:
Sofie Robertsen

Supervisors:
Simen Rød Sandve (Main supervisor)
Torgeir Rhoden Hvidsten (Co-supervisor)

Master thesis
Integrated teachers program in Science (Biology)
30 credits

Norwegian University of Life Science
Faculty of biosciences

Acknowledgment

The work presented in his thesis has been carried out at the Department of Bioscience at the Norwegian University of Life science in the period of October 2022 to May 2023.

I will first and foremost like to thank my main supervisor Simen Rød Sandve for your guidance this last year. Thank you for your encouragement and expertise throughout this process. Your sense of humor and positive energy have been especially appreciated. I also want to thank my co-supervisor Torgeir Rhoden Hvidsten for help and guidance with the machine learning. Furthermore, I would like to thank the rest of the Sandve-group for sharing your scientific expertise and for making me feel included in the group from the start. A special thanks to Arturo Vera Ponce de Leon for helping me with my coding problems.

A warm thanks to all of my fellow students, especially "Lektordamene" for all the good memories and for making these last years so fun!

Finally, thanks to my family and Kita for your support and for always believing in me. Thanks to all my friends for your support and encouragement.

Oslo, May 2023
Sofie Robertsen

Abstract

Atlantic salmon (*Salmo salar*) has an anadromous life cycle, spending the first part of its life in freshwater before migrating to seawater. Smoltification is the process where Atlantic salmon undergo several morphological, physiological and behavioral changes preparing for transition to marine environment. A major challenge in the Norwegian salmon farming industry is the high mortality (12-14%), after release of smolt into seawater. One reason is suboptimal smolt production, resulting in a state where salmon are not well adopted for life in seawater. It is therefore important to optimize smolt production protocols and develop better ways to assess seawater-readiness to ensure higher survival, growth and reduce welfare issues. Traditionally, the increased expression of the saltwater isoform nkaα1b and nkcc1a cotransporter, and a reduction in expression of the freshwater isoform nkaα1b in the gills are used as predictive markers for seawater-readiness in the salmon farming industry. The current study aimed to use Random Forest to build predictive models for growth in seawater based on gill transcriptome data from fish given different light manipulation during smolt production. The results showed poor predictive ability towards seawater growth, although superior to simple correlation with single gene expression levels. We also found that photoperiodic history had effect on the Random Forest predictions, where the Random Forest model from fish exposed to continuous light (24:0) was much better at predicting SW growth than any of the models from the fish exposed to short photoperiods (8:16 and 12:12). We extracted most influential genes for each Random Forest model and found that these differed depending on the light regime used. Based on these results the salmon farming industry should apply caution when relying on traditional smolt gene-expression markers to determine the optimal time for SW transfer.

Sammendrag

Laks (*Salmo salar)* er en anadrom fisk som tilbringer den første delen av sin livssyklus i ferskvann. Smoltifisering er prosessen der en laks gjennomgår flere morfologiske, fysiologiske og atferdsmessige endringer i forbindelse med overgangen fra ferskvann til sjøvann. En av de store utfordring i norsk lakseoppdrett er tap av fisk (12-14%) etter utsetting av smolt i sjøvann. Suboptimal smoltproduksjon er én av årsakene til dette, da laksen som et resultat ikke blir tilpasset et liv sjøvann. Viktig er det derfor å optimalisere smoltproduksjonsprotokoller, og utvikle bedre metoder for å vurdere riktig tidspunkt for sjøutsetting av smolt. En slik optimalisering vil kunne sikre høyere overlevelse og vekst, samt redusere velferdsproblemer hos laksen. Noen oppdrettere anvender molekylære markører for å evaluere «riktig» tidspunkt for sjøutsettelse, men disse har vist seg ikke å være optimale. Formålet med denne studien er å bruke Random Forest algoritmen i utviklingen av prediksjonsmodeller som måler tilvekst i sjøvann. Modellene er basert på data fra gjelle-transkriptomet fra fisk som i smoltproduksjon ble behandlet med ulike lysprotokoller. Studiens resultater viste at modellene var mindre egnet i å predikere tilvekst i sjøvann. Likevel var predikasjonen bedre enn korrelasjonen mellom tilvekst og genuttrykket til enkeltgener. Videre funn viste at lysbehandling påvirket predikasjonene, der modellen for kontinuerlig lys (24:0) ga best predikasjon, sammenlignet med modellene basert på vintersignal (8:16, 12:12). De mest innflytelsesrike genene for hver modell ble identifisert, også disse var påvirket av lysprotokol. På bakgrunn av studiens funn, bør industrien vise forsiktighet med å anvende tradisjonelle smolt-genuttrykksmarkører i vurdering av optimalt tidspunkt for sjøutsetting.

Table of content

1 Introduction

Since its start in the 1960s the Norwegian salmon farming industry has grown steadily and has within the last 50 years become the most important export industry next to oil and gas. In 2022, a total of 1.255 tons Norwegian salmon was exported to a value of 105.8 billion NOK (Norwegian Seafood Council, 2022). Today Norwegian salmon farms are the biggest producers of Atlantic salmon (*Salmo salar*) in the world (Nærings- og fiskeridepartementet, 2021).

A major challenge in the Norwegian salmon farming is high mortality (12- 14%), after release of smolt into seawater (SW) (Hjeltnes et al., 2018). One reason is suboptimal smolt production, resulting in salmon that are not well adopted for life in SW with higher salt concentration and exposure to new pathogens (Stefansson et al., 2005). To improve smolt robustness, it will be important to optimize smolt production protocols, including developing new and better ways to assess the SW-readiness and ensure higher survival, growth, and reduce welfare issues.

In this thesis we leverage data from a large smolt-experiment undertaken in 2021-2022 (FHF project #901589) to explore the use of gene expression data as predictive markers for salmon SW-performance. The experiment included three different smolt production protocols, generation of gill transcriptomes from 3000 fish at time of SW transfer, as well as tracking of individual fish phenotypes in a common garden SW-pen. In this thesis I aim to use machine learning on the gill transcriptome data to predict growth later in life on the same fish, and also identify which genes that contributes to the predictions. This knowledge can be used to help develop better smolt production protocols and enhance animal welfare and aquaculture sustainability.

2 Background

2.1 Life cycle of Atlantic salmon

Atlantic salmon is an anadromous fish, meaning that it begins its life in freshwater (FW) river systems before migrating to SW (Hoar, 1988). Atlantic salmon usually spend 1 to 6 years in FW before migrating to SW. Mature Atlantic salmon return to their river of origin, mate and lay eggs in late autumn, and the eggs hatch the following spring (Fleming, 1996). Immediately after hatching, the salmon is known as alevin, before becoming a fry that quickly grows and develops into a parr. The parr is characterized by dark marking along the side of the body and several factors influence the length of the parr stage, including size, growth rate and metabolic status (Rowe et al., 1991; Thorpe, 1994; Thorpe et al., 1998). Parr exceeding a certain size threshold in the autumn tend to undergo the smoltification process, whereas smaller parr tend to remain in the parr stage (Kristinsson et al., 1985; Thorpe et al., 1982). The adult salmon spend one to five years at sea before using geomagnetism and olfaction to guide their way back to native rivers (Hasler et al., 1978; Keefer & Caudill, 2014).

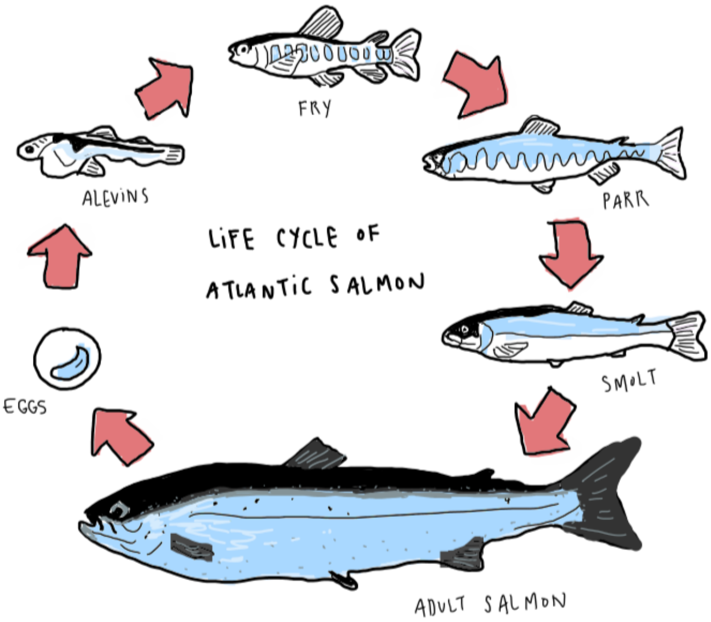

Figure 2.1 Life cycle of Atlantic salmon. Different developmental stages

2.2 Smoltification

Smoltification, hereafter referred to as smolting, is a natural process that occurs in juvenile salmon, where they undergo a series of changes in their behavior, morphology, and physiology to prepare for life in SW. This transformation is essential for their survival, as they need to adapt to the changes in the osmotic environment when they move from FW to SW (Stefansson et al., 2008). The process of smolting is under internal control by the neuroendocrine system and the timing is controlled by increasing daylight and water temperatures during spring (McCormick et al., 1995). The process of smolting is initiated by stimulation of the light-brain-pituitary axis, in a response to increasing daylength (Ebbesson et al., 2003). This results in a gradual series of physiological changes in parr that ultimately lead to adaptation for survival in SW habitats (Björnsson et al., 1989; McCormick et al., 1995). As smolting proceeds, the salmon become silvered, with a darker shade on their dorsal side and a brighter hue ventrally (Johnston & Eales, 1967; Staley & Ewing, 1992). The growth pattern also changes, turning into a more elongated body shape (Wedemeyer et al., 1980). And importantly, salmon smolts develop improved ability to tolerate high salt concentrations through remodeling the gill physiology, including expression of new genes encoding proteins involved in chloride excretion.

2.2.1 Osmoregulation

Atlantic salmon encounter distinct osmoregulatory challenges due to differences in salinity in FW and SW habitats. Salmon is hyperosmotic to FW, which means they have a higher ion concentration in their bodies than their surroundings, and as a result they face the risk of losing ions and gain water by passive diffusion and osmosis. To counteract this, the salmon actively take up ions and get rid of excess water by producing a large quantity of diluted urine. In SW, salmon is hypoosmotic to the sea water, where the internal extracellular fluids have lower salt concentration than the external environment. This leads to a state where salmon lose water and gain ions. To reduce dehydration SW teleost increase their drinking rate (Perrott et al., 1992). Osmoregulation in FW and SW requires cooperative effort of the gills, intestine and kidney of which the gill is the most studied (McCormick, 2012). In the gills, ion transport is carried out by specialized cells termed ioncytes, also called mitochondrial-rich cells (MRC) (Wilson & Laurent, 2002). Osmoregulation occurs across MRCs and their function and morphology differ between SW and FW (Evans et al., 2005; Hiroi & McCormick, 2012; Hwang & Lee, 2007). During smolting there is a notable increase in number and size of MRCs and the appearance of the associated accessory cell The MRC develops an extensive tubular system that is characteristic for SW-MRCs, continuous with the basolateral membrane giving a large surface area for transport proteins (Pisam et al., 1988).

Several genes have been demonstrated to be a part of the ion-excretory system in SW gills, cystic fibrosis transmembrane conductance regulator (CFTR), Sodium/potassium/chloride cotransporter (NKCC), and the sodium-potassium ATPase (NKA). NKA and NKCC are localized in the basolateral membrane, while the CFTR is localized in the apical membrane. Salinity tolerance is accompanied by increased activity of several ions transporters in the gill were NKA is the most studied (McCormick, 2012). It is an established marker for smolt status because of its increased activity during smolting (Hoar, 1988) and is therefore used as an indirect proxy of SW readiness of smolts in commercial salmon farming (Handeland & Stefansson, 2001). NKA is composed of two subunits α and β, where the α-subunit is the main catalytic unit and contains the binding sites for ATP, sodium and potassium. The β-subunit promotes folding and positioning of the protein into the basolateral plasma membrane. Salinity dependent isoforms of the α subunit are expressed in the gills of salmonoids (Nilsen et al., 2007; Tipsmark et al., 2011) and NKAα1a and NKAα1b are to major isoforms expressed in FW and SW (McCormick et al., 2009). There is differential expression of these isoforms during smolt development where NKA α1a is most abundant in FW while NKA α1b is most abundant in SW (Figure 2.1). Smolting is a pre-adaptive process, hence both isoforms will be present in gills during FW phase. However, as a part of the parr-smolt transformation the mRNA level of gill NKA α1a decreases and the NKA α1b increases in FW during smolting (Nilsen et al., 2007). This corresponds to an increase in salinity tolerance and is often used as an indicator of hypo-osmoregulatory capacity (McCormick, 2012; Nilsen et al., 2007).

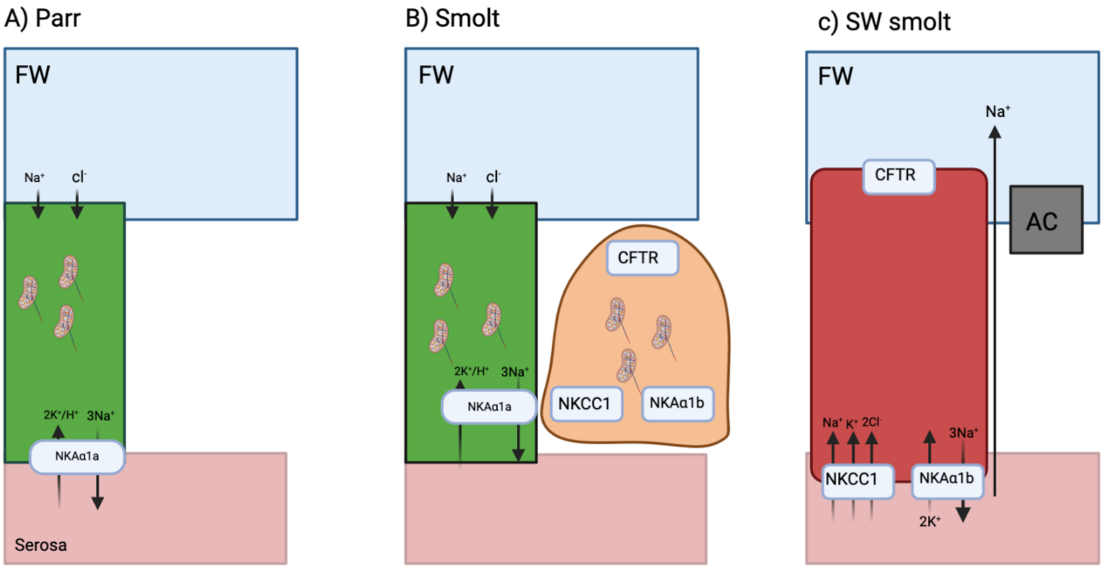

Figure 2.2 Model of the osmoregulatory components in the gills of freshwater and seawater teleost. The gill isoform of NKA changes during smoltification in FW, and after SW exposure. **A)** The expression of the NKAα1a isoform dominates in FW parr **B)** During smoltification the number of NKAα1b isoform increases as the fish is preparing for life in marine habitat. Other ion transporters (CFTR and NKCC1) are upregulated at this stage **C)** After SW exposure the NKAα1a disappears and NKAα1b are dominant. The expression of NKCC and CFTR continue to increase after SW exposure. The figure is made using BioRender and is inspired by a figure made by McCormick with colleagues (McCormick, S. et al., 2013).

NKA transport three sodium ions out of the cell for every two potassium ions it pumps into the cell, making the inside of the cell more negative and lower in sodium concentration. The extracellular concentration of sodium is higher than the intracellular concentration, while the intracellular concentration of potassium is higher than the extracellular concentration, this creates an ionic gradient across the cell membrane of the ioncytes. NKCC uses the ionic gradient to move chloride into the cell. The chloride then leaves the cell through an apical chloride channel CFTR, which is another marker related to smolt status due to its upregulation of mRNA during smolting (Kiilerich et al., 2007; McCormick, 2001; Singer et al., 2003).

2.2.2 Endocrine control

Increasing photoperiod is crucial for successful smolting, as it triggers the physiological changes necessary for adaptation to SW (Björnsson et al., 1989; McCormick et al., 1995; McCormick et al., 2007). Photoperiodic stimulation affects the neuroendocrine system in juvenile salmon, leading to changes in the expression of genes involved in osmoregulation, growth, and metabolism. Increasing photoperiod is the major factor stimulating hormones like cortisol and growth hormone (GH) (McCormick, 2001). During smolting, GH induces the liver to produce insulin-like growth factor (IGF)-1 which is involved in regulation of growth, metabolism, and osmoregulation (Björnsson, 1997; Björnsson et al., 2002). GH, IGF-1 and cortisol regulate the three major salt transporters, CFTR, NKCC and NKA during smolting (McCormick, 2001).

2.3 Smolt production in salmon farming

The salmon farming production cycle involves a 2-stage cycle which takes about 2-3 years to complete. The first stage is the FW production of smolts and the second is the grown-on stage following SW transfer of smolt. Prior to smolt production the process of fertilization and rearing is carried out (Mowi, 2022). In salmon farming, fertilized eggs are obtained from broodstock fish that have been selected for desirable traits like fast growth and disease resistance. After the eggs have been fertilized, the rearing process starts and eggs are placed into an incubator containing water where they stay until they hatch. After hatching, the production of smolt starts when fry is transferred to FW tanks on land where they stay for 8 to 16 months. During this time, the fry will develop into a parr which will develop into a smolt. Smolting usually takes up to 2 years in natural systems. In contrast, short duration in the production of large smolt is possible due to light and temperature manipulation in salmon farming.

There is no common method for smolt production in salmon aquaculture. Some producers simply grow smolt until they reach a very large size (>200g), others use feed with added salt to prime the fish for SW entry, and many producers use artificial light manipulation to stimulate and synchronize smolting. Such light manipulation imitates the natural conditions of smolting where salmon have evolved to smoltify following a period of exposure to winter photoperiods. To facilitate smolting of salmonoids, an artificial winter period is created by interrupting the constant light conditions in which they are typically reared after hatching, and exposing them to a daily light-dark cycle with limited light exposure of 12 hours or less per day, followed by a return to constant light in combination with increased temperature (Ytrestøyl et al., 2019). There are variations in the industry on the specifics of light manipulation employed in smolt production, where both the length of exposure to winter photoperiod and the number of hours light exposure per day used as winter period differs.

Numbers form the industry show that the quality of smolts produced are not optimal and it varies between facilities (Pino Martinez et al., 2023), where the quality is measured in terms of survival and growth. Producers claim that this issue can be explained by the difficulties in timing SW transfer in relation to the "smolt window", which is the time smolt has the best capacity to tolerate SW (Handeland et al., 1996). The use of intensive light regime can result in unsynchronized onset of smolting among farmed population (Pino Martinez et al., 2023; Stefansson et al., 2020). Therefor the determination of a SW ready smolt is key for the salmon farming industry. The industry defines a SW ready smolt by two gold standards. One method measures the gill NKA activity prior to SW transfer and/or changes in plasma chloride (Cl) levels after short term 24-h SW challenge (McCormick, S. et al., 2013). The second method use gene expression levels of the NKAα1a and NKA α1b isoforms to make a SW readiness score, because these isoforms are known to change in ratio during smolting (McCormick et al., 2009).

2.5 Machine learning

Machine learning (ML) is a subfield of Artificial Intelligence (AI) that involves using computer algorithms and statistical methods to identify patterns that can be used to make predictions from data (Mitchell & Mitchell, 1997). These algorithms learn from data and are able to gradually increase the accuracy of the model. A mathematical model is built on data called training data and can make predictions or decisions based on comparable data, without the need for explicit programming. ML algorithms are especially useful when predicting values/or classes containing complex patterns (interactions between features), and data from genomic research often possess these qualities, making them highly relevant to use when studying biological processes with genomics data (Greener et al., 2022). Usually, the aim of any ML model is prediction or interpretation (Libbrecht & Noble, 2015). There are two main categories of ML techniques, unsupervised learning, and supervised learning. Unsupervised learning finds patterns and make predictions using unlabeled data sets whereas supervised learning finds patters and make predictions using labeled datasets. In biology data is labeled using phenotypic information like disease status, weight or size. Supervised learning includes classification and regression, where a classification model predicts using classes and a regression model predicts using a continuous set of variables.

ML algorithms need to be trained before being able to make predictions and interpretations. This is usually done by splitting the data into a training set and a testing set, where the training set is the largest proportion of the of the original data (Greener et al., 2022). The training data is applied to the machine learning algorithm to train it, while the test set evaluates the performance of the model and make sure the model have predictive power on unseen data (Chicco, 2017). This process is called model training and the output of this process is a machine learning model. How well a machine learning model preforms and if its predictions can be trusted can be measured by model evaluation. This process uses different evaluation metrics to understand the machine learning models performance, including its strengths and weaknesses. In classification problems, the tools most frequently used include a confusion matrix (correlation between the predictions of a model and actual class labels), accuracy (measurement of how accurate the model is), precision (ratio between of true positive and total positive predictions), ROC (plot between the true positive rate and false positive rate) and AUC (area under the curve in ROC plot) (Jiang et al., 2020). In regression problems root mean squared error (RMSE) and the coefficient of determination (R^2) can be used as evaluation metrics (Sun et al., 2019). R^2 indicates whether the model is a good fit, whereas RMSE estimates how well the model was able to predict on test set outcomes and can be calculated with the following equation:

$$RMSE = \sqrt{}[\Sigma(P_i - O_i)^2/n] \qquad \text{Eq. (1)}$$

Cross validation is used to test the effectiveness of a machine learning method, and the technique can also be used as a resampling method, used to evaluate model performance (Liu, 2017). Cross-validation is a technique which splits the dataset into groups, where one group is kept aside as the test set, while training the model with the remaining groups. The process is repeated for each group held as the test set, which evaluates the performance, and the average evaluation scores are retained from the model. The skill of the model is summarized using the evaluation scores.

2.5.1 Random Forest

Random forest is a supervised ML method which can be used for both classification and regression problems. It is a commonly used algorithm introduced by Leo Breiman in 2001, where the algorithm combines the output of multiple randomized decisions trees to produce a single result (Breiman, 2001). The Random Forest algorithm is made up by an ensemble of decision trees where their predictions are combined to identify the most accurate result. Decision tree algorithms make decisions and predict values based on an if-else condition. Decision trees seek to find the best split to subset the data to create subsets that are as homogeneous as possible with respect to the target variable (Song & Lu, 2015). The Classification and Regression Tree (CART) algorithm is a common method used to build decision trees (Breiman, 2017). In CART, the algorithm iteratively selects a feature and a threshold to split the data into two groups based on the feature value. It then calculates an impurity score for each resulting subset, such as the Gini index for classification or the mean squared error for regression (Biau & Scornet, 2016). Decision trees can be prone to problems like overfitting and bias, therefore when multiple decision trees form an ensemble in the Random Forest algorithm, they are able to predict more accurate results (Qi, 2012). Bagging is a type of ensemble method where a random sample of data in a training set is selected with replacement (Breiman, 1996). Several samples of data are generated, and these models are then trained independently. The Random Forest algorithm uses both bagging and feature bagging to create an uncorrelated forest of decision trees, where feature bagging generates a random subset of features ensuring low correlation among decision trees (Biau & Scornet, 2016).

Two parameters, *M* and *mtry* are important for RF models and is usually tuned to make the model perform optimally. The *M* parameter describes the number of trees to grow and tuning this parameter may increase the computational burden, especially for big data sets containing hundreds and thousands of samples and variables (Schwarz et al., 2010). As *M* grows the variance of the forest decreases, thus more accurate predictions are most likely to be obtained by choosing a large number of trees to grow (Biau & Scornet, 2016). A trade-off between computational complexity and accuracy needs to be accomplished to achieve the best working model. The *mtry* parameter controls how much randomness is added to the decision tree process by controlling how many features are available to be considered for each new split

(Genuer et al., 2010). Tuning this parameter can have an impact on the model's performance. When building the Random Forest model, each tree in the forest only uses a subset of original data to train one tree. The data that remains are not used in the training process and can be used to measure the overall performance of the model, called the Out-of-bag (OOB) estimate (Biau & Scornet, 2016). The estimate is calculated by aggregating the predictions made by each tree on its corresponding OOB samples.

2.6 Aim of the thesis

This master project is a part of the ongoing Syncrosmolt project where main objective is to deliverer improved smolt production protocols, monitoring tools and enhanced broodstock to produce robust smolts with improved growth rates, survival, and welfare after SW transfer.

The aim of this thesis is twofold; first part is to use Random Forest to build predictive models of gill transcriptome data from fish given different light manipulation during smolt production to predict salmon smolt growth in SW. Different light treatment matter for SW performance, and there are large variations in developmental status among fish at the time of SW transfer (Strand et al., 2018; Ytrestøyl et al., 2019). Therefore, it is crucial to develop smolt markers able to predict long term SW performance. Second, use the prediction models to extract information about genes being involved in SW growth. This knowledge can be used to find new and better ways to assess SW readiness ensuring higher survival, growth and reduce welfare issues.

3 Methods

The data used in this master thesis were generated in the Synchrosmolt project funded by Fiskeri- og havbruksnæringens forskningsfinansiering (FHF). Sampling was carried out by researcher working on the project.

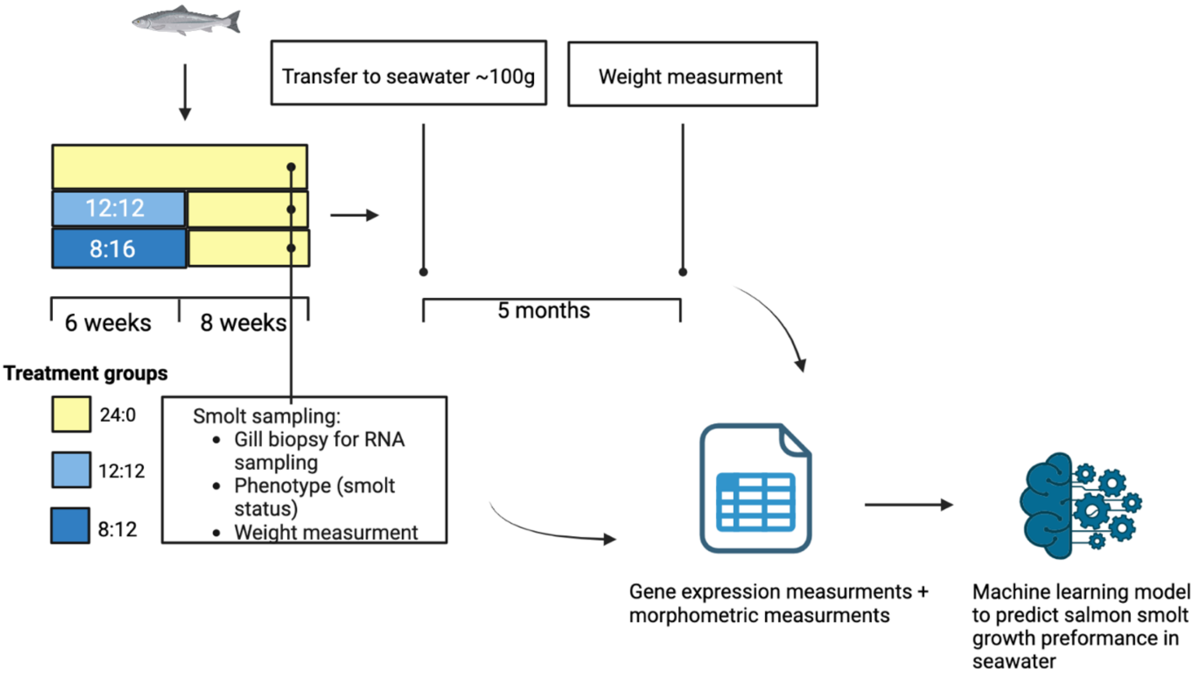
Transfer to seawater ~100g
Weight measurment
12:12
8:16
6 weeks
8 weeks
5 months
Treatment groups
24:0
12:12
8:12
Smolt sampling:
Gill biopsy for RNA sampling
Phenotype (smolt status)
Weight measurment
Gene expression measurments + morphometric measurments
Machine learning model to predict salmon smolt growth preformance in seawater

Figure 3.1 Flow of experimental set up and data analysis in the project. Production of smolt was carried out using three different photoperiods, continuous light (24:0), short-day (12:12), short-day (8:16). Prior to seawater transfer smolt sampling was carried out, here gill biopsies were taken, weight measurements and phenotyping of each salmon. After 5 months in seawater, sampling was again carried out and new weight measurement were taken. Gene expression measurements together with metadata containing the morphometric measurements was processed and applied to a machine learning algorithm to make a predictive model of salmon smolt growth performance in seawater.

3.1. Salmon smolt production

Salmon smolt was produced using a commonly used aquaculture smolt production protocol. In brief, fertilized eggs were placed on an incubator kept at 8°C, where they took 60 days to hatch. Following hatching 3000 fish was kept in one big tank on continuous light regime (24 hours light). When the weight of the salmon reached ~10g each salmon was pit-tagged and redistributed to three replicate tanks. Pit-tagging was performed on euthanized salmon that were given a small dose of benzocaine. Nano transponders were then surgically implanted into the left intraperitoneal cavity using specialized syringes. Pit-tagged salmon were further subjected into either short day photoperiod (8:16) short day photoperiod (12:12) or kept on continuous light regime (24:0), before they were transferred back on 24:0 for 8 weeks before SW transfer to initiate smolting (Figure 3.1).

3.2. Salmon smolt sampling

Prior to SW transfer salmon smolt sampling was carried out over a period of 6 days. The sampling was randomized and carried out tank by tank. Salmon were euthanized with a small dose of benzocaine before sampling. The weight and length of each salmon were measured and salmon smolt status was assessed visually and ranked from 1 to 3 by the following evaluation criteria: 1 (parr marks clearly visible), 2 (silvery appearance with faint parr marks), 3 (silver skin, no parr marks visible). All parameters were recorded electronically using Fishreader W (Trovan) and the software ZesusCapture. After 5 moths in SW, fish were taken up and morphometric measurements were taken.

3.3 Generation of smolt gill gene expression profiles

Small non-lethal gill biopsies were sampled prior to SW transfer from euthanized salmon and placed on dry ice before storing at -80°C. Pit-tag ID was recorded and linked to all gill biopsy samples. RNA isolation from gill-tissue was carried out at Qiagen (Germany) using RNeasy Fibrous Tissue Mini Kit, and RNA was sent to Novogene (Cambridge, UK) where RNA-seq libraries were prepared and RNA-sequencing was performed. RNA-seq libraries were 150 bp paired- end sequencing and sequenced to a depth of 10 million reads per sample.

3.4 Data analyses

Conversion of raw sequencing data to normalized gene expression quantification in the form of Transcripts Per Million (tpm) was done using the Salmon software (Patro et al., 2017). The Ssal.v3 genome assembly (GCA_905237065) with ENSEMBL gene annotation was used. All bioinformatics data handling prior to ML analyses were done by researchers in the Synchrosmolt project. All data analysis was performed in R (R Core Team, 2022) using the interface RStudio (Posit team, 2022). The *caret* R package (Kuhn, 2008) was used for ML and data visualization was performed using the R package ‘ggplot2’ (H. Wickham, 2016).

Prior to pre-processing, RNA-seq data was merged with metadata containing the morphometric measurements (June sampling and November sampling), smolt status (June sampling) and light treatment from each salmon-sample. Growth performance in SW, was calculated by subtracting the sampling weight in June from the sampling weight in November and samples that contained NA values were removed from the analysis.

3.4.1 Pre-processing and quality control of RNA-seq data

Initially, the data set was log transformed to reduce skewness and expression ratio of the NKAα1b and NKAα1a isoforms were calculated. By only keeping rows where the row means was larger than zero, non-expressed genes were filtered from data sets to lighten the computational load. Principal Component Analysis (PCA) was used to identify patterns and relationships in the data, and therefore also identify potential sample outliers. First, the function standardizes the values and make new principal components where the first principal component (PC1) corresponds to the direction with the maximum variation in the data set. The second principal component (PC2) corresponds to the direction with the second maximum amount of variation in the dataset. For visualization, samples were colored according to light treatment, and samples that deviated from the expected pattern in the dataset were identified as sample outliers and subsequently removed. Lastly, the data set was split into three subsets according to light treatment, 8:12, 12:12 and 24:0.

A statistical test for differences in weight (June sampling), growth and NKA ratio according to light treatment was preformed using a one-way Analysis of Variance (ANOVA) followed by a post-hoc Tukey HSD.

3.4.2 Feature selection

Prior to machine learning, feature selection was carried out to reduce the number of genes used as input in the machine learning algorithm. By reducing the number of genes, the machine learning model may be more accurate, and it may prevent the algorithm from crashing. Two different steps of feature selection (near zero variance and correlation) were applied to the data sets (8:16, 12:12 and 24:0). For near zero variance, variables with little variance were removed from the data set. The cutoff ratio of the most common value to the second most common value was set at 40 and the cutoff of the percentage of distinct values out of total samples was set at 20. In the second step a correlation matrix with the correlation coefficients for all genes were calculated, genes with a pair-wise correlation were removed using a cutoff set at 0.35. After feature selection was carried out a correlation matrix between all genes and growth performance in SW was made to assess the correlation between single genes and the response variable used in the machine learning algorithm.

3.4.3 Model training

A Random Forest algorithm was trained for each light treatment group separately (8:16, 12:12 and 24:0) to predict salmon smolt growth performance in SW. The Random Forest algorithm was trained using the complete data set to ensure the algorithm had enough training data to learn form. Model training, with k-fold cross validation as the resampling method was conducted on the data sets where the number of folds were set to 10. The method used different partitions of the data set to train and test the model on different iterations. The algorithm was initially trained by using the default parameters of the Random Forest algorithm, in carets train function, before parameter tuning was carried out using grid search for *mtry* and manual search for *ntree*. The metrices used to evaluate the performance of the model was root mean square error (RMSE) and the coefficient of determination (R^2), both based on cross validation. The performance metrices of the different models was then compared and a final predictive model was chosen for each light treatment group.

3.4.4 Variable importance

The variable importance for each model was calculated using the *VarImp()* function from the caret package (Kuhn, 2008). The function calculates the variable importance based on the mean decrease in accuracy (calculated across all trees) resulting from removing a particular feature from the Random Forest model. The mean decrease of accuracy is a measure used to assess the contribution of each feature to the accuracy of the models predictions. From the output of the *VarImp()* function the top 10 important features were chosen, and the geneIDs of the features were changed to their corresponding gene name in the ensemble annotation for easier interpretation. In the case where genes were annotated as novel, a Basic Local Alignment Search (BLAST) search were applied to identify their potential homologs and gain insight into their function (Altschul et al., 1990). The Blastp tool from NCBI with default settings were used to compare the translated protein sequence of the novel gene to a protein database. From the list with the closest matches to the query sequence, the most abundant description of gene function was chosen.

Finally, a functional enrichment analysis of the top 50 important features. These features were retrieved from the variable importance analysis was performed using the bioinformatic webtool g:profiler (Raudvere et al., 2019).

4 Results

4.1. Gene expression data quality and filtering

Mean normalized expression level (tpm) across all 1816 fish for each gene (Figure 4.1) ranged from 3.09 x 10^{-5} to 14.9 (median tpm = 1,02). A total of 1016 genes had a zero expression and were removed from further analyses.

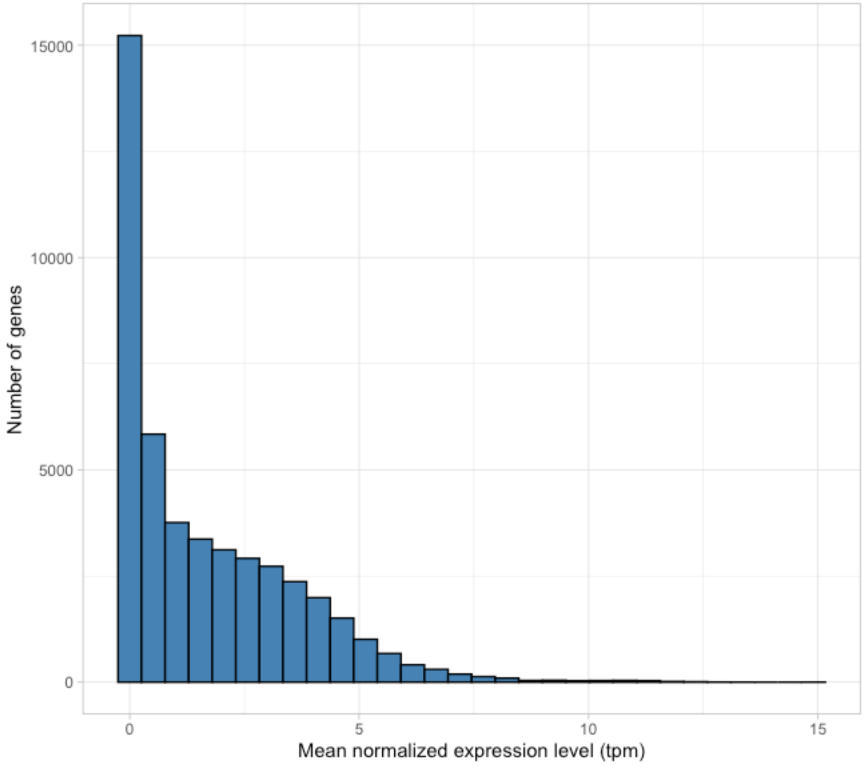

Figure 4.1 Histogram visualizing the mean distribution of gene expression in all samples. The mean gene expression of each gene is log transformed and calculated on the x-axis. The gene expression is measured in TPM. The y- axis shows the number of genes corresponding to the mean gene expression value. Most genes are lowly expressed.

To get a better overview of the data and identify potential sample outliers we performed PCA analyses on gene expression levels for each light treatment separately (Figure 4.2). PC1 and PC2 explained 18-15% and 5-9% of the variance across the three treatments respectively, and no extreme sample outliers were detected.

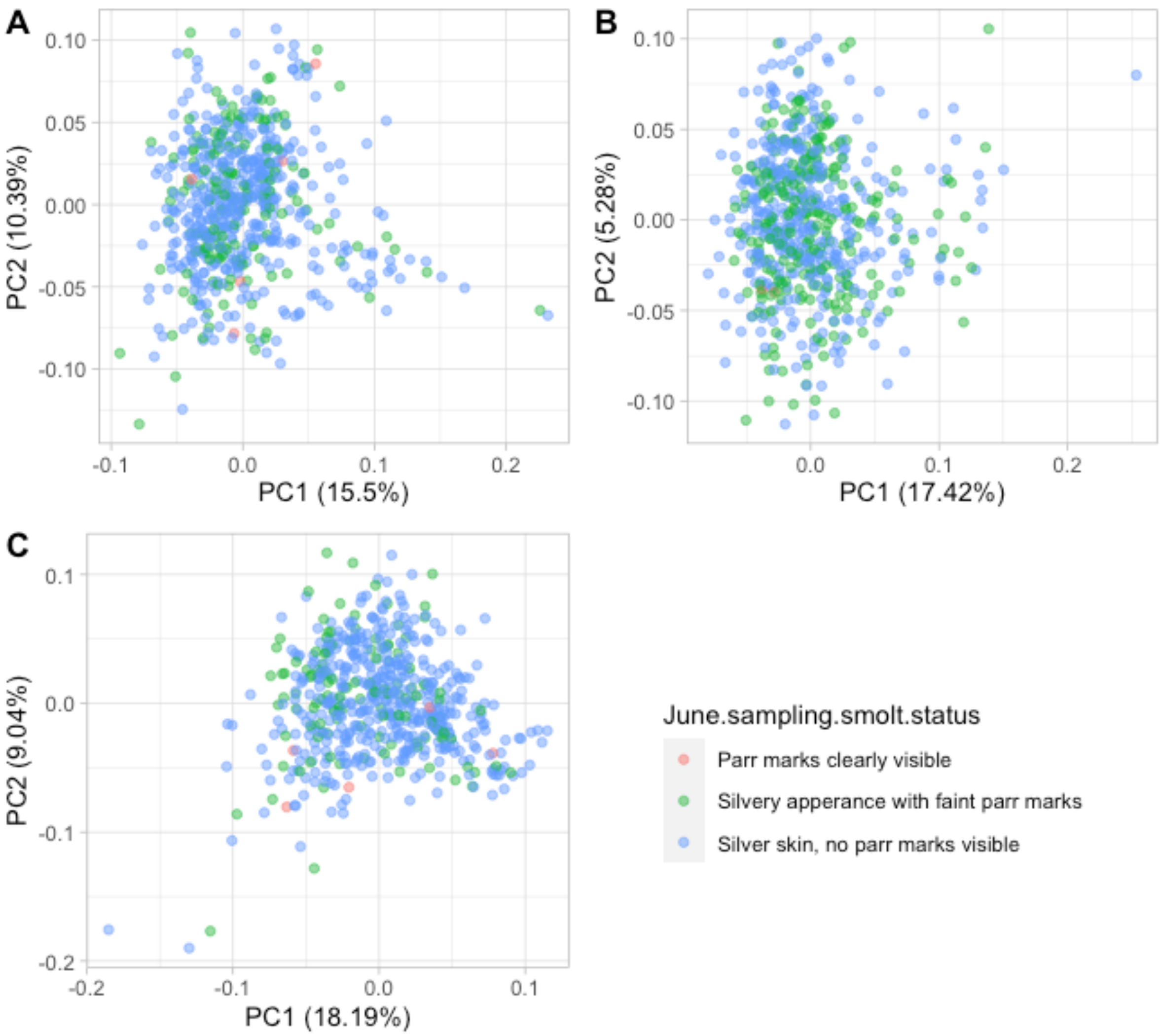

Figure 4.2 PCA plot of samples being exposed to three different light treatments. Samples were plotted along the PC1 and PC2 axis and colored according to June sampling smolt status. **A)** Gene expression levels for samples exposed to 8:16 photoperiod. PC1 explains 15.5% of the variation in the data set, while PC2 explains 10.39 % of the variation in the data set. **B)** Gene expression levels for samples exposed to 12:12 photoperiod, PC1 = 17.42% and PC2 0 5.28% **C)** Gene expression levels for samples exposed to 24:0 photoperiod, PC1 = 18.19% and PC2 = 9,04%.

4.2 Growth and NKA ratio expression in groups exposed to different light treatment

To determine whether photoperiodic history influenced growth in FW phase we compared the weight prior to SW transfer between groups exposed to different light treatment (Figure 4.3). Salmon exposed to 24:0 photoperiod showed significantly higher weight (Tukey-test, p-adjusted = 0.00) (median = 171g) compared to salmon exposed to winter photoperiod, 8:16 (median 153g) and 12:12 (median = 154g). The 24:0 group showed lower growth in SW compared to 8:16 and 12:12 group (Figure 4.3B). However, no significant difference was detected between the groups (ANOVA, p = 0.99).

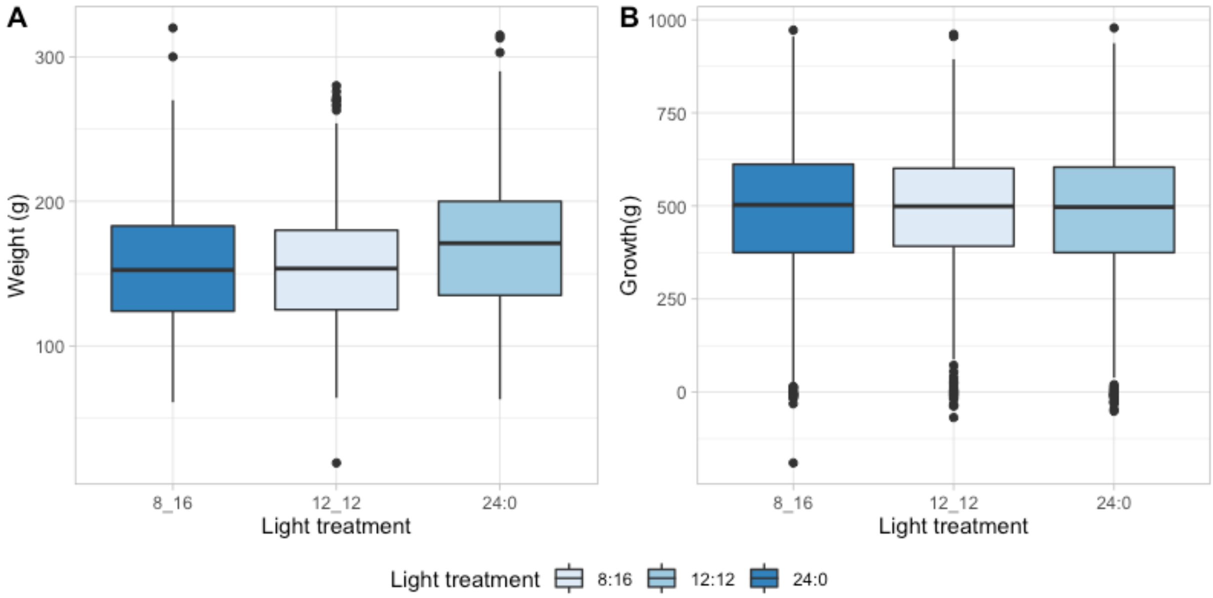

Figure 4.3. Boxplot visualizing the distribution of salmon smolt weight(g) measured prior to SW transfer and growth (g) in SW. **A)** The x-axis shows light treatment with the corresponding weight in on the y-axis. Salmon exposed to continuous light have a higher weight than salmon exposed to winter photoperiod. **B)** The x- axis show light treatment with the corresponding growth in SW on the y-axis. Growth in SW was the same for the three groups.

We were interested in how mortality in the SW phase varied according to light treatment. Fish exposed to 24:0 was the group with the highest mortality (total dead = 431). In 24:0 the mortality was highest compared to the 12:12 and 8:16 groups where the mortality was 394 and 398 respectively (Figure 4.4).

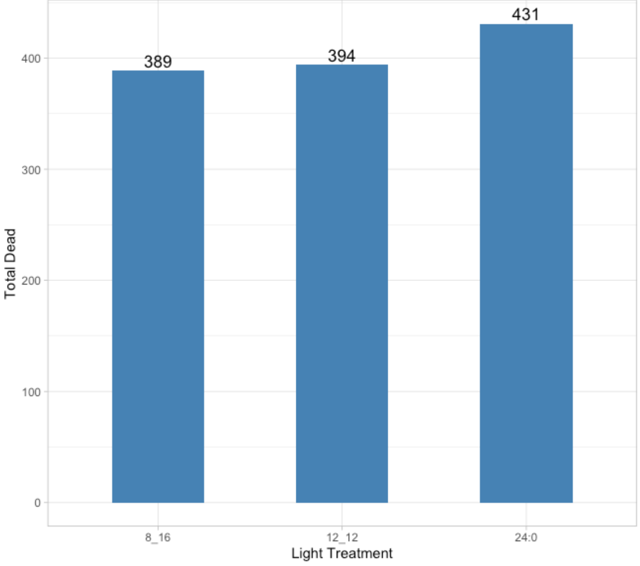

Figure 4.4 Mortality according to light treatment: The x-axis shows three different light treatments (8:16, 12:12 and 24:0) with the corresponding mortality of the group. The mortality was highest in the 24:0 group.

Since the ratio of NKAα1b to NKAα1a is a commonly used marker for smolt status in commercial smolt production, we wanted to test if light treatment were associated with different NKA-ratios (Figure 4.4). The median NKA-ratios were significantly higher in 8:16 and 12:12 (1.19, 1.18) compared to the 24:0 (0.88) (Tukey-test, p-adjusted = 0.00).

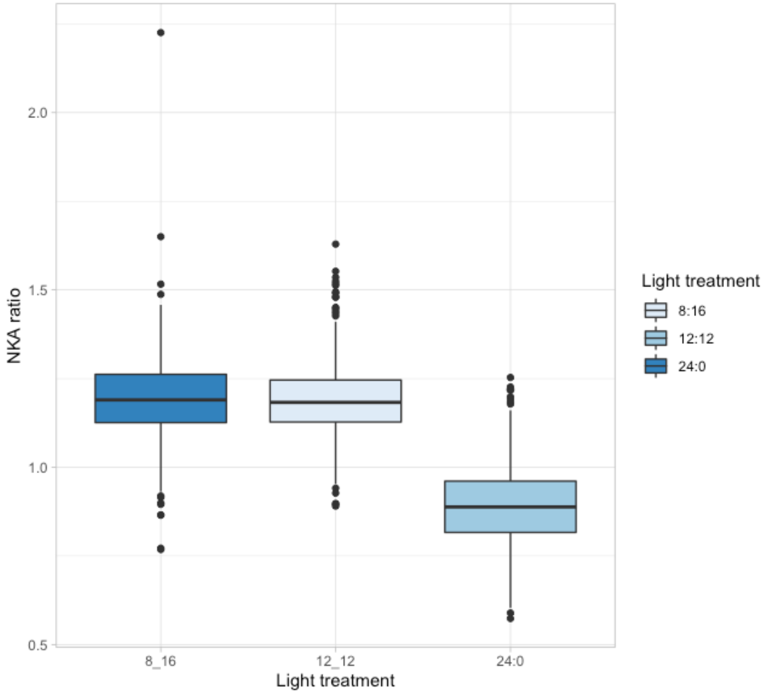

Figure 4.5 Boxplot visualizing expression for the NKA ratio for salmon exposed to different light regimes. The x-axis show which light treatment salmon is exposed to, 12:12, 8:16 and 24:0 while the y-axis show expression of NKA ratio. Salmon exposed to winter photoperiod expression ratio of the NKA is higher than for salmon exposed to continuous light.

4.3 Feature selection

To make the machine learning model more accurate and reduce computational costs, several steps of feature selection was performed on the data sets (8:16, 12, 12:12, 24:0). Genes with a near zero variance, meaning that they have few unique values and features with low variance were removed. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. In cases where two features have high correlation one of the features may be eliminated, the threshold was set at 0.35. The two steps of feature selection reduced the number of genes from 45.647 to 7825 for 8:12, to 8686 for 12.12 and to 7037 for 24:0.

4.4 Random Forest prediction models for growth in SW

A Random Forest algorithm was trained using three data sets 8:12, 12:12 and 24:0 to investigate if different photoperiod had effect on predicting growth in SW. Since growth is a continuous variable we chose a RF-algorithm with a regression model.

4.4.1 Salmonoids exposed to 8:16 photoperiod

The algorithm was trained using 610 samples and 7825 genes. First, a baseline model was trained using the default values of the rf algorithm, *ntree* = 500 and *mtry* = 70. This resulted in a model where the coefficient of determination (R^2) was 0.04. A low R^2 value reflects low explanatory/predictive power, and lead to inadequate predictions of the outcome variable, SW growth performance. The RMSE was calculated to be 185, compared to the mean growth in SW at 479h the models prediction error was estimated to be approximately 38% on average. To attempt to increase the model performance we tuned the mtry parameter, however this resulted in no changes of RMSE (=185) and R^2 (=0.04) (Figure 4.6). The best value for this parameter was *mtry*=2500 which was used for the analyses of the 8:12 photoperiod fish.

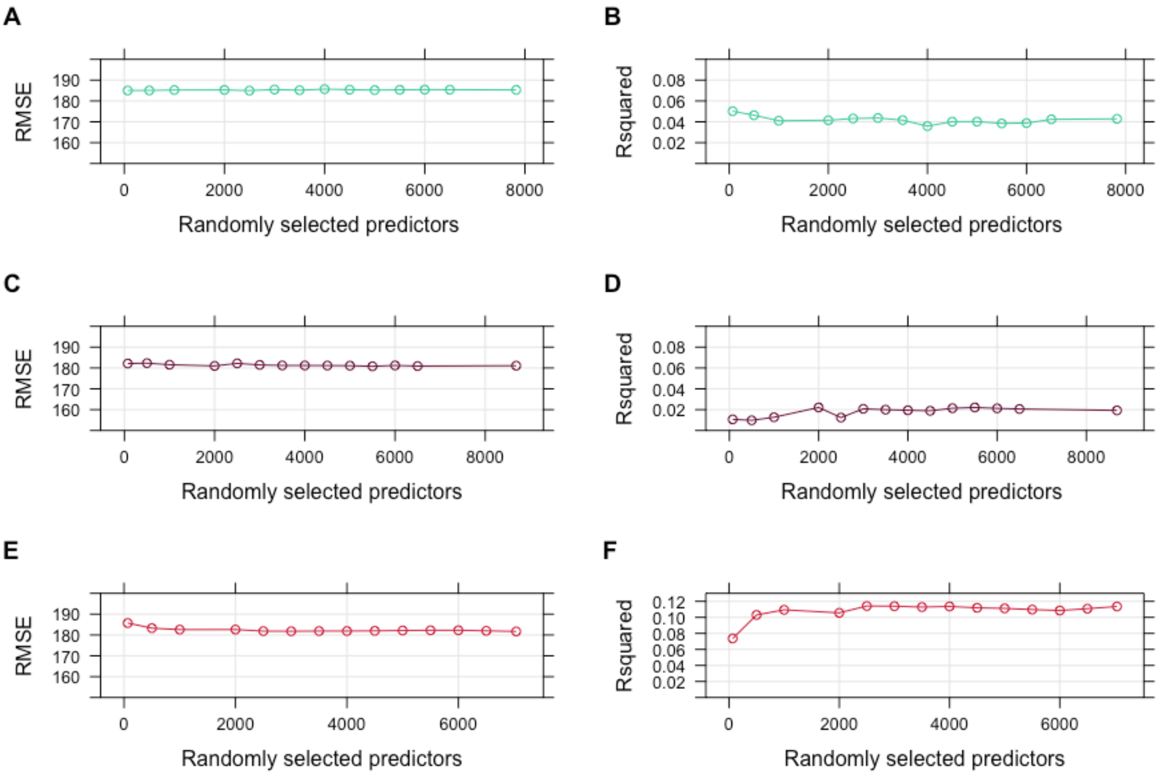

Figure 4.6 The best mtry numbers on models RMSE and R^2 produced by gridsearch. A-C) Number of selected predictors plotted on the x axis with the corresponding RMSE of the model. **A)** 8:16 **B)** 12:12 **C)** 24:0. RMSE ranges from 182-185 for the three groups. **D-F)** Number of selected predictors plotted on the x axis with corresponding R^2 of the model. **D)** 8:16 **E)** 12:12 **F)** 24:0. The 24:0 model has the highest R^2 (0.11) out of the tree models.

4.4.2 Salmonoids exposed to 12:12 photoperiod

The Random Forest algorithm was trained using 630 samples and 8686 features. Similar to the 8:16 group, a baseline model was trained with, ntree = 500 and mtry = 74. This resulted in a model where R^2 was 0.01 and RMSE was 183. Compared to the mean growth in SW at 481g the prediction error of the model was 38%. Attempts were made to increase the model performance by tuning the mtry parameter, however this resulted in very small changes in RMSE (=181) and R^2 (=0.02) (Figure 4.6). The best value for this parameter was *mtry* = 5500 which was used for the analyses of the 12:12 photoperiod fish.

4.4.3 Salmonids exposed to 24:0 photoperiod

The 24:0 data set contained 576 samples and 7037 features, which was used to train the algorithm. A baseline model was trained (RMSE = 185) and (R^2 = 0.07) before the mtry parameter was tuned in attempts to increase model performance. This resulted in small changes for RMSE (=182) (Figure 4.6) however the models R^2 increased (=0.11) (Figure 4.6). The best value for this parameter was *mtry*=7037 which was used for the analyses of the 24:0 photoperiod fish. The model could explain 11% of the variation in growth in SW and the RMSE of the model was 182, the prediction error of the model was 38%.

4.5 Variable importance

The variable importance analysis was used to quantify the contribution of each gene to the regression models, specifically to determine the genes that play a crucial role in the growth of salmon in SW. This approach allowed for identifying the genes with the highest impact on the model, providing insights into the key genetic factors that influence SW growth in salmon.

4.5.1 Salmonoids exposed to 8:16 photoperiod

From the variable importance analysis of the Random Forest model of salmon exposed to 8:16 photoperiod, the gene with the highest impact on the model was IRF1-2 (importance = 6.68) (Figure 4.6). The top second gene was annotated as uncaractherized gene (importance = 3.12), where Zink finger protein was the BLAST homolog (Accession number = AKP41000, e-value= 0.0, % identity = 78.90). The top third feature of the model was the plg gene (plasminogen) with a variable importance of 3.12. The known smolt gene-expression markers NKA and CFTR had a variable importance of -0.62 and 2.41 respectively.

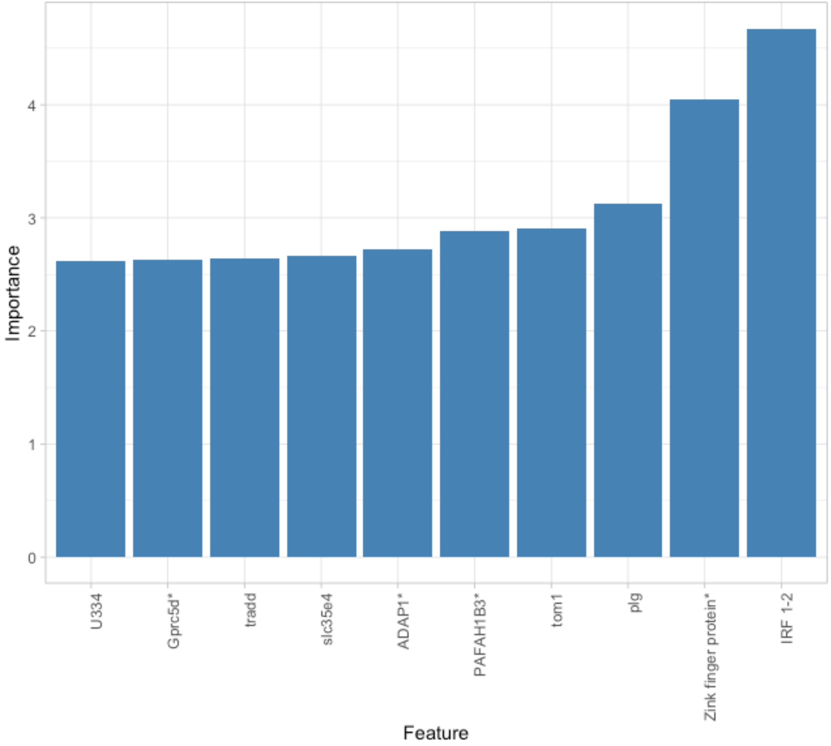

Figure 4.7 Variable importance plot for salmonoids exposed to 8:12 photoperiod. The 10 top genes contributing to the Random Forest model made from salmonoids exposed to short day photoperiod (8:16). The features are plotted on the x-axis with the corresponding importance on the y-axis. The gene which had the highest variable importance was a IRF 1-2 and its variable importance was 4.68. *Sequence homology BLAST

4.5.2 Salmonoids exposed to 12:12 photoperiod

From the variable importance analysis of the 12:12 Random Forest model, the gene with the highest impact on the model (importance = 7.89) was an uncharacterized gene, however its blast homolog was transposase (Accession number = ABV31710, e-value=1x10^{-74}, % identity = 78.06), and this gene had an importance of 7.89 (Figure 4.7). The top second gene of the model was also an uncharacterized gene (importance = 4.33), and its BLAST homolog was zink finger protein (Accession number = XP_045555642, e-value= 0, % identity = 99.11). In comparison, the known smolt-gene expression markers NKA and CFTR had very low variable importance of 0.13 and -0.33 respectively.

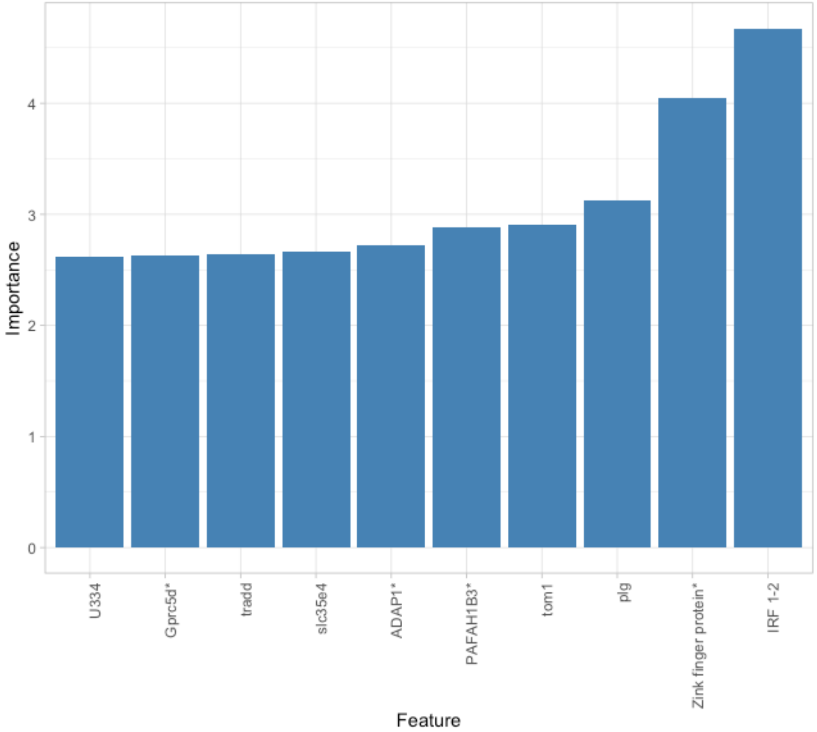

Figure 4.8 Variable importance plot for salmonoids exposed to 12:12 photoperiod. The top 10 genes contributing to the Random Forest model made from salmonoids exposed to 12:12 photoperiod. The genes are plotted on the x-axis with the corresponding importance on the y-axis. The gene which had the highest variable importance was an uncharacterized gene and its BLAST homolog was transposase, its variable importance was 4.92. *Sequence homology BLAST

4.5.3 Salmonids exposed to 24:0 photoperiod

From the variable importance analysis of the 24:0 Random Forest model, the known smolt-gene expression marker CFTR ranked top second with a variable importance of 10.5 (Figure 4.8), however NKA had a variable importance of -0.06. The BLAST homolog of the top third gene was the sodium/potassium-transporting ATPase subunit alpha-3 isoform (Accession number = XP_014055887, e-value= 0, % identity = 99.68), with a variable importance of 9.50.

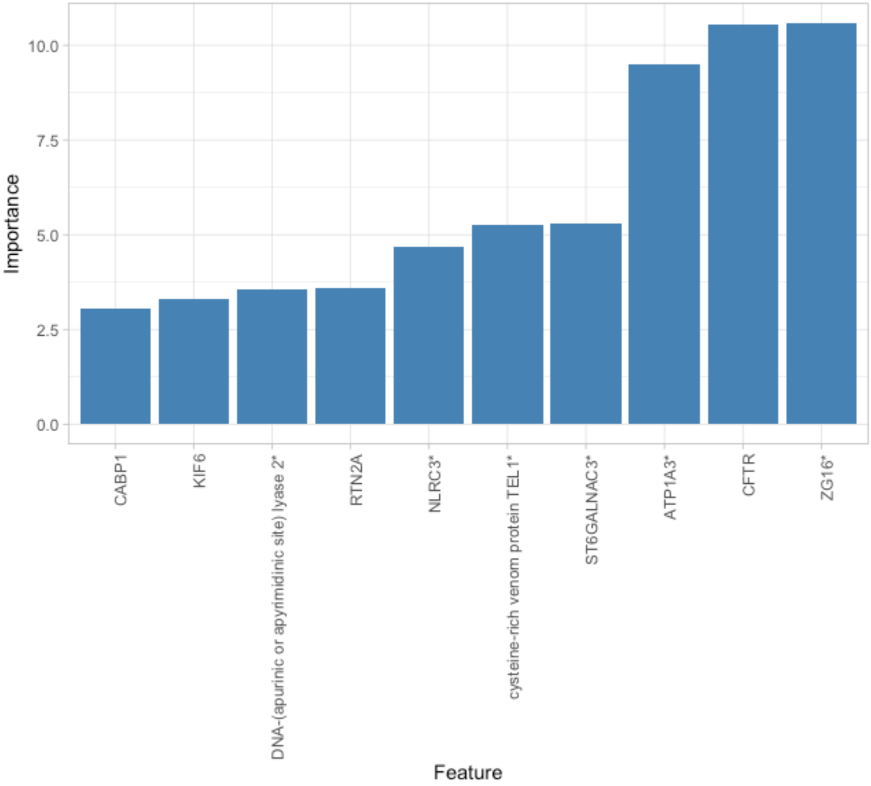

Figure 4.9 Variable importance plot for salmonoids exposed to 24:0 photoperiod. The top 10 genes contributing to the Random Forest model made from salmonoids exposed to 24:0 photoperiod. The features are plotted on the x-axis with the corresponding importance on the y-axis. CFTR was the gene with the second highest variable importance at 10.57. *Sequence homology BLAST

To test if there was any common pathways or biological functions associated with the top 50 ranked genes of each Random Forest model (8:16, 12:12, and 24:0) a functional enrichment analysis was performed in g:Profiler (Raudvere et al., 2019), however this gave no significant enrichments of gene ontologies or KEGG pathways.

4.6 Correlation between single genes and growth performance in seawater

Although the Random Forest models trained in this study performed poorly, we wanted to investigate how poor these Random Forest predictions were compared to using gene expression from single genes. This is relevant as the gene expression levels of single genes (or the ratio between two genes) are used in aquaculture production to assess smolt development status. We therefore computed the correlation between the expression levels of all genes individually and growth in SW and compared this with the correlation coefficients (sqrt(R^2)) from the Random Forest models~~.~~

The correlation coefficient of the 8:16 Random Forest model was 0.20, while the correlation coefficients of NKA and CFTR were 0.07 and 0.08, respectively (Figure 4.10). The gene with the strongest correlation to the SW growth was the JunD gene (JunD proto-oncogene, AP-1 transcription factor subunit) with a correlation coefficient of 0.14. (Figure 4.10). The correlation coefficient of the 12:12 Random Forest model was 0.14, while the correlation coefficients of the NKA and CFTR were 0.06 and 0.05, respectively (Figure 4.10). The 24:0 Random Forest model had a notably higher correlation coefficient of 0.33, while the correlation coefficients of CFTR and NKA were 0.27 and 0.22 respectively (Figure 3.10).

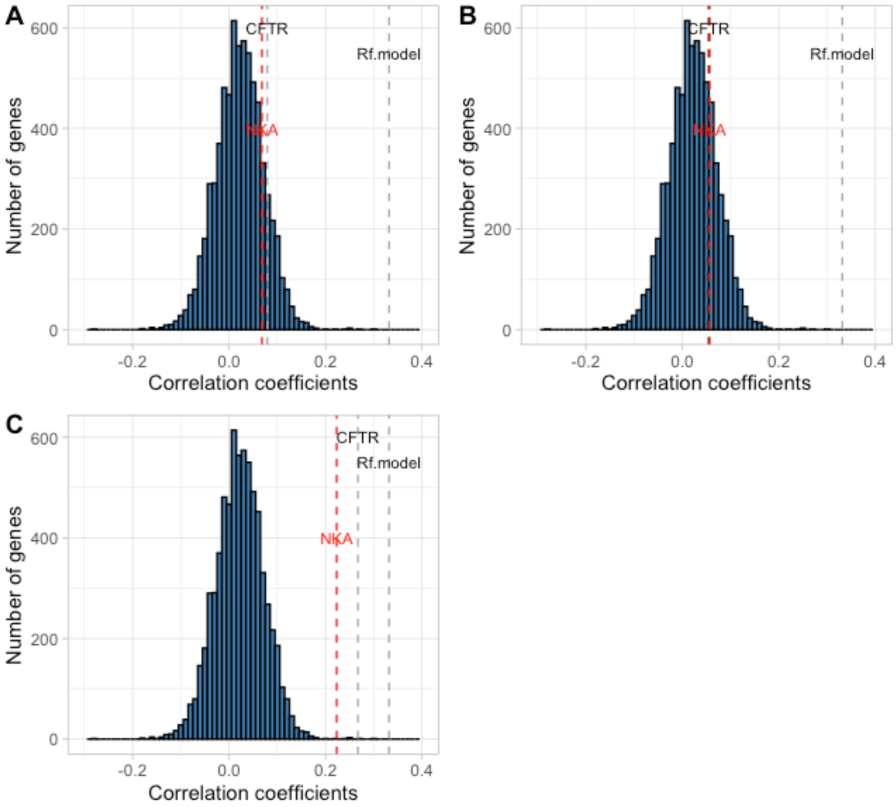

Figure 4.10 Distribution of genes according to correlation coefficient between genes and SW growth. The x- axis shows the correlation coefficient between the expression level of single genes growth and growth in SW. The y-axis shows the number of genes. CFTR and NKA are known gene-expression markers for SW readiness in salmon smolt and their correlation coefficient to SW growth is marked with a vertical line on the figure. Most genes show a low correlation coefficient meaning that they do not correlate strong with the growth variable. For all Random Forest models trained in this study **A)** 8:16, **B)** 12:12, and **C)** 24:0 the models correlation coefficient is higher than the correlation coefficient of single gene expressions.

5 Discussion

The two main objectives of this master thesis were to (i) test if we can use machine learning to predict smolt performance in SW based on FW gill transcriptomes from the same fish and (ii) identify the genes that will best predict SW smolt performance. In the following discussion, I will first focus on three aspects of the results, namely the poor performance of the machine learning predictions, the clear impact of photoperiodic history (i.e. light treatments) on the Random Forest models and the Random Forest method and its suitability for gene expression based trait predictions.

5.1 What can cause the poor prediction of growth in sea water?

Successful smolting requires the synchronization of developmental and physiological processes to produce a smolt phenotype ready for SW. The development of salinity tolerance is one important aspect of smolt development and is achieved by increasing the activity of several ion transporters in the gill, including the NKA protein (McCormick, 2012). Several isoforms of the *alpha (a)* subunit of this protein have been found expressed in salmonids (Richards et al., 2003). In gill tissue of Atlantic salmon, the expression of four distinct isoforms of the *nkaa* genes have been identified (*-1a, -1b, -1c, -3*) as well as one *nkab1* gene. The *nkaa1a* has been found highly expressed in FW whereas nkaa1b has been found highly expressed in SW (McCormick et al., 2009; Nilsen et al., 2007). Hence, gill transcriptome profiles have been used as an indirect proxy of SW readiness of smolts in both academic experiments (Kiilerich et al., 2007; McCormick, 2001; McGowan et al., 2021; Singer et al., 2003; Striberny et al., 2021) and commercial salmon farming (Handeland & Stefansson, 2001). However, in our study we find that Random Forest models have extremely poor predictive ability for SW growth, although superior to single genes involved in smolt gill development (Figure 4.10.). What could this mean?

One obvious interpretation is that poor model predictions reflect that gill transcriptome holds little information about longer term future growth. It is clear from numerous studies that gill genes involved in salinity regulation clearly distinguish small parr from larger fish ready to smolt (Nilsen et al., 2007; Pelis et al., 2001; Tipsmark et al., 2002). However, as suggested by Iversen with colleagues (Iversen et al., 2020), it is possible that most large smolts (>50-100g) have the capacity to suppress blood chloride levels following SW entry, despite variable levels of 'smolt gill' gene expression markers. In other words, it is perhaps not dysregulated gill function that results in suppressed growth in the sea for most fish, but rather other factors which does not reflect well in the gill transcriptome.

Another reason for poor model performance could be related to technical aspects. Here we choose a Random Forest algorithm which offers several advantages of little data preprocessing and have few parameters for the user to adjust. Random Forest is also powerful when number of observations (p) is much lower than features (n) as is the case in our study. In other studies of gene expression highly accurate rf models have been trained with n<100 with 10s of thousands of gene expression phenotypes (Chen & Ishwaran, 2012). The way of one-step-at-a-time node splitting enables the forest to impose regularization for effective analysis in cases where p >> n, also the grouping properties of trees enables RF to adeptly deal with correlation and interactions among variables (Ishwaran et al., 2010). Hence, we do not believe the choice of the Random Forest algorithm *per se* (compared to another more complicated model such as neural networks) has hampered our predictive ability. Another potential factor that could have impacted the Random Forest model is the way we quantify growth in SW. It is possible that by representing growth as a percentage of weight at smolt would capture some correlations to gill gene expression that we did not have in our dataset, however, due to time constraints we could not evaluate this in this thesis.

In this study the cross-validation method was used to measure model performance, which were evaluated based on the average performance across multiple iterations. We used the complete dataset for training because p >> n, to ensure the algorithm had enough training data to learn from. The tuning of model parameters (i.e *mtry*) was in our case done by grid search, this method could potentially lead to overfitting when tuning multiple times. However, by only tuning one parameter we do not think this have affected the performance our predictive model.

5.2 Photoperiodic history impacts the rf predictions

The growth performance of salmonoids is influenced by a number of abiotic and biotic factors with light and temperature being the primary environmental factors that regulate various physiological processes and therefore affect the life trajectories of this species (McCormick, S. D. et al., 2013). Previous studies have concluded that smolt development is highly impacted by photoperiodic history (Duncan & Bromage, 1998; Iversen et al., 2020; Saunders et al., 1985), including growth performance in SW (Striberny et al., 2021). It is perhaps not surprising that our results showed that the three Random Forest models from different light regimes had different predictive power and non-identical top lists of most influential features (Figures 4.7, 4.8, 4.9).

In this study we found that the Random Forest model for the 24:0 group were much better at predicting SW growth than any of the models from the fish exposed to short photoperiods (Figure 4.6). At first glance, this result is possibly a bit counterintuitive, as a large bulk of scientific work has demonstrated the importance of exposure to winter photoperiods in smolt production protocols (Björnsson et al., 1989; Björnsson et al., 2000; Handeland & Stefansson, 2001; McCormick et al., 1987; McCormick et al., 1995). Increased levels of GH are normally observed after transfer from short to long photoperiods, and salmon exposed to winter period can therefore develop a better scope for growth in SW (Björnsson et al., 1989; McCormick et al., 2007). However, when dissecting the data further, it is perhaps not as puzzling after all. If the 24:0 treatment was the least optimal in the context of gill physiology development and resulted in a less homogenous smolt population, then this could explain the better model performance in 24:0 fish. This interpretation is also indirectly supported by 24:0 fish also having the highest early SW mortality rate (Figure 4.4). These interpretations are consisted with other studies suggesting that extended use of continuous light regimes deprives the juvenile salmon of seasonal cures and thus critically interfere with the completion of parr-smolt-transformation. These negative effects include reduced hypo-osmoregulatory ability (McCormick et al., 1987), smolt-related endocrine signaling (Björnsson et al., 2000) and growth rate after transfer to SW (Striberny et al., 2021). Hence, we believe that poorer Random Forest model performance on fish exposed to winter photoperiod could be due to these fish having a more synchronized smolt gill development. This may indicate that building a predictive model based on gene expression from FW is not necessary reflecting the salmon status when it comes to survival, welfare and performance in SW.

The top 10 of associated genes ranked by variable importance were non identical between our three Random Forest models. For example, the gene expression of CFTR were only ranked top 10 important features in the 24:0 model (Figure 4.10), in comparison CFTR had a negative variable importance in the 12:12 model (Figure 4.9), and removing this feature would likely improve the model's performance (Cutler et al., 2012). The gene ranked top one differed for the three rf models, however for the genes included in the 8:16 and 12:12 model a sequence homology to Zink finger protein which was detected as the second top feature. Zink finger proteins represent the most abundant class of DNA binding proteins, often as transcription factors and they therefore play a significant role in gene regulation (Laity et al., 2001). NKA had a low variable importance in all three models < 0.13, and for the 8:16, and 24:0 the variable importance was negative. Our variable importance analysis has highlighted an important consideration when using gill transcriptome profiles as an indirect proxy for SW readiness of smolts in commercial salmon farming. Specifically, we found that the markers identified as important differed depending on the light regime used to produce smolts.

5.3 Immune system at smolt impact growth performance

Production of smolt in the 8:16 group resulted in a model where the IRF1-2 gene was the top feature associated with growth in SW. According to our results IRF1-2, was the strongest predictor of smolt growth in SW. In humans IRF1-2 is an activator of genes involved in the immune system. This gene activates transcription of genes involved in response to viruses and bacteria, as well as playing a role in the immune responses (Oshima et al., 2004; Su et al., 2007) The top third gene was the plg gene, encoding a protein circulating blood plasma which is converted to plasmin. Plasmin is a protease enzyme involved in the breakdown of blood clots, and it also has functions in the immune system, such as clearance of apocopic cells and regulation of inflammatory processes (Stelzer et al., 2016). The gill is the major mucosal immune barrier with lymphoid tissue, named gill associated lymphoid tissue (GIALT) (Koppang et al., 2015; Rességuier et al., 2020). This tissue is rich in T cells, natural killer cells and macrophages. The process of smolting has shown to suppress immune functions and a previous study on Atlantic salmon revealed a reduction in several types of immune-related cells (West et al., 2021). Our results suggest that the expression of specific genes associated with the immune system during smolting serves as a significant predictor of smolt growth in SW for samples exposed to winter photoperiod (8:16). These genes exhibit considerable predictive power, indicating their potential role in regulating immune responses are essential for growth in SW.

The results from the correlation analysis between gene expression and growth in SW showed that the JunD gene had the strongest correlation to growth for fish exposed to short photoperiod (8:16). This gene is important in cellular differentiation and proliferation (Hernandez et al., 2008), and could therefore potentially play a role in the remodeling of gill cell types during smolting. During smolting it is hypothesized that a reorganization of the gill immune system needs to coincide with the physiological changes happening, due to exposure to novel pathogens which they have previously not been exposed to (Johansson et al., 2016). The previous studies point towards an adaptive immunological reprogramming that helps to avoid immune shock when salmon transition between the distinctive pathogen complements of FW and SW habitats (Lee & Eom, 2016; Wang et al., 2012; West et al., 2021). Our findings suggest that the JunD gene could be involved in important cellular changes associated with growth in SW.

5.4 NKA subunits influence on growth

Results from the 24:0 Random Forest model showed ATP1A3 as the top third gene when predicting growth in SW, this gene encodes the α-subunit of NKA. This gene is a paralogue of the NKAα1a and NKA α1b, however it is not used as a smolt gene-expression marker. The α3 isoform was first found expressed in gills of rainbow trout (*Oncorhynchus mykiss*) along with three other NKA α-isoforms (α1a, α1b, α1c and α3) (Richards et al., 2003). The expression levels of NKA α1c- and α3-isofoms was found to be low in FW and their expression pattern did not change following transfer to SW. However, in a study performed by Nilsen and colleagues (Nilsen et al., 2007) they found that expression of NKA-α3 increased towards smolting. Our results showed this gene a strong predictor of growth in SW, which may be consisted with the previous findings where the α3 isoform may be important in the functional differences in NKA.

6 Conclusion

In this thesis we tested the use of a Random Forest model to predict salmon smolt growth in SW based on FW transcriptome from the same fish.

The results from the Random Forest model indicate low predictive power when using gill transcriptome to predict growth in SW. Furthermore, the top genes associated with growth in SW varied depending on the photoperiodic history of the sample. Therefore, the salmon farming industry should apply caution when relying on traditional smolt gene-expression markers to determine the optimal time for SW transfer, especially when using a winter photoperiod in the smolt production process.

7 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215 (3): 403-410. doi: https://doi.org/10.1016/S0022-2836(05)80360-2.

Biau, G. & Scornet, E. (2016). A random forest guided tour. *TEST*, 25 (2): 197-227. doi: 10.1007/s11749-016-0481-7.

Björnsson, B. T., Thorarensen, H., Hirano, T., Ogasawara, T. & Kristinsson, J. B. (1989). Photoperiod and temperature affect plasma growth hormone levels, growth, condition factor and hypoosmoregulatory ability of juvenile Atlantic salmon (Salmo salar) during parr-smolt transformation. *Aquaculture*, 82 (1): 77-91. doi: https://doi.org/10.1016/0044-8486(89)90397-9.

Björnsson, B. T. (1997). The biology of salmon growth hormone: from daylight to dominance. *Fish Physiology and Biochemistry*, 17 (1): 9-24. doi: 10.1023/A:1007712413908.

Björnsson, B. T., Hemre, G.-I., Bjørnevik, M. & Hansen, T. (2000). Photoperiod regulation of plasma growth hormone levels during induced smoltification of underyearling Atlantic salmon. *General and Comparative Endocrinology*, 119 (1): 17-25.

Björnsson, B. T., Johansson, V., Benedet, S., Einarsdottir, I. E., Hildahl, J., Agustsson, T. & Jönsson, E. (2002). Growth Hormone Endocrinology of Salmonids: Regulatory Mechanisms and Mode of Action. *Fish Physiology and Biochemistry*, 27 (3): 227-242. doi: 10.1023/B:FISH.0000032728.91152.10.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24: 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, 45 (1): 5-32.

Breiman, L. (2017). *Classification and regression trees*: Routledge.

Chen, X. & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99 (6): 323-9. doi: 10.1016/j.ygeno.2012.04.003.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10 (1): 35. doi: 10.1186/s13040-017-0155-3.

Council., N. S. (2022). *Nøkkeltall for norsk sjømateksport*. Available at: https://nokkeltall.seafood.no/ (accessed: 10.01.23).

Cutler, A., Cutler, D. R. & Stevens, J. R. (2012). Random Forests. In Zhang, C. & Ma, Y. (eds) *Ensemble Machine Learning: Methods and Applications*, pp. 157-175. New York, NY: Springer New York.

Duncan, N. J. & Bromage, N. (1998). The effect of different periods of constant short days on smoltification in juvenile Atlantic salmon (Salmo salar). *Aquaculture*, 168 (1-4): 369-386.

Ebbesson, L. O. E., Ekström, P., Ebbesson, S. O. E., Stefansson, S. O. & Holmqvist, B. (2003). Neural circuits and their structural and chemical reorganization in the light–brain–pituitary axis during parr–smolt transformation in salmon. *Aquaculture*, 222 (1): 59-70. doi: https://doi.org/10.1016/S0044-8486(03)00102-9.

Evans, D. H., Piermarini, P. M. & Choe, K. P. (2005). The multifunctional fish gill: dominant site of gas exchange, osmoregulation, acid-base regulation, and excretion of nitrogenous waste. *Physiological reviews*, 85 (1): 97-177.

fiskeridepartementet, N.-o. (2021). *Norsk havbruksnæring*. Available at: https://www.regjeringen.no/no/tema/mat-fiske-og-landbruk/fiskeri-oghavbruk/1/oppdrettslaksen/Norsk-havbruksnaring/id754210/ (accessed: 10.03.2023).

Fleming, I. A. (1996). Reproductive strategies of Atlantic salmon: ecology and evolution. *Reviews in Fish Biology and Fisheries*, 6 (4): 379-416. doi: 10.1007/BF00164323.

Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31 (14): 2225-2236. doi: https://doi.org/10.1016/j.patrec.2010.03.014.

Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23 (1): 40-55. doi: 10.1038/s41580-021-00407-0.

Handeland, S., Järvi, T., Fernö, A. & Stefansson, S. (1996). Osmotic stress, antipredatory behaviour, and mortality of Atlantic salmon (Salmo salar) smolts. *Canadian Journal of Fisheries and Aquatic Sciences*, 53 (12): 2673-2680.

Handeland, S. & Stefansson, S. (2001). Photoperiod control and influence of body size on off-season parr–smolt transformation and post-smolt growth. *Aquaculture*, 192 (2-4): 291-307.

Hasler, A. D., Scholz, A. T. & Horrall, R. M. (1978). Olfactory Imprinting and Homing in Salmon: Recent experiments in which salmon have been artificially imprinted to a synthetic chemical verify the olfactory hypothesis for salmon homing. *American Scientist*, 66 (3): 347-355.

Hernandez, J. M., Floyd, D. H., Weilbaecher, K. N., Green, P. L. & Boris-Lawrie, K. (2008). Multiple facets of junD gene expression are atypical among AP-1 family members. *Oncogene*, 27 (35): 4757-4767. doi: 10.1038/onc.2008.120.

Hiroi, J. & McCormick, S. D. (2012). New insights into gill ionocyte and ion transporter function in euryhaline and diadromous fish. *Respiratory Physiology & Neurobiology*, 184 (3): 257-268. doi: https://doi.org/10.1016/j.resp.2012.07.019.

Hjeltnes, B., Bang Jensen, B., Bornø, G., Haukaas, A. & Walde, C. S. (2018). *Fiskehelserapporten*.

Hoar, W. S. (1988). 4 The Physiology of Smolting Salmonids. *Fish Physiology*, 11: 275-343.

Hwang, P.-P. & Lee, T.-H. (2007). New insights into fish ion regulation and mitochondrion-rich cells. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 148 (3): 479-497. doi: https://doi.org/10.1016/j.cbpa.2007.06.416.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. & Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105 (489): 205-217.

Iversen, M., Mulugeta, T., Gellein Blikeng, B., West, A. C., Jørgensen, E. H., Rød Sandven, S. & Hazlerigg, D. (2020). RNA profiling identifies novel, photoperiod-history dependent markers associated with enhanced saltwater performance in juvenile Atlantic salmon. *PLOS ONE*, 15 (4): e0227496. doi: 10.1371/journal.pone.0227496.

Jiang, T., Gradus, J. L. & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, 51 (5): 675-687. doi: https://doi.org/10.1016/j.beth.2020.05.002.

Johansson, L.-H., Timmerhaus, G., Afanasyev, S., Jørgensen, S. M. & Krasnov, A. (2016). Smoltification and seawater transfer of Atlantic salmon (Salmo salar L.) is associated with systemic repression of the immune transcriptome. *Fish & shellfish immunology*, 58: 33-41.

Johnston, C. E. & Eales, J. G. (1967). Purines in the Integument of the Atlantic Salmon (Salmo salar) During Parr–Smolt Transformation. *Journal of the Fisheries Research Board of Canada*, 24 (5): 955-964. doi: 10.1139/f67-085.

Keefer, M. L. & Caudill, C. C. (2014). Homing and straying by anadromous salmonids: a review of mechanisms and rates. *Reviews in Fish Biology and Fisheries*, 24 (1): 333-368. doi: 10.1007/s11160-013-9334-6.

Kiilerich, P., Kristiansen, K. & Madsen, S. S. (2007). Cortisol regulation of ion transporter mRNA in Atlantic salmon gill and the effect of salinity on the signaling pathway. *Journal of endocrinology*, 194 (2): 417-428.

Koppang, E. O., Kvellestad, A. & Fischer, U. (2015). Fish mucosal immunity: gill. In *Mucosal health in aquaculture*, pp. 93-133: Elsevier.

Kristinsson, J. B., Saunders, R. L. & Wiggs, A. J. (1985). Growth dynamics during the development of bimodal length-frequency distribution in juvenile Atlantic salmon (Salmo salar L.). *Aquaculture*, 45 (1): 1-20. doi: https://doi.org/10.1016/0044-8486(85)90254-6.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28: 1-26.

Laity, J. H., Lee, B. M. & Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology*, 11 (1): 39-46. doi: https://doi.org/10.1016/S0959-440X(00)00167-6.

Lee, S.-Y. & Eom, Y.-B. (2016). Analysis of microbial composition associated with freshwater and seawater. *Biomedical Science Letters*, 22 (4): 150-159.

Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Briefings in Bioinformatics*, 20 (4): 1280-1294. doi: 10.1093/bib/bbx165.

McCormick, S., Regish, A., Christensen, A. & Björnsson, B. T. (2013). Differential regulation of sodium-potassium pump isoforms during smolt development and seawater exposure of Atlantic salmon. *The Journal of experimental biology*, 216: 1142-51. doi: 10.1242/jeb.080440.

McCormick, S. D., Saunders, R. L., Henderson, E. B. & Harmon, P. R. (1987). Photoperiod control of parr–smolt transformation in Atlantic salmon (Salmo salar): changes in salinity tolerance, gill Na+, K+-ATPase activity, and plasma thyroid hormones. *Canadian Journal of Fisheries and Aquatic Sciences*, 44 (8): 1462-1468.

McCormick, S. D., Björnsson, B. T., Sheridan, M., Eilerlson, C., Carey, J. B. & O'Dea, M. (1995). Increased daylength stimulates plasma growth hormone and gill Na+, K+-ATPase in Atlantic salmon (Salmo salar). *Journal of Comparative Physiology B*, 165 (4): 245-254. doi: 10.1007/BF00367308.

McCormick, S. D. (2001). Endocrine Control of Osmoregulation in Teleost Fish. *American Zoologist*, 41 (4): 781-794, 14.

McCormick, S. D., Shrimpton, J. M., Moriyama, S. & Björnsson, B. T. (2007). Differential hormonal responses of Atlantic salmon parr and smolt to increased daylength: A possible developmental basis for smolting. *Aquaculture*, 273 (2): 337-344. doi: https://doi.org/10.1016/j.aquaculture.2007.10.015.

McCormick, S. D., Regish, A. & Christensen, A. (2009). Distinct freshwater and seawater isoforms of Na+/K+-ATPase in gill chloride cells of Atlantic salmon. *Journal of Experimental Biology*, 212 (24): 3994-4001.

McCormick, S. D. (2012). 5 - Smolt Physiology and Endocrinology. In McCormick, S. D., Farrell, A. P. & Brauner, C. J. (eds) vol. 32 *Fish Physiology*, pp. 199-251: Academic Press.

McCormick, S. D., Regish, A. M., Christensen, A. K. & Björnsson, B. T. (2013). Differential regulation of sodium–potassium pump isoforms during smolt development and seawater exposure of Atlantic salmon. *Journal of Experimental Biology*, 216 (7): 1142-1151. doi: 10.1242/jeb.080440.

McGowan, M., MacKenzie, S., Steiropoulos, N. & Weidmann, M. (2021). Testing of NKA expression by mobile real time PCR is an efficient indicator of smoltification status of farmed Atlantic salmon. *Aquaculture*, 544: 737085. doi: https://doi.org/10.1016/j.aquaculture.2021.737085.

Mowi. (2022). *Salmon Farming Industry Handbook*. Available at: https://mowi.com/wp-content/uploads/2022/07/2022-Salmon-Industry-Handbook-1.pdf.

Nilsen, T. O., Ebbesson, L. O., Madsen, S. S., McCormick, S. D., Andersson, E., Björnsson, B. r. T., Prunet, P. & Stefansson, S. O. (2007). Differential expression of gill Na+, K+-ATPaseα-and β-subunits, Na+, K+, 2Cl-cotransporter and CFTR anion channel in juvenile anadromous and landlocked Atlantic salmon Salmo salar. *Journal of Experimental Biology*, 210 (16): 2885-2896.

Oshima, S., Nakamura, T., Namiki, S., Okada, E., Tsuchiya, K., Okamoto, R., Yamazaki, M., Yokota, T., Aida, M., Yamaguchi, Y., et al. (2004). Interferon regulatory factor 1 (IRF-1) and IRF-2 distinctively up-regulate gene expression and production of interleukin-7 in human intestinal epithelial cells. *Mol Cell Biol*, 24 (14): 6298-310. doi: 10.1128/mcb.24.14.6298-6310.2004.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14 (4): 417-419.

Pelis, R. M., Zydlewski, J. & McCormick, S. D. (2001). Gill Na(+)-K(+)-2Cl(-) cotransporter abundance and location in Atlantic salmon: effects of seawater and smolting. *Am J Physiol Regul Integr Comp Physiol*, 280 (6): R1844-52. doi: 10.1152/ajpregu.2001.280.6.R1844.

Perrott, M. N., Grierson, C. E., Hazon, N. & Balment, R. J. (1992). Drinking behaviour in sea water and fresh water teleosts, the role of the renin-angiotensin system. *Fish Physiology and Biochemistry*, 10 (2): 161-168. doi: 10.1007/BF00004527.

Pino Martinez, E., Imsland, A. K. D., Hosfeld, A.-C. D. & Handeland, S. O. (2023). Effect of Photoperiod and Transfer Time on Atlantic Salmon Smolt Quality and Growth in Freshwater and Seawater Aquaculture Systems. *Fishes*, 8 (4): 212.

Pisam, M., Prunet, P., Boeuf, G. & Jrambourg, A. (1988). Ultrastructural features of chloride cells in the gill epithelium of the atlantic salmon, Salmo salar, and their modifications during smoltification. *American Journal of Anatomy*, 183 (3): 235-244. doi: https://doi.org/10.1002/aja.1001830306.

Qi, Y. (2012). Random Forest for Bioinformatics. In Zhang, C. & Ma, Y. (eds) *Ensemble Machine Learning: Methods and Applications*, pp. 307-323. Boston, MA: Springer US.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47 (W1): W191-W198. doi: 10.1093/nar/gkz369.

Rességuier, J., Dalum, A., Pasquier, L., Zhang, Y., Koppang, E., Boudinot, P. & Wiegertjes, G. (2020). *Lymphoid tissue in teleost gills: variations on a theme. Biology (Basel) 9: 127*.

Richards, J. G., Semple, J. W., Bystriansky, J. S. & Schulte, P. M. (2003). Na+/K+-ATPase alpha-isoform switching in gills of rainbow trout (Oncorhynchus mykiss) during salinity transfer. *J Exp Biol*, 206 (Pt 24): 4475-86. doi: 10.1242/jeb.00701.

Rowe, D. K., Thorpe, J. E. & Shanks, A. M. (1991). Role of Fat Stores in the Maturation of Male Atlantic Salmon (Salmo salar) Parr. *Canadian Journal of Fisheries and Aquatic Sciences*, 48 (3): 405-413. doi: 10.1139/f91-052.

Saunders, R. L., Henderson, E. B. & Harmon, P. R. (1985). Effects of photoperiod on juvenile growth and smolting of Atlantic salmon and subsequent survival and growth in sea cages. *Aquaculture*, 45 (1): 55-66. doi: https://doi.org/10.1016/0044-8486(85)90257-1.

Schwarz, D. F., König, I. R. & Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, 26 (14): 1752-1758.

Singer, T. D., Finstad, B., McCormick, S. D., Wiseman, S. B., Schulte, P. M. & Scott McKinley, R. (2003). Interactive effects of cortisol treatment and ambient seawater challenge on gill Na+,K+-ATPase and CFTR expression in two strains of Atlantic salmon smolts. *Aquaculture*, 222 (1): 15-28. doi: https://doi.org/10.1016/S0044-8486(03)00099-1.

Song, Y. Y. & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*, 27 (2): 130-5. doi: 10.11919/j.issn.1002-0829.215044.

Staley, K. B. & Ewing, R. D. (1992). Purine levels in the skin of juvenile coho salmon (Oncorhynchus kisutch) during Parr-smolt transformation and adaptation to seawater. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 101 (3): 447-452. doi: https://doi.org/10.1016/0305-0491(92)90026-N.

Stefansson, S., Bth, B., Ebbesson, L. & McCormick, S. (2008). Smoltification. In, pp. 639-681.

Stefansson, S. O., Baeverfjord, J., Handeland, S. O., Hansen, T., Nygård, S., Rosseland, B. O. & et.al. (2005). *Fiskevelferdsmessigvurdering av produksjon av 0-års smolt*.

Stefansson, S. O., Björnsson, B. T., Ebbesson, L. O. & McCormick, S. D. (2020). Smoltification. In *Fish larval physiology*, pp. 639-681: CRC Press.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, 54 (1): 1.30.1-1.30.33. doi: https://doi.org/10.1002/cpbi.5.

Strand, J. E. T., Hazlerigg, D. & Jørgensen, E. H. (2018). Photoperiod revisited: is there a critical day length for triggering a complete parr–smolt transformation in Atlantic salmon Salmo salar? *Journal of Fish Biology*, 93 (3): 440-448. doi: https://doi.org/10.1111/jfb.13760.

Striberny, A., Lauritzen, D. E., Fuentes, J., Campinho, M. A., Gaetano, P., Duarte, V., Hazlerigg, D. G. & Jørgensen, E. H. (2021). More than one way to smoltify a salmon? Effects of dietary and light treatment on smolt development and seawater growth performance in Atlantic salmon. *Aquaculture*, 532: 736044. doi: https://doi.org/10.1016/j.aquaculture.2020.736044.

Su, Z. Z., Sarkar, D., Emdad, L., Barral, P. M. & Fisher, P. B. (2007). Central role of interferon regulatory factor-1 (IRF-1) in controlling retinoic acid inducible gene-I (RIG-I) expression. *J Cell Physiol*, 213 (2): 502-10. doi: 10.1002/jcp.21128.

Sun, S., Wang, C., Ding, H. & Zou, Q. (2019). Machine learning and its applications in plant molecular studies. *Briefings in Functional Genomics*, 19 (1): 40-48. doi: 10.1093/bfgp/elz036.

Thorpe, J. E., Talbot, C. & Villarreal, C. (1982). Bimodality of growth and smolting in Atlantic salmon, Salmo salar L. *Aquaculture*, 28 (1): 123-132. doi: https://doi.org/10.1016/0044-8486(82)90015-1.

Thorpe, J. E. (1994). Performance Thresholds and Life-History Flexibility in Salmonids. *Conservation Biology*, 8 (3): 877-879.

Thorpe, J. E., Mangel, M., Metcalfe, N. B. & Huntingford, F. A. (1998). Modelling the proximate basis of salmonid life-history variation, with application to Atlantic salmon, Salmo salar L. *Evolutionary Ecology*, 12 (5): 581-599. doi: 10.1023/A:1022351814644.

Tipsmark, C., Breves, J., Seale, A., Lerner, D., Hirano, T. & Grau, E. (2011). Switching of Na+, K+-ATPase isoforms by salinity and prolactin in the gill of a cichlid fish. *Journal of Endocrinology*, 209 (2): 237.

Tipsmark, C. K., Madsen, S. S., Seidelin, M., Christensen, A. S., Cutler, C. P. & Cramb, G. (2002). Dynamics of Na(+),K(+),2Cl(-) cotransporter and Na(+),K(+)-ATPase expression in the branchial epithelium of brown trout (Salmo trutta) and Atlantic salmon (Salmo salar). *J Exp Zool*, 293 (2): 106-18. doi: 10.1002/jez.10118.

Trovan. *FishReader*. Available at: https://www.trovan.com/en/aquaculture/products/FishReader-W-1
(accessed: 9.05.22).

Wang, Y., Sheng, H.-F., He, Y., Wu, J.-Y., Jiang, Y.-X., Tam, N. F.-Y. & Zhou, H.-W. (2012). Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. *Applied and environmental microbiology*, 78 (23): 8264-8271.

Wedemeyer, G., Saunders, R. L. & Clarke, W. C. (1980). Environmental factors affecting smoltification and early marine survival of anadromous salmonids. *Marine Fisheries Review*, 42 (6): 1-14.

West, A. C., Mizoro, Y., Wood, S. H., Ince, L. M., Iversen, M., Jørgensen, E. H., Nome, T., Sandve, S. R., Martin, S. A. M., Loudon, A. S. I., et al. (2021). Immunologic Profiling of the Atlantic Salmon Gill by Single Nuclei Transcriptomics. *Frontiers in Immunology*, 12. doi: 10.3389/fimmu.2021.669889.

Wilson, J. M. & Laurent, P. (2002). Fish gill morphology: inside out. *Journal of experimental Zoology*, 293 (3): 192-213.

Ytrestøyl, T., Baeverfjord, G., Kolarevic, J., Solheim, M., Hjelle, E., Mørkøre, T. & Brunsvik, P. (2019). *Hva betyr fremtidens produksjonsstrategier for ytelse, helse og velferd i sjøfasen (BENCHMARK) Faglig sluttrapport*.

Appendices

Code

R-scripts used for machine learning are available at:
https://github.com/sofierob/Master_thesis

Feature importance

Appendix 1: Feature importance of the 10 top features of the Random Forest model trained on samples exposed to 8:16 photoperiod. The gene product corresponding to the geneID from ensemble annotation. *BLAST homology

Gene ID	Product/Human readable gene name	Importance
ENSSSAG00000067957	IRF 1-2	4.675085
ENSSSAG00000096115	Zink finger protein*	4.044526
ENSSSAG00000048657	plg gene (plasminogen)*	3.123647
ENSSSAG00000091826	tom1 (target of myb1 membrane trafficking protein)	2.906224
ENSSSAG00000121032	PAFAH1B3*	2.880318
ENSSSAG00000104042	ADAP1*	2.719710
ENSSSAG00000098144	slc35e4 (solute carrier family 35 member E4)	2.663777
ENSSSAG00000081374	tradd (tnfrsf1a-associated via death domain)	2.646980
ENSSSAG00000059823	Gprc5d*	2.630080
ENSSSAG00000040556	U334 (nucleoside-triphosphatase, cancer-related)	2.613816

Appendix 2: Feature importance of the 10 top features of the Random Forest model trained on samples exposed 12:12 photoperiod. The gene product corresponding to the geneID from ensemble annotation. *BLAST homology

GeneID	Product/Human readable gene name	Importance
ENSSSAG00000108652	Transposase*	7.888520
ENSSSAG00000109564	Zink finger protein*	4.336683
ENSSSAG00000115630	Chordc1 undefined	4.307205
ENSSSAG00000106818	Zink finger (SCAN) protein	3.516217
ENSSSAG00000008608	GAS2L3*	3.341893
ENSSSAG00000067562	ZFAND4 (zinc finger AN1-type containing 4)	3.172849
ENSSSAG00000045361	nrxn3b (neurexin 3b)	3.011871
ENSSSAG00000112083	Eosinophil peroxidase-like*	2.929402
ENSSSAG00000042328	TLR undefined	2.832087
ENSSSAG00000039803	CFAP206 (cilia and flagella associated protein 206)	2.822662

Appendix 3: Feature importance of the 20 top features of the Random Forest model trained on samples exposed to 24:0 photoperiod. The gene product corresponding to the geneID from ensemble annotation. *BLAST homology

GeneID	Product/Human readable gene name	Importance
ENSSSAG00000089982	ZG16*	10.585890
ENSSSAG00000051689	CFTR (CF transmembrane conductance regulator)	10.536953
ENSSSAG00000093085	ATP1A3*	9.504433
ENSSSAG00000066983	ST6GALNAC3*	5.301288
ENSSSAG00000093757	TEL1 (Cystein-rich venom protein) *	5.246348
ENSSSAG00000108467	NLRC3*	4.668199
ENSSSAG00000050238	RTN2A (reticulon 2)	3.604681
ENSSSAG00000066030	DNA- (apurinic or apyrimidinic site) lysase 2*	3.559809
ENSSSAG00000028036	KIF6 (kinesin family member 6)	3.309607
ENSSSAG00000008547	CABP1 (calcium binding protein 1b)	3.049088

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway