CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

*MASTER'S THESIS*

# Analysis of gene expression patterns in wild eelpout from the Baltic Sea
## *Identification of expression patterns for differentially expressed genes connected to pollution*

EMMA WIJKMARK

*Department of Mathematical Sciences*
*Division of Mathematical Statistics*
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2014

Thesis for the Degree of Master of Science

# Analysis of gene expression patterns in wild eelpout from the Baltic Sea

## *Identification of expression patterns for differentially expressed genes connected to pollution*

Emma Wijkmark

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE – 412 96 Gothenburg, Sweden
Gothenburg, September 2014

**Abstract**

In a world where the human makes large toxic footprints the anthropogenic pollution of the aquatic environment is evident. It is therefore of great importance to understand what and how pollution adversely affect organisms in the environmental ecosystems. The aim of this study is to understand microarray gene expression patterns in the fish eelpout and how they change in connection to pollution. The analysed dataset is unique due to its extensive set-up with 158 wild fish from 16 sites located in four large geographical regions in the Baltic Sea.

The expression profiles were found to be highly varying. Clustering analysis showed that fish from the same region and site had a high tendency to group together. Comparison between fish from reference and polluted sites showed significant differences in gene expression but the effects were in general small. This is likely an indication that there are few shared effects between the polluted sites. Similarly, there were small effects on the gene expression between fish with a high and low reproduction success. Large and significant effects were however seen when comparing fish with low and high values of the known biomarker EROD. We also observed strong correlations between measured gene expression for the *CYP1A* gene and EROD. The strength of the correlation varied between regions and the highest correlation was found at the Swedish west coast. Finally, we assessed the independence between sampled fish and found that several genes had a high within-site correlation (median correlation over all genes were 0.18).

Our results suggest that there are small effects on gene expression connected to pollution and reproduction success. We could, however, identify large effects connected to region and site which may indicate that ecological factors and population parameters had a substantial impact on the observed gene expression profiles.

## Acknowledgements

First of all I want to express my gratitude to my supervisor Erik Kristiansson for his excellent guidance into the world of applied statistics and for his never-ending time for questions. Then I would like to thank my collaborators at the University of Gothenburg who not only made this thesis possible but also much more interesting by giving a biological view of the statistical results. It wouldn't have been as fun without you. Lastly, I want to turn my gratitude to everyone who not contributed directly to the thesis but indirectly by creating a warm atmosphere in which to spend the breaks.

*Emma Wijkmark, Gothenburg ,July 2014*

# Contents

# 1

# Introduction

In a world where the human makes large toxic footprints it is of great importance to understand the impact this has on the environment. The pollution is today widespread and many of the emissions are accumulated in the coastal environment. The anthropogenic pollution of the aquatic environment is evident and it is therefore a main issue to investigate what and how the pollution adversely affects organisms in the aquatic ecosystems. To do this, fishmonitoring is being developed as a tool for investigation of adverse effects caused by pollution and the traditional strategies in biomonitoring by using histopathology can today be complemented by microarrays. This makes it possible to investigate the relation between pollution and the aquatic organisms' gene expressions which as a next step might be used as early warning signs.

Within the project called Balcofish gene expression profiles of the fish eelpout (*Zoarces viviparus*) have been analysed using microarrays. The species eelpout has shown to overall be a valuable bio-indicator. One major advantage compared to other fish species is that it gives birth to live young, or rather the eggs develops into fry within the female. Thus it is possible to directly examine the reproduction for each female (Hedman et al., 2011). The Balcofish project is a large EU-BONUS financed project with the aim to investigate chemical pollution by developing fish monitoring and by providing solid information to the management of the Baltic Sea. To achieve this, fish have been sampled from 16 sites located in four large geographical regions in the Baltic Sea. The Balcofish project is a joint research project involving scientists from Denmark, Germany and Sweden. This thesis is performed in collaboration with the scientists at the University of Gothenburg, Department of Biological & Environmental Sciences and Department of Infectious Diseses, and is part of the Balcofish project.

The DNA microarrays, developed in the 1990s, revolutionised the field of gene expression analysis since it gave the possibility to investigate a very large number of genes simultaneously. The gene expression microarrays that have been used to retrieve the data for this study consist of a solid surface with microscopic DNA spots, each of them

containing many copies of a unique DNA sequence called probe. The probe can be a part of a known gene or an unidentified sequence. mRNA from each individual fish is isolated and reverse transcribed into cDNA (a more stable molecule) and fluorescently labeled. The cDNA from a sample is then hybridized to the microarray. cDNA that have not attached to any probe is washed away. The amount of cDNA sequences that attaches to each probe tells how much of that specific sequence that was found in the sample and by that how much this gene was expressed. This amount of attached sequences is usually examined by detection of fluorescent using a laser scanner (Alberts, 2002).

Since microarrays is used to investigate a large number of genes simultaneously the retrieved data will be high dimensional. The raw intensities are often noisy and this together with the high dimensionality urges the need for statistical tools. Common preprocessing of the intensities to aid the succeeding analysis is quantile normalization and log transformation. The log transformed gene expression data is often close to normal and therefore linear models can be used. The linear model gives a flexible frame work and there is a variety of statistical tests that can be used for inference. Among these we find simpler methods such as the classical t-test but alternatively more robust estimators can be applied for example the estimators found in the `limma`-package for the software `R` (Smyth, 2004).

Many of previous studies of gene expression differences in fish have investigated the short term impacts in controlled studies . The step from this into field implementation is not done without complications. Sellin Jeffries et al. (2012) looked at taking microarrays to the field by deploying fatheaded minnows (*Pimephales promelas*) to be caged for 7 days at four sites with different anthropogenic impact. It was shown that microarrays can be utilized to discriminate between sites with different contamination loads. The gene expression profiles were found to be site specific and that fish from low- and high-impact sites aggregated into distinct groups. In Williams et al. (2014) long term impact of contamination exposure was investigated when European flounder (*Platichthys flesus*), all sampled from the same site, were kept for 7 months in mesocosms with sediment from sites with different contamination loads. Small but statistically significant alterations in transcriptomic and metabolomic responses in liver tissue were detected between fish exposed to different sediments. Neither of these studies concerned fish from different populations. To be able to study gene expression for chronically exposed wild fish the fish need to be sampled at site and for large-scale studies the population effect cannot be avoided. For the final step in taking microarrays into field implementation the long term impact of exposure, ecological parameters and population effects will all be needed to be taken into account. In Falciani et al. (2008) European flounder sampled from six widely spread sites were compared and a wide range of genes seemed to be differently expressed between the sites. It was demonstrated that gene expression signatures in livers can predict the geographical sampling site but that the accuracy is limited to specific sites. The question is if discrimination between field sites with different levels of contaminations still is apparent when the study concerns wild sampled fish from different populations. And furthermore, if gene expression signatures can predict geographical sampling site and origin even when the number of sites is increased.

## 1.1 Aim

In this master's thesis project gene expression patterns in wild sampled fish will be investigated and questions concerning how to handle data from large-scale sampling will be adressed such as impact of site and region. The aim of this study is to analyse gene expression patterns in eelpout liver and find possible relation between gene expression differences and exposure of pollution as well as relation between gene expression differences and biomarkers.

The specific objectives are:

1. Identify tendencies of groupings of the fish. Investigate the general appearance of the gene expression profiles and discrimination between samples based on sampling site, region and exposure of pollution.

2. Find differentially expressed genes connected to pollution. Identify consistent differences between the gene expression profiles of the fish in relation to the contamination load at the sites.

3. Find differentially expressed genes connected to biomarkers such as reproduction success. This by defining a measure of reproduction success and then identify consistent differences in gene expression profiles between groupings of the fish based on reproduction success.

# 2

# Theory

The analysis of the gene expression data will rely on statistical methods and the following chapter will introduce the methods used for the data analysis. The structure of the linear model used for inference and the basics for handling the arising problem of high dimensionality will be explained. Moreover, an introduction to the unsupervised analysis methods hierarchical clustering and principal component analysis is given.

## 2.1 Linear models with empirical Bayes

### 2.1.1 Formulation of linear model using design and contrast matrix

Let $\mathbf{y}_g$ denote the log transformed gene expression for probe $g$ and denote the coefficient vector as $\alpha_g$ holding the coefficients representing the exploratory variables. These two vectors will be

$$\mathbf{y}_g = \begin{bmatrix} y_{g,1} \\ y_{g,2} \\ : \\ y_{g,n} \end{bmatrix}, \qquad \alpha_g = \begin{bmatrix} \alpha_{g,1} \\ \alpha_{g,2} \\ : \\ \alpha_{g,p} \end{bmatrix}$$

where $n$ is the number of samples, for this study $n = 158$, and $p$ is the number of covariates included in the model. The so-called design matrix, denoted as $\mathbf{X}$, is constructed as follows and has dimension $n \times p$,

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & . & x_{1,p} \\ x_{2,1} & x_{2,2} & . & x_{2,p} \\ : & : & . & : \\ x_{n,1} & x_{n,2} & . & x_{n,p} \end{bmatrix}.$$

The design matrix is such that each column holds data for one of the included covariates. Each row represents a sample since the $i$:th row is the covariate profile for the $i$:th sample and will not change with the probes. This profile holds information for this specific sample about the properties included in the model which can be both individual properties and properties of its environment. The linear model can now be formulated

$$\mathbf{Y}_g \sim \mathcal{N}_n(\mathbf{X}\alpha_g,\sigma_g^2\mathbf{I}_n)$$

where all samples are assumed to be independent.

For each covariate that is included in the design matrix we get an estimation of its coefficient and by that the possibility to remove its impact in the model or to examine its influence. This choice of which coefficients to examine can be made using a contrast matrix $\mathbf{C}$. It holds information about which coefficients that are to be compared and how these should be weighted. For the following setting each column should sum to zero. The number of rows in the contrast matrix is the same as the length of the coefficient vector $\alpha_g$. In the same way as for the design matrix, the contrast matrix will be the same for all probes. The contrast of interest is given by $\beta_g = \mathbf{C}^T\alpha_g$ where for this study the contrast matrix is always one dimensional

$$\mathbf{C} = \begin{bmatrix} c_1 \\ c_2 \\ : \\ c_p \end{bmatrix}.$$

Let the null hypothesis $H_0$ be no difference in gene expression and let the alternative hypothesis $H_A$ be that there is difference in gene expression according to

$$H_0 : \beta_g = 0$$
$$H_A : \beta_g \neq 0.$$

The covariates that are included in the design matrix but have entries zero in the contrast matrix are covariates which effects the model tries to compensate for. Covariates that have non-zero values in the contrast matrix are the ones which effects are being examined. The magnitude and sign of the elements in the contrast matrix tell how the features are weighted together. For example, if site effects are being examined the weights can be by the number of samples in the sites (Smyth, 2004).

### 2.1.2 Empirical Bayes and modified t-test

The variances for each probe are not assumed to be equal. But if estimating each probe's variance using a frequentist approach we will run into problems due to the high dimensionality. When having a very large number of tests, such as the microarrays will give, the risk is almost certain that some of the estimated variances will be too small due

to randomness. The result of this might be that even if the absolute difference in gene expression is too small to be of biological interest the result will still seem significant due to the by randomness grossly underestimated variance. Due to this the ordinary t-test might give a bad rating of the biological effects. To avoid this problem an empirical Bayes method can be used which borrows information across the probes to fit hyperparameters $\xi$ and $\eta$. Consider the following for testing the contrast $\beta_g$. The variance $\sigma^2$ is a random variable thought to be independent between the probes. $\mathbf{Y}_g$ is the log transformed gene expression for probe $g$.

$$\begin{aligned}\mathbf{Y}_g|\sigma_g^2 &\sim \mathcal{N}_n(\mathbf{X}\alpha_g,\sigma_g^2\mathbf{I}_n)\\ \sigma_g^2 &\sim \Gamma^{-1}(\xi,\eta)\end{aligned}$$

The inverse gamma distribution has not only been empirically shown to fit microarray data well but has good statistical properties since it is the conjugate prior to the normal distribution. The hyperparameters $\xi$ and $\eta$ are point estimated using all the given data across the probes. Since the number of probes is large the parameters are well estimated and considered as known. They decide how the variance behaves by setting the shape and rate of the inverse gamma distribution.

Let us consider the null hypothesis $H_0 : \beta_g = 0$ and the alternative hypothesis $H_A : \beta_g \neq 0$. Without further explanation and only giving an intuitive feeling for the result of empirical Bayes we can say that we have now moved from the ordinary t-test to the moderated t-test

$$T_g = \sqrt{\frac{n-p+2\xi}{\mathbf{C}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{C}}}\frac{\overline{\mathbf{Y}}_g}{\sqrt{S_g^2+2\eta}}.$$

where $\overline{\mathbf{Y}}_g$ is the projection of $\mathbf{Y}_g$ on the subspace spanned by $H_0$, $\overline{\mathbf{Y}}_g = \mathbf{C}(\mathbf{X}^T\mathbf{X})^-\mathbf{Y}_g$, and $S_g^2$ is the variance according to $S_g^2 = \mathbf{Y}_g^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X})\mathbf{Y}_g$. The distribution of $T_g$ follows a t-distribution with $n - p + 2\xi$ degrees of freedom (Smyth et al., 2003; Kristiansson et al., 2005). One way to interpret this is that a large variance will lead to a large $\eta$ and the variance $S_g^2$ is less trusted in the empirical Bayes setting. This approach makes the model more robust against underestimated variances.

### 2.1.3 Multiple testing correction

The above described model and computation of test-statistic will be done for each probe separately. The large number of probes tested in the microarray analysis gives a multiple testing situation where each comparison gives roughly 135 000 test-statistics. Therefore we expect that many test-statistics will have large magnitudes and thus small p-values by chance. This problem is referred to as the multiple testing problem. To correct for this the p-values can be adjusted by making them stricter, in other words less significant. There are different ways of adjusting the p-values which differ in what type of error that is being controlled and how strictly.

One simple and strict correction is the Bonferroni correction which controls the family-wise error rate (FWER). The FWER is the probability of at least one false rejection of the null hypothesis in favour of the alternative hypothesis for a chosen threshold $\alpha$ according to

$$\text{FWER} = P(V \geq 1) \leq \alpha$$

where $V$ is the number of falsely rejected null hypotheses. The Bonferroni corrected p-value is the non-adjusted p-value multiplied by the number of tests and this should be smaller than $\alpha$ for the change to be called significant. If the number of tests is large, as for microarray experiments, the Bonferroni correction along with other methods for FWER correction are too conservative which will result in too few genes called significantly changed.

Another method is Benjamini-Hochberg which instead controls the false discovery rate (FDR). The FDR is the number of tests incorrectly called significant according to

$$\text{FDR} = \mathbf{E}[V/R] \leq \alpha$$

where $V$ again is the number of falsely rejected null hypotheses in favour of the alternative hypotheses and $R$ is the total number of rejected null hypotheses (Hastie, 2009). This means that the Benjamini-Hochberg adjusted p-values are such that for a chosen threshold, $\alpha$, all probes with a FDR adjusted p-value lower than $\alpha$ are called differentially expressed. Among these the expected proportion of false discoveries, i.e. probes that are falsely called differentially expressed, are less than the threshold value $\alpha$ (Smyth, 2005).

## 2.2 Hierarchical clustering

Hierarchical clustering is used to visualise groupings in the data using a binary tree structure. Each level of the tree describes a hierarchy and represents a particular grouping of the observations into disjoint clusters. At the lowest level each cluster contains a single observation and at the highest level there is one cluster which contains all of the observations. The clustering is often visualised in a binary tree called dendrogram, see figure 2.1 for an example. A dendrogram is a tree where the nodes at the first level is at height zero and the consecutive levels of the dendrogram have branch lengths proportional to the dissimilarity between the joined clusters. Long branch lengths between clusters indicate dissimilar clusters while short branch lengths imply more similar clusters. The monotonicity property possessed by dendrograms means that dissimilarity between joined clusters is monotone increasing with the height at which the clusters are joined. By cutting the tree at different levels we get different groupings of the observations and if long branches are cut it is more likely that these groupings are representations of natural clusters.

A distance metric is defined for the pairwise distances between the observations which for example can be Euclidean distance or correlation distance. Left to determine is the meaning of two clusters being dissimilar and the answer to this question will change depending on the analysis. Three measures often used are single linkage, complete linkage

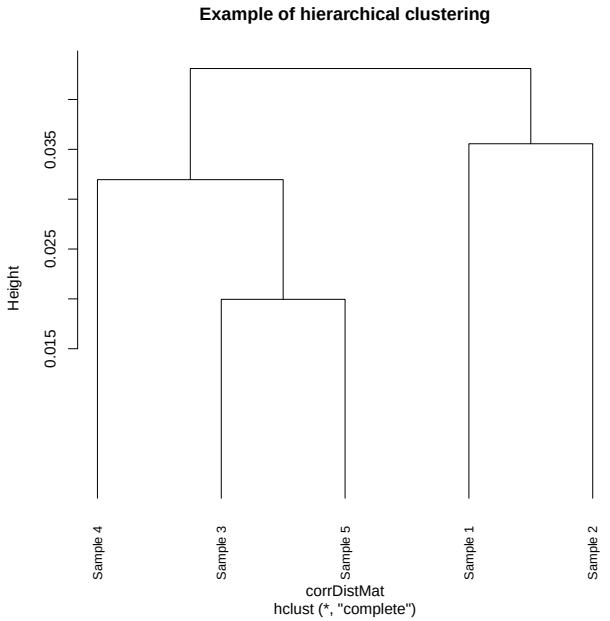


**Figure 2.1:** Example of a dendrogram from hierarchical clustering where 5 samples are being clustered based on correlation distance metric and using complete linkage. The clusters are recursively joined and the height at which clusters are joined describes the proportional dissimilarities between them.

and average linkage which all compare the pairwise dissimilarities between clusters and joins the two clusters with smallest between cluster dissimilarity. Single linkage defines the between cluster dissimilarity to be the smallest distance between two observation belonging to different clusters. In contrast, complete linkage defines the between cluster dissimilarity to be the distance of the most dissimilar pair of observations belonging to different clusters. Complete linkage will result in compact clusters since the algorithm favours clusters where all observations are as similar as possible. The trade-off for this is that some observations might be assigned to a cluster while being closer to some observations from another cluster. The average linkage can be seen as a compromise between the above described linkage choices since it defines the between cluster dissimilarity to be the average over all pairwise distances between the observations of different clusters. It creates clusters that are relatively compact where all observations are relatively close. The drawback is that the result is dependent on the numerical scale of the dissimilarity measure of the observations. Applying a monotone strictly increasing transformation to the dissimilarities can change the result while single and complete linkage only depend on the ordering of the dissimilarities. The average linkage has however statistical consistency property in contrast to single and complete linkage. For single and complete linkage it is unclear what aspects of the population distribution that are being estimated by the clustering algorithm.

There are two different clustering strategies. One of them is the agglomerative, a

bottom-up method which starts with each observation as a single cluster. Then the two least dissimilar clusters are recursively merged and the method ends with one cluster containing all observations. The other strategy, namely divisive method, is a top-down method which starts with all observations in one cluster and then recursively split the clusters until each observation is a cluster. Using divisive methods it is not guaranteed to get the monotonicity property required for the result to be shown as a dendrogram (Hastie, 2009).

## 2.3 Principal component analysis

Principal component analysis is a projection of the data onto a subspace spanned by uncorrelated variables. It is used for shrinking dimensions by an orthogonal transformation to components which sequentially captures the most variance. The principal components are in short the linear combinations of the original variables that give uncorrelated components with maximum variance among all sequential linear combinations of the original components.

Let $\mathbf{X}$ be an $p$-dimensional random vector holding the observed data having a positive semidefinite covariance matrix $\Sigma$. Let further, $\Gamma$ be an orthogonal matrix such that $\Sigma = \Gamma\Delta\Gamma^T$ where

$$\Delta = \begin{pmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_p \end{pmatrix}$$

and $\delta_i$ are the ordered roots of $|\Sigma - \delta\mathbf{I}| = 0$ i.e. $\delta_1 \geq \delta_2 \geq ..\delta_p \geq 0$. Define $\mathbf{U}$ to be $[U_1,...,U_p]^T$ and $\mathbf{U} = \mathbf{\Gamma}^T\mathbf{X}$. Then $\mathbf{U}$ is called a vector of principal components of $\mathbf{X}$ and $U_i$ is called the $i$th principal component of $\mathbf{X}$. The covariance matrix of $\mathbf{U}$ is $\Delta = \mathbf{\Gamma}^T\Sigma\mathbf{\Gamma}$ and therefore the components $U_i$ are uncorrelated with variance $\delta_i$ i.e. the variance of the $i$th component is equal to the $i$th eigenvalue of $\Sigma$. Since $\mathbf{\Gamma}$ is orthogonal, observing $\mathbf{U}$ is equivalent to observing $\mathbf{X}$.

The covariance matrix $\Sigma$ is rarely known and the principal components needs to be found by estimations (Arnold, 1981). Keeping the above in mind and omitting the details when having unknown $\Sigma$, let us consider $\mathbf{x}$ to be the vector of observed components. Denote the rotation matrix by $R$ which by having the principal components as columns and each variable as a row works as a matrix of basis vectors. By $\mathbf{Z} = \mathbf{x}R$ we get $\mathbf{Z}$ the matrix holding new coordinates for each sample in the subspace spanned by the principal components.

PCA is often used for variable reduction. By only using the first $k$ components of $\mathbf{Z}$ for further analysis of the data the number of variables are reduced to the $k$ variables that captures as much of the variance as possible. The choice of how many variables to use for analysis can be based on the amount of explained variance or by exterior restriction of the maximum number of dimensions. Interpretation of the principal components is not always obvious since they are linear combinations of the observed variables. It might be the case that the feature of interest is not captured by the first principal components

but by the $i$th and $j$th components instead. Therefore, principal components are often analysed by plotting the data using pairwise combinations of the principal components and the components giving rise to groupings in the data will be further investigated. The components of $\mathbf{x}$ are often standardized unless the numerical scale is on its own thought to be valuable information. The reason is that the same data but using different units will give completely different principal components (Johnson and Wichern, 1998).

# 3

# Materials / Method

## 3.1 Sampling set-up and data

### 3.1.1 Experimental design

The fish, later called samples, are of the species European eelpout (*Zoarces viviparus*) which has shown to overall be a valuable bio-indicator. Its stationary behaviour makes it possible to investigate long term ecological and anthropogenic impact in contrast to other fish species. One major advantage compared to other fish species is that the reproduction for each female can be directly examined since it gives birth to live young, or rather the eggs develops into fry within the female (Hedman et al., 2011).

The fish were sampled at 16 coastal sites in four regions. The sites are located such that four sites are on the Swedish west coast (Fjällbacka, Stenungsund, Göteborg, Vendelsö), four sites on Swedish east coast (Kvädöfjärden, Marsö, Gåsö, Slakmöre), five sites in Denmark (Agersø, Karrebæk Fjord, Isefjord, Frederiksværk, Roskilde Fjord) and three sites in Germany (Wismar, Eggers Wiek, Slazhaff), see table 3.1. The sites have been classified based on pollution level, some as polluted, some as medium polluted and some as reference sites (Albertsson et al., 2011). From each site, 8 to 11 female samples have been hybridized to microarrays. It is thought to be different populations of eelpout in the different regions but also that Denmark might be divided into two different populations. The samples taken from the same site are assumed to be from the same population even if there is a possibility that occasional exchanges of individuals between populations have occurred (Hedman et al., 2011).

Workshops have been held to get the sampling procedure as undiversified as possible but there are some aspects of the sampling procedure with large differences. One example is that the samples from German sites were moved by car before testing was performed. Considerable effort has been put into keeping the number of affecting factors at a minimum and as an example all microarray and enzyme analyses have been performed by the same laboratory.

**Table 3.1:** Sampling sites included in the study divided into four geographical regions and classified by the prior belief of their pollution levels.

| Region | Site | Classification |
|---|---|---|
| Sweden - West coast | Fjällbacka | Reference |
| | Stenungsund | Polluted |
| | Göteborg | Polluted |
| | Vendelsö | Reference |
| Sweden - East coast | Kvädöfjärden | Reference |
| | Marsö | Reference |
| | Gåsö | Polluted |
| | Slakmöre | Medium polluted |
| Denmark | Agersø | Reference |
| | Karrebæk Fjord | Polluted |
| | Isefjord | Medium Polluted |
| | Frederiksværk | Polluted |
| | Roskilde Fjord | Polluted |
| Germany | Wismar | Polluted |
| | Eggers Wiek | Medium polluted |
| | Slazhaff | Medium polluted |

### 3.1.2 Data included in the study

The gene expression data is one-channel microarray intensities and has been preprocessed by log transforming and normalizing each sample. Roughly 135 000 data points for each sample were analysed from the DNA microarrays as 2-4 probes had been designed from approximately 50 000 eelpout liver sequences (contigs).

Data for ethoxyresorufin-o-deethylase (EROD) activity and reproduction success for each fish is also included in the study. The reproduction data is in the form of number of dead and abnormal fry together with the total brood size.

## 3.2 Exploratory analysis

First unsupervised analysis was performed on all gene expression profiles using hierarchical clustering and principal component analysis. This was done to identify and visualise patterns in the general appearance.

### 3.2.1 Hierarchical clustering

Hierarchical clustering is an easy way to visualise groupings in data. An agglomerative (bottom-up) clustering procedure was used and implemented using `hclust` in `R`. The samples were clustered and each sample's gene expression profile was compared with the other samples' profiles. Since the absolute levels were not of interest but the patterns of the gene expressions, correlation distance was chosen instead of Euclidean distance. The distance metric used was $(1 - \mathbf{P})/2$ where $\mathbf{P}$ denotes the correlation matrix with $\mathbf{P}_{ij}$ defined as the Pearson correlation between the $i$th and $j$th samples' gene expression profiles.

When using average linkage a monotone strictly increasing transformation applied to the dissimilarities can change the result while single and complete linkage only depend on the ordering of the dissimilarities. The chaining effect that single linkage might have is not a desirable effect in this case. Instead it is more appealing to have compact clusters for the trade-off of risking that a few observations might be assigned to the wrong cluster. Hence complete linkage was chosen to overall have compactness such that all observations in a cluster were alike and still not having to depend on any transformation.

The samples were clustered using different gene expression profiles. These gene expression profiles were created by filtering based on how much each probe varied. For example, a high variance profile was created using the top 10% most varying probes and a low variance profile based on the 10% least varying probes. The clustering was performed using no information about site belonging or pollution level. This information was however used when visualising the results in dendrograms.

### 3.2.2 Principal component analysis

The scales of the variables in PCA have a huge impact on the result. If there is no information in scale or the variables have been measured using different unit systems the general recommendation is to scale the variables. In this case however, all the variables included were of the same type and by scaling we would risk getting an increase of noise by giving low variance probes to much impact. Therefore the variables were not scaled or centred since we rather risk losing some informative low expressed probes in favour for not bringing up to much noise.

First principal component analysis was performed by `prcomp` in `R` using all samples. Principal components were pairwise plotted against each other and the samples marked by region or pollution. This was then repeated for each region such that now only samples belonging to the same region were used in each of the regional principal component analyses. The samples were marked either by pollution level or site belonging.

## 3.3 Gene rank analysis

### 3.3.1 The linear model and empirical Bayes

The main outline for the following analysis was to fit a linear model to the log transformed expression data for each probe and then use empirical Bayes method to fit hyperparameters. Inference was done using moderated t-test and p-values were adjusted by FDR (Benjamini Hochberg) for cut-off 0.01.

The statistical model used in microarray analysis needs to be able to handle the hierarchical structure of the data and the inference should take the high dimensionality and the multiple testing situation into account. The linear model has shown to work well for microarray data and Sjögren et al. (2007) showed that it has a higher power to detect differentially expressed genes compared to standard methods. Log transformed gene expression data is often, as well as in this case, close to normal which makes the linear model favourable and we have robust estimators in the `limma`-package (Smyth, 2004). Based on which comparison that was made different aspects were taken into account and these features included in the design matrix. The samples were divided into different groups and compared using the contrast matrix. This allowance for general experiments to be analysed using the same framework by only changing the design and contrast matrix gives a very flexible and appealing model (Smyth, 2005).

The linear model was fitted using least squares approach and then empirical Bayes was used. Empirical Bayes can be seen as borrowing information across all probes and through this we did not have to assume equal variance across the probes. Instead the variances were thought to be from the same distribution with the fitted hyperparameters.

To correct for the multiple testing problem Benjamini Hochberg correction was used to adjust the p-values which controls the false discovery rate. The FDR adjusted p-values are interpreted such that for the chosen threshold 0.01, all probes with adjusted p-values lower than 0.01 are called significantly differentially expressed. Among these, the expected proportion of probes falsely called differentially expressed are less than the threshold value i.e. 1% (Smyth, 2005).

The above described steps were done in the software `R` using the `limma` package in Bioconductor which calculates the log fold change and adjusted p-value.

### 3.3.2 Differences in gene expression based on pollution classification

For comparing sites with different pollution classification the design matrix consisted of covariates indicating site, meaning that an overall site effect was estimated for each site and this was done for each probe separately. The interpretation of a site coefficient with large magnitude is that all together something in the environment or population of the fish from this site had a large impact on the expression of this particular gene in comparison to fish from other sites. Using the contrast matrix two groups were then formed. One consisted of all samples from reference sites and the other of samples from polluted sites. By comparing groups chosen to be reference and polluted we compared, for each probe, if there was any difference in effects between these groups. In other words,

we test how likely it is that the regulation of this gene was caused by the fish being in reference or polluted environment. This was done by having positive weights in the contrast for reference sites and negative weights for polluted sites. For this comparison the site coefficients were weighted by the number of samples from each site. The medium polluted sites had elements zero in the contrast matrix. The group sizes for this contrast were 51 samples in the reference group and 68 samples in the polluted group.

When looking at the regional comparison the sites could not be divided as easily into one reference group and one polluted group. The reason for this was the number of sites and the unbalanced partition of sites with different pollution classifications, see table 3.1. To overall get as good groups as possible and to stay consistent the group division for the regional comparisons were made such that medium polluted sites were combined with reference sites and compared against polluted sites. Notice that this means that for the Swedish west coast reference sites were compared against polluted sites while for Germany medium polluted sites were compared against the polluted site. The group sizes using above contrast were for: Swedish west coast 21-21, Swedish east coast 30-8, Denmark 20-29 and for Germany 19-10. Also for the regional comparisons all site coefficients were estimated but now only site effects for sites belonging to the same region were compared. This by setting all but one regions coefficients in the contrast matrix to zero and reweight.

### 3.3.3 Gene expression in relation to the biomarker EROD

In an aquatic system the type or mixture of contamination may be hard to know and a biomarker, indicator of some biological state or condition, can be of good use. Hepatic EROD induction in fish has shown to be useful in such situations and has successfully been used as a biomarker of general contamination (Whyte et al., 2000). EROD measures the enzyme activity of the *CYP1A* gene and its level of gene expression has been measured using four probes.

#### Correlation between EROD activity and expression of *CYP1A*

The within-site correlation between the log transformed EROD values and the log transformed measured expression of *CYP1A* was computed as Pearson correlation coefficient. This was done for each site separately and the regional estimations of the correlation were retrieved by averaging the within-site correlation for each region. The reason to first compute the within-site correlations is to get the correlation at individual level. Otherwise it might be the case that EROD level and expression of *CYP1A* might correlate at site level but that this correlation is not representative for the correlation at individual level.

#### Differences in gene expression for samples with high and low EROD activity

To compare the difference in gene expression in relation to levels of EROD two groups were formed. The high EROD group consisted of the 5% of the samples with highest

EROD values, values above 0.400, and the low EROD group were the 5% of the samples with lowest EROD values, values 0.037 and lower. This gave a group size of 8 in the high level group and 7 in the low level group. No other factors were taken into account for this comparison and only samples within these groups were used to fit the linear model.

This was repeated with high and low EROD individuals from the Swedish west coast. This time the limit for high EROD values were 0.35 and low EROD values were considered to be values 0.06 and under, giving group sizes of 3 and 4.

### 3.3.4 Difference in gene expression based on reproduction success

Reproduction success was measured using rate of total abnormal fry which was defined to be the number of dead and malformed fry divided by the total brood size. A low rate of abnormal fry means good reproduction success. In the same way as for the high and low EROD comparison the only features included in the comparison were indication of belonging to the high or low rate of abnormal fry group and only samples within these groups were included in the model. High rate of abnormal fry was chosen to be rates 0.05 or more while low rates were values 0.03 or lower. This to create two large groups of size 60 for the high rate group and 77 for the low rate group.

## 3.4 Within-site correlation

The common assumption of independence between samples is most likely not true since there are common factors affecting all fish from the same site such as ecological parameters and pollution. The assumption was tested by estimating $\rho$ describing the correlation between samples belonging to the same site and $\sigma^2$ describing the variance. This was done for each probe separately and the subscript indicating probe has therefore been omitted. A simplifying assumption made was that these parameters were the same for all sites and that there were no correlations between samples belonging to different sites. The covariance matrix was as follows

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_M \end{bmatrix}$$

where each $\Sigma_i$, for sites 1,...,$M$, is the submatrix describing the covariance for site $i$ and has dimension $n_i \times n_i$ where $n_i$ is the number of samples at site $i$. The submatrix for each site of the covariance matrix $\Sigma$ was

$$\Sigma_i = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

We hence assumed that there existed within-site correlation for the samples but no between-site correlation.

The estimation was done by maximizing the multivariate normal log likelihood

$$l(\sigma^2,\rho) = -\frac{n}{2}ln(2\pi) - \frac{1}{2}ln(det(\Sigma)) - \frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu).$$

The average of the log transformed gene expressions was subtracted for each sample by replacing $\mu$ with $\bar{y}$ where $\bar{y} = \sum_{i=1}^{N} y_i$ and $N$ is the total number of samples. To ease the optimization this was rewritten by taking advantage of the structure of the covariance matrix $\Sigma$, using the matrix determinant lemma, Sherman-Morrison formula and the multinomial theorem (Bernstein (2009) Fact 2.16.1 concerning determinant and invers and Fact 1.17.1), yielding

$$\begin{aligned} l(\sigma^2,\rho) = &-\frac{1}{2}\left(Nlog(2\pi\sigma^2) + (N-M)log(1-\rho) + \sum_{i=1}^{M} log(1-\rho+\rho\cdot n_i)\right) \\ &-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{M}(a_iA_i + b_iB_i)\right) \end{aligned}$$

where

$$a_i = \frac{-(n_i-2)\rho - 1}{((n_i-1)\rho+1)(\rho-1)}, \quad b_i = \frac{\rho}{((n_i-1)\rho+1)(\rho-1)}$$

and

$$A_i = \sum_{k=1}^{n_i}(y_{ik}-\mu_i)^2, \quad B_i = \left(\sum_{k=1}^{n_i} y_{ik} - \mu_i\right)^2 - A_i$$

where $y_{ik}$ is the $k$th sample in the $i$th site. For maximization of the likelihood numerical optimization in `R` using `optim` was performed.

# 4

# Results

The result section is divided into three parts. First is the exploratory analysis which investigates the general appearance and similarities between samples. In the next part changes in the gene expression patterns are analysed. The final part investigates independence between the samples.

## 4.1 Exploratory analysis

To get an understanding of the data the unsupervised analysis methods hierarchical clustering and principal component analysis were performed on the samples' gene expression profiles.

### 4.1.1 Hierarchical clustering

The dendrogram in figure 4.1 shows a distinct clustering by region. A German cluster containing two thirds of the German samples and only one non-German sample was found. The Swedish east coast samples were also well clustered except the samples from the site Marsö. These were clustered into a separate cluster and then joined with clusters for other regions while the large Swedish east coast cluster was late joined with other clusters. Two clear Swedish west coast clusters were formed. However, the cluster consisting of samples from Fjällbacka and Göteborg in the lower part of figure 4.1 is likely due to technical variation since all these were measured on the same physical chip. The Danish samples were more scattered than the samples from the other regions and no complete Danish clusters are seen even though some Danish samples were clustered together.

A clear grouping was seen for the clustering based on high variance probes, figures 4.1 and 4.2, in comparison to clustering using low variance probes, figures A.1 and A.2 in appendix.

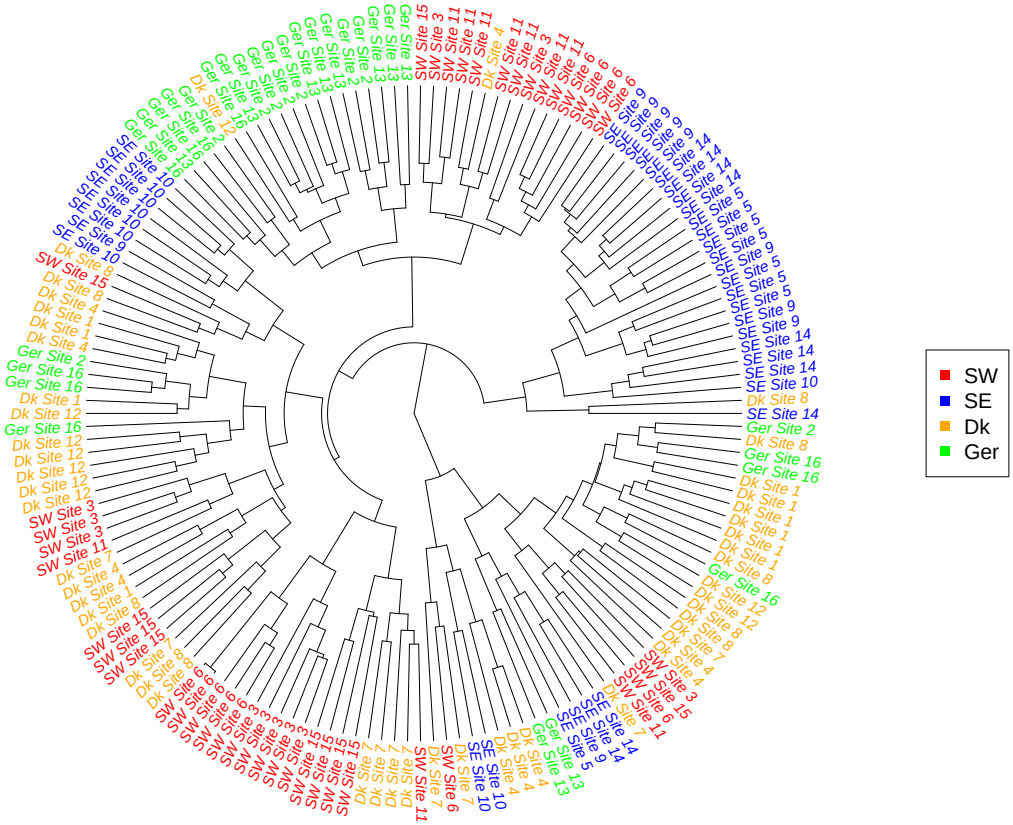


**Figure 4.1:** Hierarchical clustering using correlation distance metric and complete linkage based on the 10% most varying probes. Dendrogram coloured by region and sites numbered in alphabetic order which shows a clear regional clustering of the samples.

A large part of the regional clustering came from clustering by site which was found at a lower hierarchical level than the regional clustering and is seen in figure 4.2 where the dendrogram is coloured by site. Even if not all samples from the same site were clustered together small clusters tended to be formed by samples from the same site.

We did not identify any clustering based on levels of pollution, figure 4.3. The small tendencies seen seem more to be site or region effects that make some pattern appear.

### 4.1.2 Principal component analysis

The principal component analysis gave no distinct division of the samples into regions or pollution classification when looking at two dimensional principal component plots. The first 10 principal components captured 57 % of the variance and how the variance captured by the principal components declines is shown in figure B.1 in appendix. As in line with the hierarchical clustering there seemed to be more grouping by region than

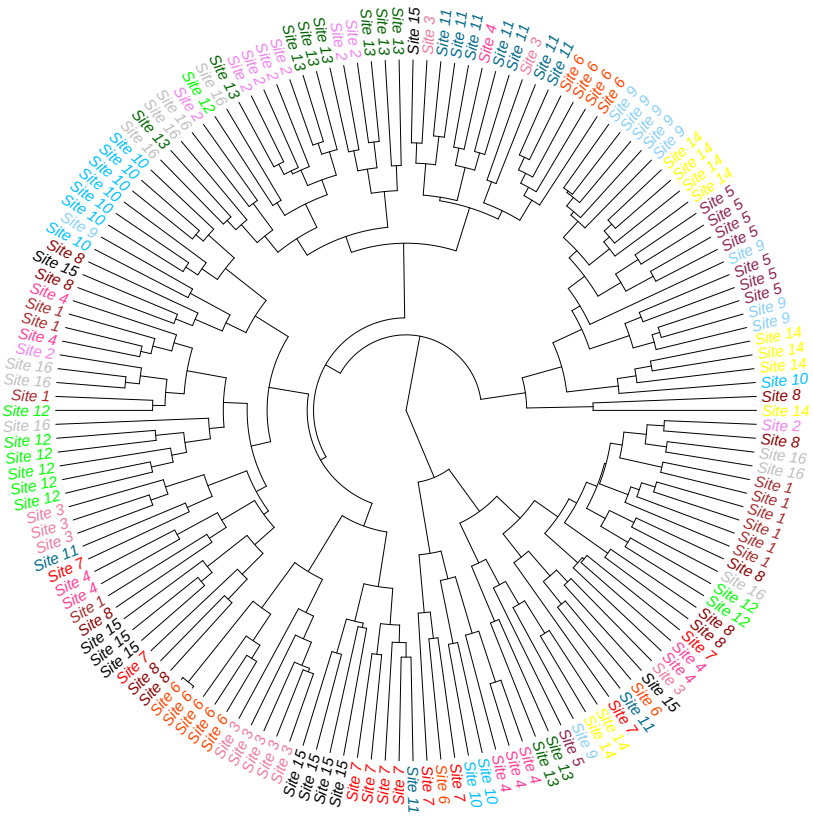

**Figure 4.2:** Hierarchical clustering using correlation distance metric and complete linkage based on the 10% most varying probes. Dendrogram coloured by site and numbered in alphabetic order visualising that the samples were clustered by site.

pollution.

Figure 4.4 shows an example of the principal component analysis. In the left plot it is seen that some discrimination between regions can be found but the overlap was quite large for the confidence ellipses at confidence level 75%. The right plot in the same figure shows the same principal components but this time the samples were marked by pollution level for which the confidence ellipses coincide. Comparison by the left and right subplots tells us that these two principal components divided better between regions than pollution.

Even when performing PCA within each region the overlap for samples from different sites were considerable. It was however possible to find principal components that to some extent divided the samples by site, for an example see figure 4.5. As for the PCA using all samples there were no distinct segregation based upon pollution classification but tendencies could be found. For example figure 4.6 shows Danish samples marked either as reference or medium polluted and polluted where a division between polluted

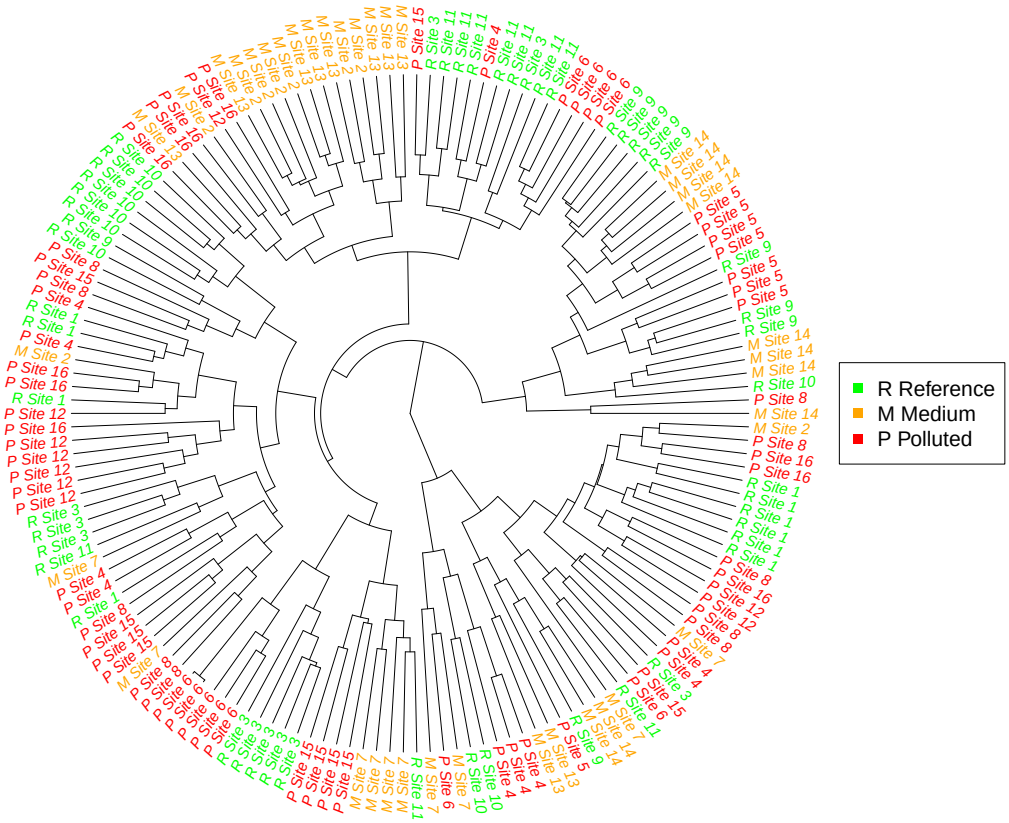


**Figure 4.3:** Hierarchical clustering using correlation distance metric and complete linkage based on the 10% most varying probes. Dendrogram coloured by pollution classification and sites numbered in alphabetic order showing that no clustering based on pollution was found.

and non-polluted samples are seen. This could however be a result of aggregation by site that occasionally looks like grouping by pollution.

## 4.2 Gene rank analysis

For the following analysis the linear statistical model presented in theory has been used for each probe assuming independent normally distributed log transformed gene expressions. The empirical Bayes method in the `limma`-package was then used to estimate the variance (see theory chapter for details).

The results were visualised in volcano plots where each probe is a point represented by its log fold change on the horizontal axis and its false discovery rate adjusted p-value on the vertical axis. These are interpreted such that large magnitudes in horizontal direction are probes which differ a lot in expression between the groups while points close to zero in horizontal direction do not alter in expression between the groups. The

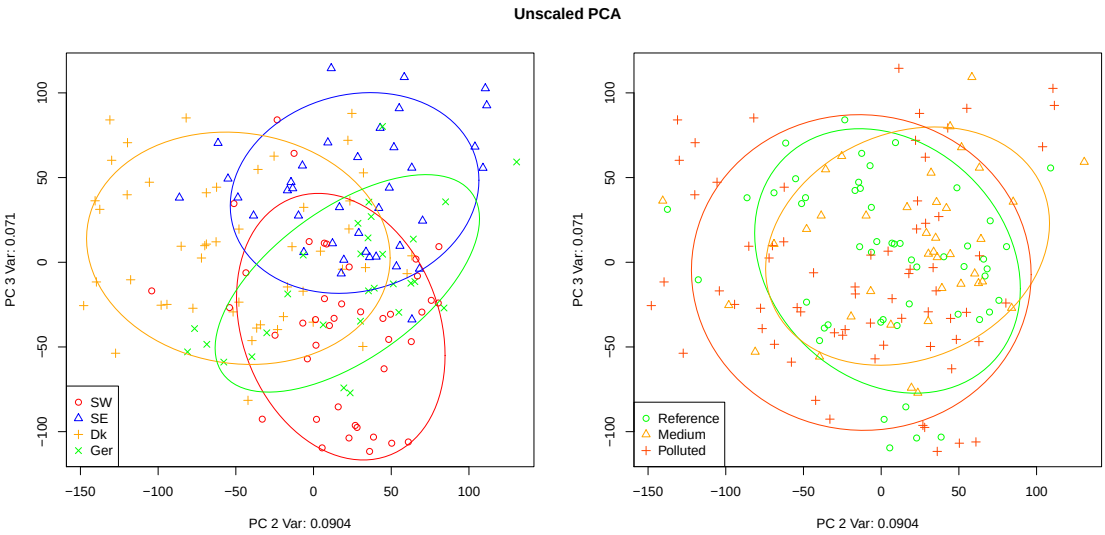


**Figure 4.4:** Unscaled principal component analysis performed on all samples with confidence ellipses at confidence level 75%. Same principal components are visualised in both subplots and samples coloured by region to the left and coloured by pollution to the right. This shows that the samples were to some extent divided into region but not into pollution classification by these components.

vertical axis describes the significance in the changes represented as minus the base ten logarithm of the FDR adjusted p-values. This means that points far up on the vertical axis have very small p-values and high significance. Probes in the top left and right corners are the probes that altered the most and had changes with high significance.

### 4.2.1 Differences in gene expression based on pollution classification

Since reference sites were compared against polluted sites a negative log fold change means that the probe was upregulated in the polluted sites compared to in the reference sites. The effects seen were not very strong and a FDR cut-off at 0.01 gave 1553 significantly differentially expressed probes. Figure 4.7 shows that the comparison between samples from reference sites against samples from polluted sites had more downregulated than upregulated probes in the polluted sites. There was however a larger number of high significance changes for the in polluted sites upregulated probes than the number of high significance changes for the downregulated probes

When moving to regional comparisons the log fold changes and significance were overall larger (figures 4.8-4.11). The number of significantly differentially expressed probes for FDR cut-off 0.01 were 5276 for Swedish west coast, 7085 for Swedish east coast, 3112 for Denmark and 5172 for Germany. The volcano plots were quite symmetrical for all but Swedish east coast (figure 4.9) which had much more high significance probes with positive log fold changes than negative log fold changes. This means that Swedish

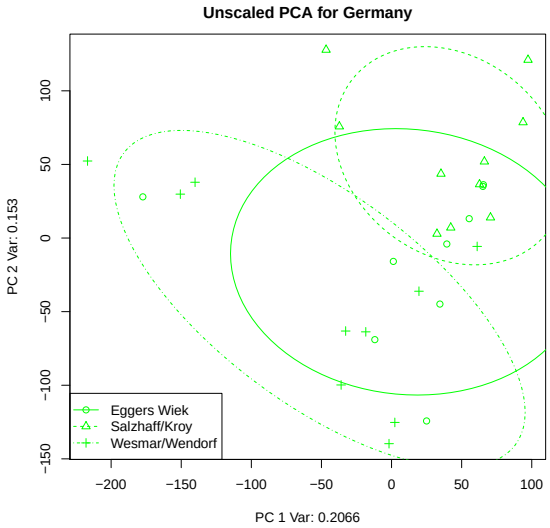


**Figure 4.5:** Unscaled principal component analysis performed on German samples with confidence ellipses at confidence level 75%. Samples are marked by site which shows that these principal components gave some division of the German samples into site.

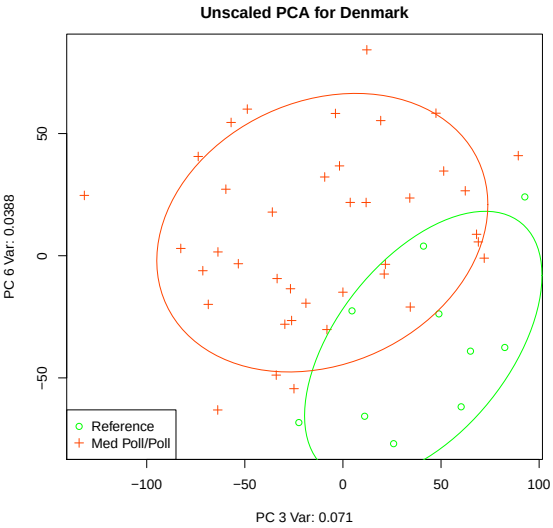


**Figure 4.6:** Unscaled principal component analysis performed on Danish samples with confidence ellipses at confidence level 75%. The samples are coloured by pollution classification where medium polluted sites have been grouped with polluted sites which gave a division of the samples.

east coast seems to have had more significantly downregulated genes in the polluted sites compared to the other regions but the cause for this is not known.

### 4.2.2 Gene expression in relation to the biomarker EROD

#### Correlation between EROD activity and expression of *CYP1A*

The regional correlations were formed by taking average of the site correlations for each of the four probes and are summarised in table 4.1 while all site correlations can be seen in table C.1 in appendix. There was a clear positive correlation between measured expression of *CYP1A* and levels of EROD, measure of the enzyme activity of *CYP1A*. Two of the sites had however negative correlations. In figure 4.12 log transformed EROD values are shown at the horizontal axis and log transformed expression for *CYP1A* for one of the probes on the vertical axis. Linear regression was performed for each site to get a visualisation of the relation at individual level. The mean of the regional correlations were 0.45 for Swedish west coast, 0.35 for Swedish east coast, 0.25 for Denmark and 0.26 for Germany. The correlation was highest for the Swedish west coast. If not for the site Gåsö the correlation for the Swedish east coast would have been high as well. It can be noticed that the German samples had lower EROD levels compared to the other regions but the same levels of *CYP1A* expression.

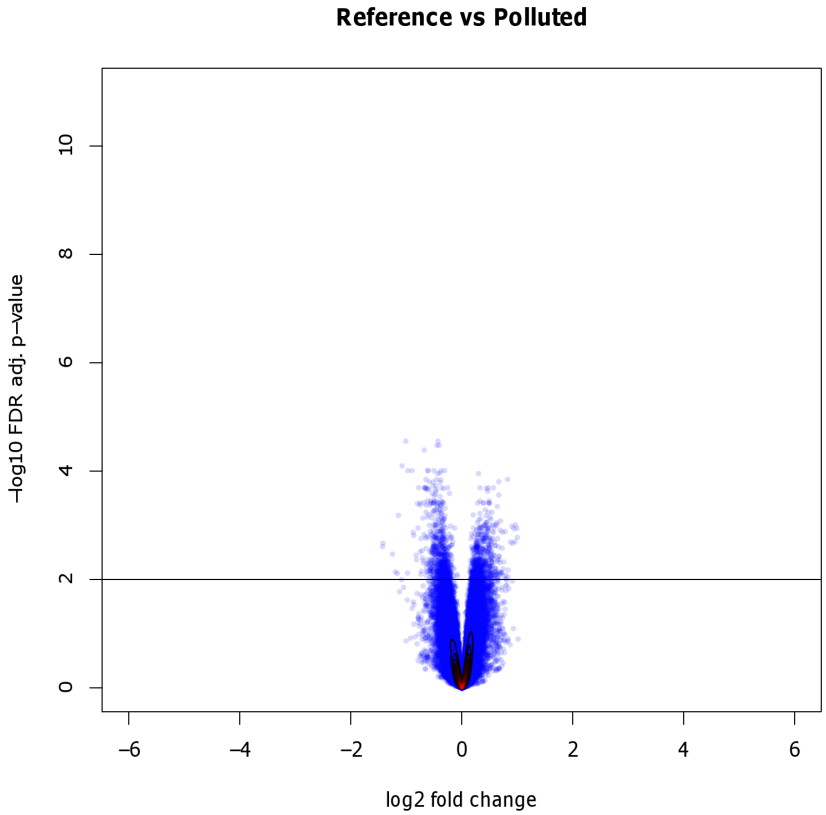


**Figure 4.7:** Volcano plot for comparison of hepatic gene expression for samples from reference sites against polluted sites. A negative log fold change tells that an upregulation was seen in the fish from polluted sites. The line represents a FDR cut-off at 0.01 and shows that small but significant effects were seen.

### Differences in gene expression for samples with high and low EROD activity

As expected all *CYP1A* probes were upregulated in the high EROD group and we saw log fold changes between 0.56 to 1.39 with adjusted p-values between 0.08 and 0.195 (table 4.2). It was however not the probes for gene expression of *CYP1A* that were the probes with most extreme log fold changes even if the groups were chosen based on EROD. Instead other probes had log fold changes with very large magnitudes. When using a FDR cut-off at 0.01 there were 1178 significantly changed probes, the probes for *CYP1A* not being any of them.

There were less significantly differentially expressed probes in the comparison of high and low EROD individuals from the Swedish west coast than in the all regions comparison of high and low EROD individuals. The number of significantly differentially expressed probes was 51 for the within Swedish east coast comparison. The log fold changes in the comparison between high and low EROD samples from the Swedish west coast had

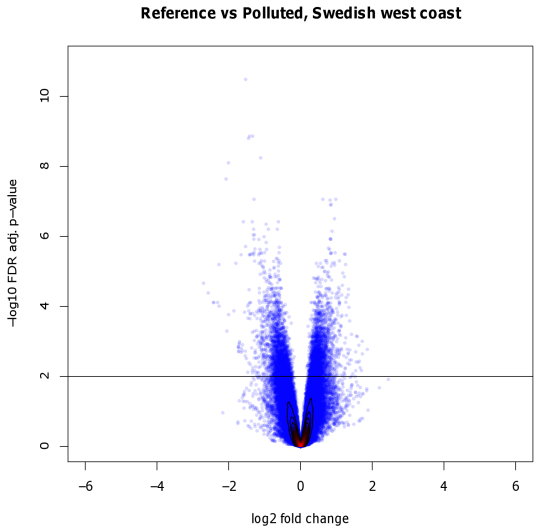


**Figure 4.8:** Volcano plot for comparison of hepatic gene expression for samples from Swedish west coast where samples from reference sites were compared against samples from polluted sites. Negative log fold changes indicate upregulations in fish from the polluted sites and the line represents a FDR cut-off at 0.01. Larger effects were seen here compared to the all regions comparison of reference against polluted.

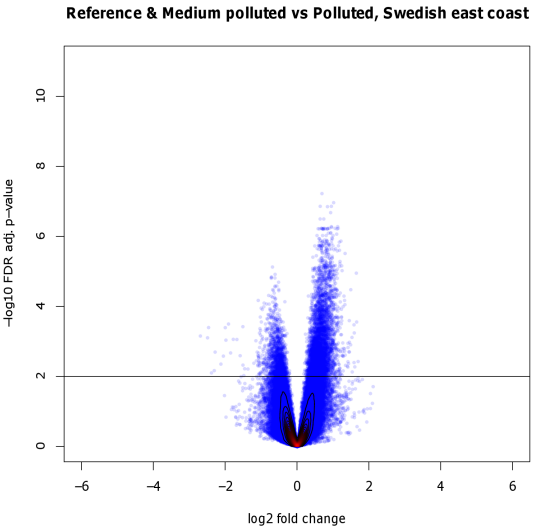


**Figure 4.9:** Volcano plot for comparison of hepatic gene expression for samples from Swedish east coast where samples from reference and medium polluted sites were compared against samples from polluted sites. Negative log fold changes indicate upregulations in fish from the polluted sites and the line represents a FDR cut-off at 0.01. A large proportion of the probes were downregulated in the polluted sites.

**Table 4.1:** Table showing the regional correlations of log transformed EROD levels and log transformed gene expression of *CYP1A* measured using four different probes. All regions had clear positive correlations for all probes. The strongest correlation was found at the Swedish west coast.

| | Probe 13544 | Probe 13545 | Probe 13546 | Probe 13547 | Mean |
|---|---|---|---|---|---|
| SW | 0.386 | 0.473 | 0.375 | 0.580 | 0.453 |
| SE | 0.292 | 0.347 | 0.401 | 0.368 | 0.352 |
| Dk | 0.220 | 0.210 | 0.332 | 0.250 | 0.253 |
| Ger | 0.270 | 0.267 | 0.171 | 0.322 | 0.257 |

larger magnitudes than in the all regions comparison and also the log fold changes for the *CYP1A* probes were larger. The p-values were smaller but still not significant (table 4.3). Notice the difference in rank of the *CYP1A* probes between the all sample comparison and the within Swedish west coast comparison (table 4.2 and table 4.3).

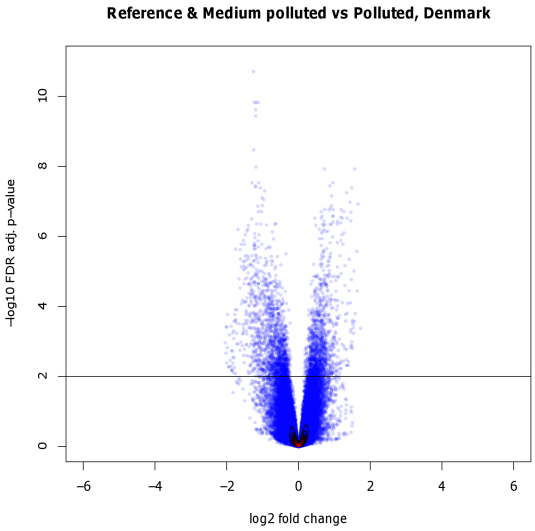


**Figure 4.10:** Volcano plot for comparison of hepatic gene expression for samples from Denmark where samples from reference and medium polluted sites were compared against samples from polluted sites. Negative log fold changes indicate upregulations in fish from the polluted sites and the line represents a FDR cut-off at 0.01 showing large and significant changes.

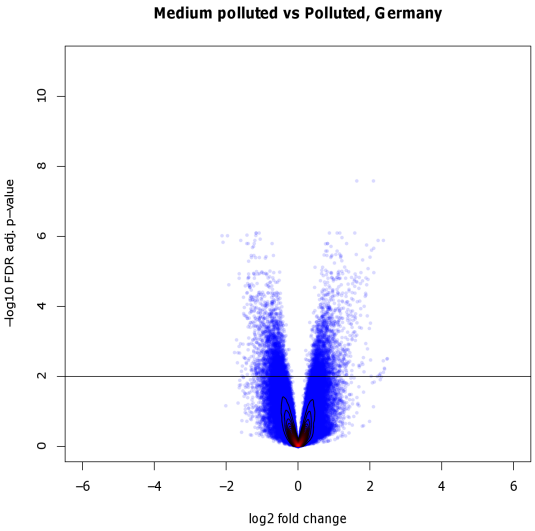


**Figure 4.11:** Volcano plot for comparison of hepatic gene expression for samples from Germany where samples from medium polluted sites were compared against samples from polluted sites. Negative log fold changes indicate upregulations in fish from the polluted sites and the line represents a FDR cut-off at 0.01. Large and significant changes were seen despite the absence of reference sites.

**Table 4.2:** Table of changes for the probes measuring expression of *CYP1A* in the comparison between the 5% of the individuals with highest levels of EROD against the 5 % with lowest EROD levels. There was large upregulation of the *CYP1A* gene in the high EROD group but using a FDR cut-off at 0.01 none of these changes were considered as significant.

| Probe | Rank | log FC | FDR adj. p-val |
|---|---|---|---|
| 13547 | 12372 | 1.39 | 0.08 |
| 13545 | 26170 | 0.56 | 0.18 |
| 13544 | 27235 | 0.41 | 0.19 |
| 13546 | 27931 | 1.01 | 0.20 |

**Table 4.3:** Table of changes for the probes measuring expression of *CYP1A* in the comparison between high and low EROD individuals at the Swedish west coast. A larger upregulation was seen in this comparison than in the all regions comparison based on EROD but still, using a FDR cut-off at 0.01, none of the changes were considered to be significant.

| Probe | Rank | log FC | FDR adj. p-val |
|---|---|---|---|
| 13546 | 3206 | 2.18 | 0.08 |
| 13544 | 4213 | 0.95 | 0.09 |
| 13547 | 4472 | 2.45 | 0.10 |
| 13545 | 9442 | 1.15 | 0.15 |

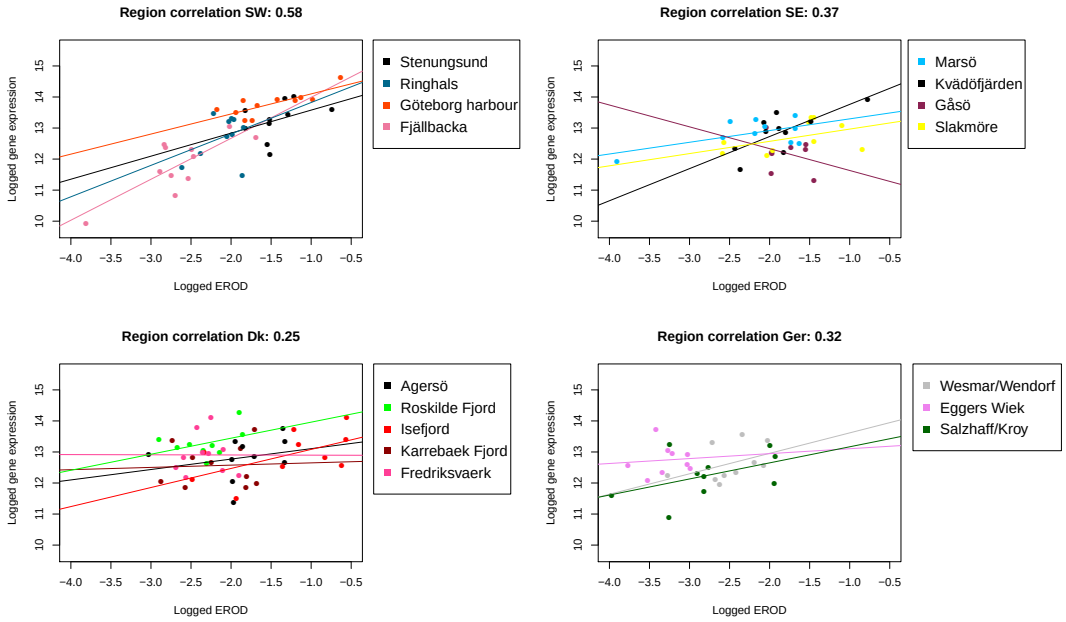


**Figure 4.12:** Scatterplots of log transformed EROD levels against log transformed gene expression for one of the probes measuring expression of *CYP1A*. Linear regression has been plotted for each site. The region correlation is the average of this probe's within-site correlations for each region. All sites but Frederiksværk and Gåsö had positive correlations and most correlations were strong. Notice the low levels of EROD in the German samples compared to the other regions.

### 4.2.3 Difference in gene expression based on reproduction success

A change in gene expression as a result of external impact is on its own not a bad sign for the fish. Instead it could be a confirmation that the species has the necessary tools to cope with a changing environment. What it all comes down to in the end is the possibility to successfully reproduce. By looking at differences in gene expression in relation to rate of abnormal fry we got a linking between the processes at mRNA level in the fish and impact on reproduction ability.

Figure 4.14 shows the volcano plot for the comparison based on rate of abnormal fry. There was a high density of probes with low significance and close to zero log fold change. The number of significantly changed probes using a FDR cut-off at 0.01 was 100. Very few probes had a log fold change with larger magnitude than one which makes this comparison the one with the smallest differences in gene expression. For the differences that were found there was approximately the same amount of upregulated as downregulated genes.

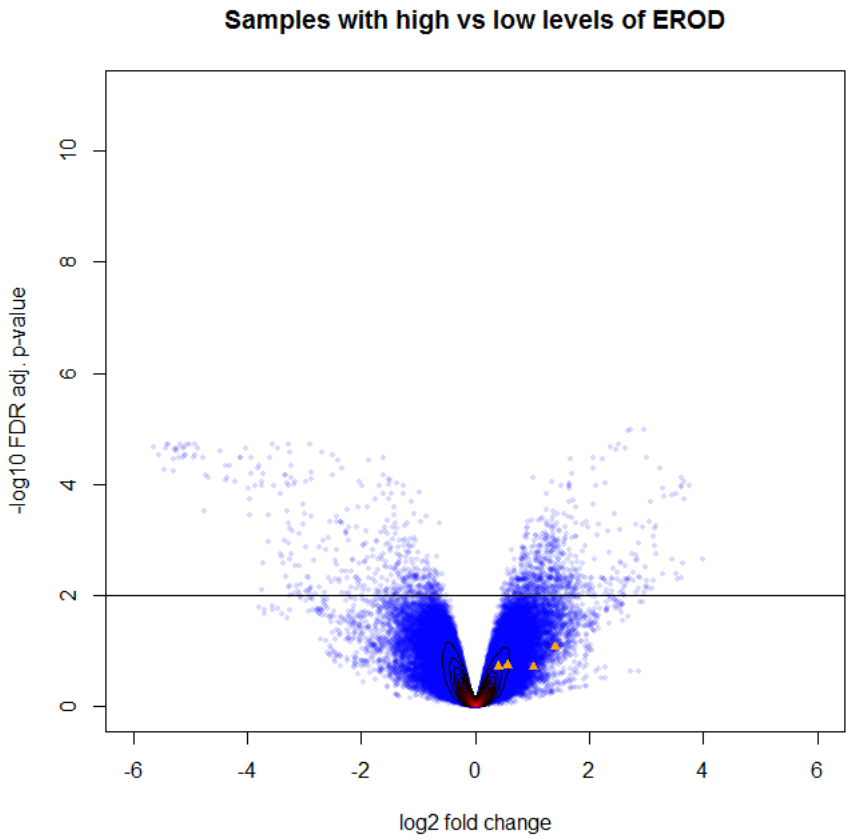


**Figure 4.13:** Volcano plot for comparison of hepatic gene expression for the 5% of the individuals with highest levels of EROD against the 5 % with lowest EROD levles. A positive log fold change tells that an upregulation was seen in fish in the high EROD group. Large effects were found and the line representing a FDR cut-off at 0.01 shows that many of these were significant but the four probes measuring expression of *CYP1A*, listed in table 4.2 and here marked with triangles, were not.

## 4.3 Within-site correlation

To investigate the within-site relationship between the samples a within-site correlation parameter $\rho$ was estimated together with an estimation of the variance $\sigma^2$ for each probe.

The estimated variances ranged between 0.006 and 10.3 with median 0.183. In empirical Bayes for microarray data the variances are assumed to be from an inverse gamma distribution which the histogram of $\widehat{\sigma^2}$ (figure 4.15) follows. This is consistent with results of Smyth et al. (2003) and Kristiansson et al. (2005) which concluded that inverse gamma is a good distribution assumption in microarray analysis. In the histogram the 412 probes with $\widehat{\sigma^2}$ larger than 3 have been left out.

Often the within-site correlation is assumed to be zero but the clustering gave indices

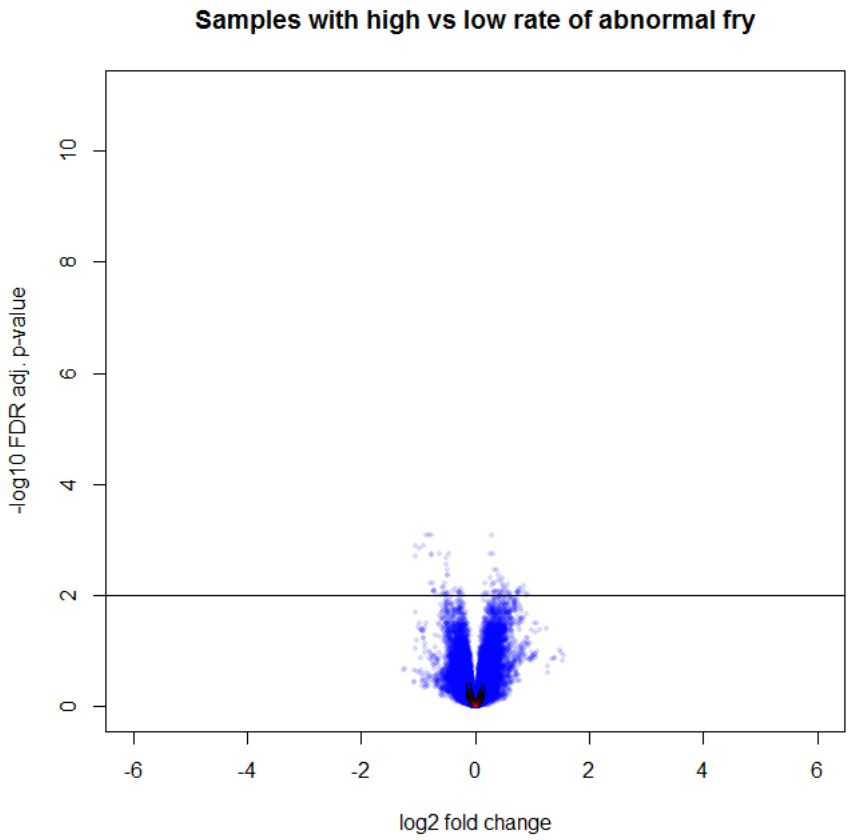


**Figure 4.14:** Volcano plot for comparison of hepatic gene expression based on rate of abnormal fry which was defined to be the number of dead and abnormal fry didvided by the total brood size. Two large groups contaning almost all samples were formed and the high rate of abnormal fry group was compared aganist the low rate group. Small effects were seen and the line representing a FDR cut-off at 0.01 shows that few probes were significantly changed.

that there are apparent similarities between samples belonging to the same site and here we saw that independence between samples from the same site was most of the times not the case. A majority of genes seem to have been affected by within-site correlation. The within-site correlation ranged between -0.09 and 0.840 with median 0.175 (figure 4.16). Out of the 135 091 probes the number of negative $\widehat{\rho}$ was 4980.

The relation between the variance and the within-site correlation for each probe is seen in figure 4.17. The highest density of probes was found for $\widehat{\sigma^2}$ close to 0.11 and $\widehat{\rho}$ close to 0.16. Noticeable is that for both high and low variance probes large values of $\widehat{\rho}$ were found and the within-site correlations were spread over the same range.

There were numerous probes with large within-site dependence. The probes that represent known genes and with largest $\widehat{\rho}$ are listed in table 4.4, for a more extensive

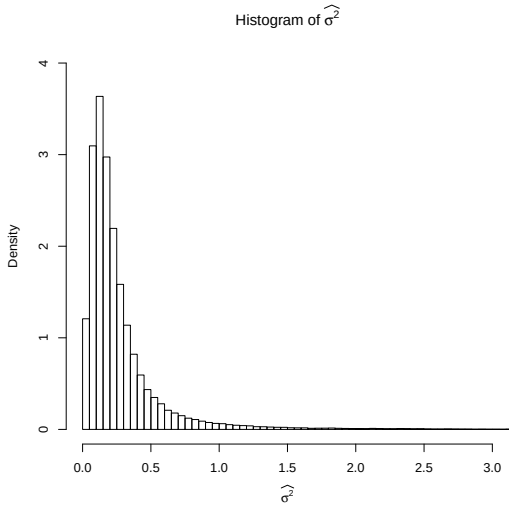


**Figure 4.15:** Truncated histogram of estimated variance, $\widehat{\sigma^2}$, of each probe. The histogram seems to follow an inverse gamma distribution which is consistent with the assumption made in empirical Bayes.

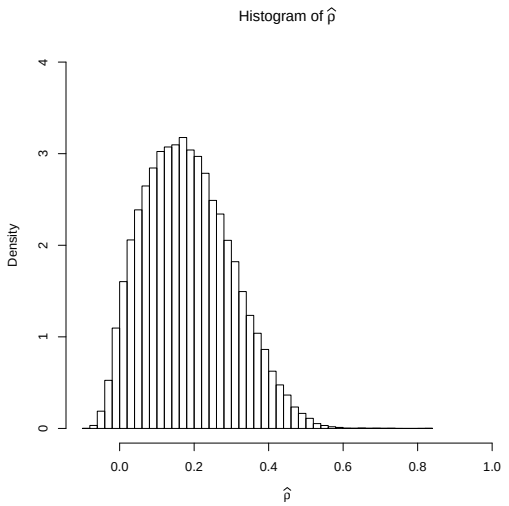


**Figure 4.16:** Histogram of estimated within-site correlation, $\widehat{\rho}$, for each probe. This shows that many genes seem to have been affected by within-site correlation and several strong within-site correlations were seen.

listing see table D.1 in appendix. At the top we find genes associated with tryptase-2 precursor, ATP binding cassette, ice structuring protein and antifreeze proteins. Genes with smallest $\widehat{\rho}$ were connected to ribosomal functions.

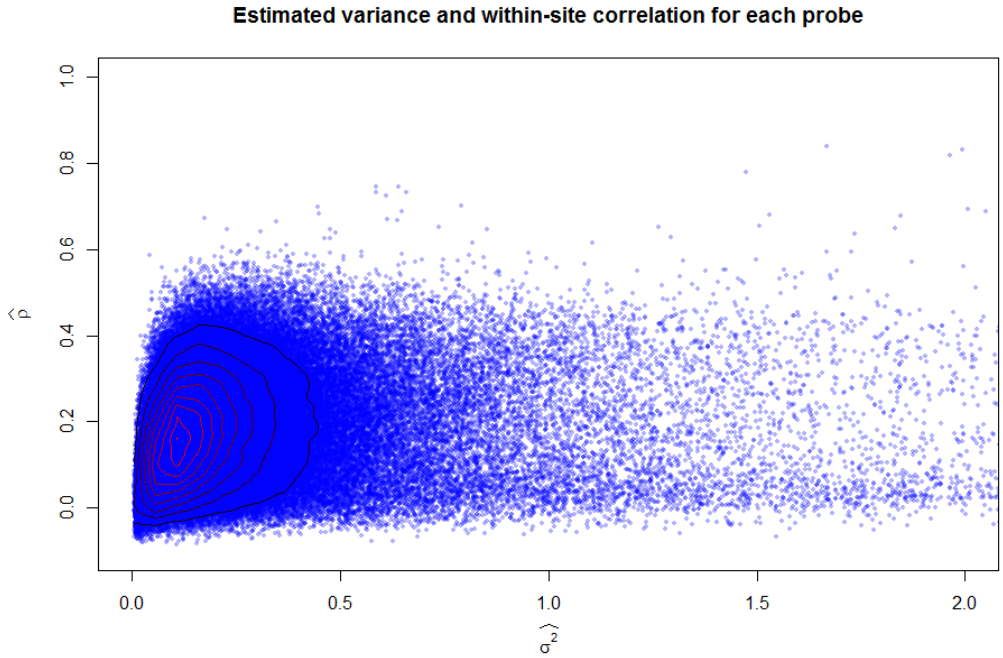


**Figure 4.17:** Scatterplot showing the relation between estimated variance, $\widehat{\sigma^2}$, and estimated within-site correlation, $\widehat{\rho}$, for the probes. The dispersion of the points, each point representing a probe, shows that the range of estimated within-site correlations was the same for all variances.

**Table 4.4:** List of the probes representing known genes and which had largest estimated within-site correlations. Genes appearing multiple times are represented as the probe with largest estimated within-site correlation.

| Rank | $\hat{\rho}$ | Annotation |
|---|---|---|
| 1 | 0.84 | tryptase-2 precursor |
| 15 | 0.73 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 22 | 0.72 | ANP2_ANALU Type-3 ice-structuring protein |
| 31 | 0.69 | type III antifreeze protein |
| 45 | 0.66 | cytochrome P450, family 2, subfamily J, polypeptide 2 |
| 51 | 0.65 | type III antifreeze protein |
| 58 | 0.64 | protein FAM110B-like |
| 71 | 0.61 | PAX3 and PAX7 binding protein |
| 78 | 0.60 | RNA-binding protein 6-like |
| 81 | 0.59 | PAX3- and PAX7-binding protein |
| 82 | 0.59 | hypothetical protein |
| 87 | 0.59 | splicing regulatory glutamine/lysine-rich protein |
| 99 | 0.58 | ANP3_ZOAAM Ice-structuring protein |

# 5

# Discussion

In this study hepatic gene expression microarray data from wild fish of the species eelpout was analysed. The fish, 158 in total, were sampled at 16 sites located in four large geographical regions in the Baltic Sea. The aim was to understand microarray gene expression patterns in the fish eelpout and how they change in connection to pollution. Discrimination between samples based on field site, population and exposure of pollution was investigated. Differences in gene expression of the fish in relation to exposure of contamination and the biomarkers EROD and reproduction success were identified using linear model and empirical Bayes. Using numerical optimization under the assumption of normal distributed gene expressions the within-site correlation between the gene expressions was estimated for each probe separately. Clustering analysis showed that fish from the same region and site had a high tendency to group together. The comparison between fish from reference and polluted sites had significant differences in gene expression but the effects were in general small. This is likely an indication that there were few shared differentially expressed genes between the polluted sites. The effects where however larger and more significant for the regional comparisons based pollution classification which could either be due to differences in population that was blurring out the signal or that the pollution in each region was more homogeneous. The large effects identified in connection to region and site may indicate that ecological factors and population parameters had a substantial impact on the observed gene expression profiles. Large and significant effects were also seen when comparing fish with low and high values of the known biomarker EROD. We observed strong correlations between measured gene expression for *CYP1A* gene and levels of EROD. The strength of the correlation varied between regions and the highest correlation was found at the Swedish west coast. There were, however, small effects on the gene expression between fish with high and low rate of abnormal fry and a more thorough investigation of the reproduction success and its connection to changes in gene expression should be performed. We assessed the independence between sampled fish and found that several genes had a high

within-site correlation. This underlines the importance to use a model that models site specific effects and/or correlations.

Large-scale field studies will always have factors that affect the sampling procedure and retrieving of data. Such factors are sampling bias and systematic effects caused by different sampling circumstances for the regions. An example from this study is that testing was done at site for some regions and for others the samples needed to be moved before testing was performed. It is often hard to remove the impact of these factors and larger and more complex studies will typically result in more varying conditions. The extensive set-up of this study is unique and much care has been put into making the sampling and testing procedure for all sites as similar as possible such as performing all microarray and enzyme analyses by the same laboratory.

The general appearance was analysed by hierarchical clustering, using correlation distance metric and complete linkage, as well as by unscaled PCA. This showed that site and region had a high tendency to group together. According to the hierarchical clustering of the samples, region and site seem to be the factors that had largest influence on the general appearance of the gene expression profiles of the fish. For discrimination into site and region the high variance probes contained more information than the low variance probes. Three of the regions, namely Swedish west coast, Swedish east coast and Germany, had clear regional grouping while the Danish fish did not seem to stand out compared to fish from the other regions. The variable reduction into two variables using principal component analysis was not able to capture the division into region or site in a credible way. Summing up the explanatory analysis there was apparent division into region and site but no perfect segregation was seen. It could also be concluded that neither the clustering nor the PCA showed clear grouping based on pollution which indicate that there were likely not a large number of genes collectively induced by different types of contamination.

Often large-scale gene expression studies address short term hypotheses for controlled experiments in laboratory. In contrast, this study considered wild sampled fish that, due to its stationary, had lived at the same site for a long time. Therefore, the fish was assumed to have been chronically exposed to the ecological and anthropogenic conditions at the site and many of the short term stress responses could be assumed to have settled down. Using a linear model, gene expression profiles between samples from reference and polluted sites were compared. There were significant differences in gene expression but the effects were in general small. The effects were however larger and more significant when the pollution classification based comparisons were performed within each region. The in general small effects in the comparison between fish from reference and polluted sites could be due to various reasons. One of them is the difficulty of prior pollution classification. It is for example possible that some reference sites are not completely clean. It is also possible that the level of exposure differed between the polluted sites, which may induce variation in the gene response. The regional comparisons showed larger effects despite that reference and medium polluted sites now were combined into one group and compared against polluted sites. This indicated that there was another cause for the small effects in the all regions comparison than solely the pollution classification.

It should also be noted that the polluted sites are likely to have been exposed in different ways which led to differences in gene expression profiles within the group of polluted. This implied that there were significant effects in the data but either the differences in population was blurring out the signal or there was more similar types of pollution within the same region.

The regional averages of the within-site Pearson correlation coefficients between measured expression of *CYP1A* and levels of EROD, measure of the enzyme activity of *CYP1A*, showed strong positive correlations. The clear positive correlations we saw for all regions, average between 0.26 and 0.45, lies in line with the correlations seen in Guo et al. (2008) which were between 0.21 and 0.64. In Guo et al. (2008) the relation between mRNA levels and protein expression in human circulating monocytes was investigated and even though this study investigated mRNA levels in the liver of eelpouts compared to activity of protein this gave an indication of what to be expected. Considering the difference in measuring protein expression and activity of protein it seems as the correlations seen here were almost a bit stronger than expected especially for the Swedish west coast. It can be noticed that the German samples had lower levels of EROD compared to the other regions but the same measured expression of *CYP1A*. This could have been a cause of inhibition or population effects.

The gene expression profiles of the 5% of the individuals with highest levels of EROD were compared against the 5% of the individuals with lowest levels of EROD. It was found to be large and significant differences in gene expression between the groups. EROD is used as a general biomarker for pollution but its biological function is complex and involves detoxification of foreign chemical substances which includes both ecological impact and anthropogenic pollution. The more specialized biological function of EROD than as protection against general pollution could be the reason for the larger effects in the comparison based on EROD than the effects seen in the comparisons based on pollution classification. We saw upregulation for many genes in the high EROD group and it is likely that other genes than *CYP1A* also were upregulated due to the same external influence.

As an indicator of reproduction success the rate of abnormal fry was chosen and defined to be the number of dead and abnormal fry divided by the total brood size. Two large groups were formed containing almost all samples and fish with high rates of abnormal fry were compared against samples with low rates. Small differences in the gene expression profiles were seen for these groups. The reproduction process involves many biological systems and therefore it might be too complex to be captured by a comparison of this type. This would be the case if having many different reasons for the high abnormal rate which means that different genes are changing for the samples within the same group. It is likely that the number of fry and the rate of abnormal fry were affected by physiological and ecological parameters. The number of fry was for example strongly correlated with the length of the fish which was biased by region. Furthermore, small brood sizes tended to have a lower rate of abnormal fry. There was therefore no clear way how to choose the groups for the reproduction comparison. Based on the comparison performed it cannot be concluded that the data do not contain information

about the relationship between gene differences and reproduction success but rather that the comparison needs to be done more carefully.

By maximizing the normal log likelihood for each probe with respect to variance and within-site correlation the dependencies between the samples were investigated. This was performed under the assumptions that the parameters were the same for all sites and that there were no correlation between samples from different sites. It is a common assumption in microarray analysis using linear models that the samples are independent but it was here shown that many genes were affected by within-site correlation and that it was strong for several genes. Due to this dependence the p-values calculated under the assumption of independence cannot be assured to be accurate. For many biological question the most interesting is however the ordering of the genes and even if the magnitude of the p-values was not reliable due to correlation the ranking of the probes may be more robust.Two approaches that could be used to take care of the within-site correlation are either by modifying the covariance matrix in the linear model or to use mixed models. Out of the roughly 135 000 probes 4% had negative within-site correlation, all with small magnitudes. Among the genes with negative within-site correlation there was an overrepresentation of genes associated with basal functions such as ribosomal processes. These processes are thought not to be easily affected by environmental impact and therefore not subject of within-site correlation. Considering the large number of estimations performed, one for each probe, the negative within-site correlations are likely to be due to variations around zero seen when having a large number of independent samples. To be able to make inference on individual level the dependence should be taken into account. One way to interpret the impact of within-site correlation is in the context of sampling design. If a specific gene is to be investigated and its within-site correlation is low fish from the same site can be considered to be independent replicates.

To conclude, this work shows that large-scale gene expression is a viable tool for assessing differentially expressed genes in wild fish. Our results suggest there are small effects on gene expression connected to pollution and rate of abnormal fry in the case of large-scale sampling and that factors such as population might interfere. We could, however, identify large effects connected to region and site which may indicate that ecological factors and population parameters have a substantial impact on gene expression profiles. Finally, we demonstrated that genes tend to have similar expression levels in fish from the same site which underlines the need of taking dependences into account if inference on individual level is to be made.

# Bibliography


Bruce Alberts. *Molecular biology of the cell*. Garland Science, New York, 2002. ISBN 0815340729.

E Albertsson, J Gercken, J Strand, N Asker, S Bergek, I Holmqvist, U Kammann, J. Parkkonen, and L Förlin. Biomarker responses in eelpout from different coastal sites in Sweden, Denmark and Germany. *Manuscript*, 2011.

Steven Arnold. *The theory of linear models and multivariate analysis*. Wiley, New York, 1981. ISBN 0471050652.

Dennis Bernstein. *Matrix mathematics theory, facts, and formulas*. Princeton University Press, Princeton, N.J, 2009. ISBN 9780691132877.

F. Falciani, A.M. Diab, V. Sabine, T.D. Williams, F. Ortega, S.G. George, and J.K. Chipman. Hepatic transcriptomic profiles of European flounder (Platichthys flesus) from field sites and computational approaches to predict site from stress gene responses following exposure to model toxicants. *Aquatic Toxicology*, 90(2):92–101, 2008.

Yanfang Guo, Peng Xiao, Shufeng Lei, Feiyan Deng, Gary Guishan Xiao, Yaozhong Liu, Xiangding Chen, Liming Li, Shan Wu, Yuan Chen, Hui Jiang, Lijun Tan, Jingyun Xie, Xuezhen Zhu, Songping Liang, and Hongwen Deng. How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. 40 (5):426–436, 2008. doi: 10.1111/j.1745-7270.2008.00418.x.

Trevor Hastie. *The elements of statistical learning data mining, inference, and prediction*. Springer, New York, 2009. ISBN 0387848576.

J.E. Hedman, H. Rüdel, J. Gercken, S. Bergek, J. Strand, M. Quack, M. Appelberg, L. Förlin, A. Tuvikene, and A. Bignert. Eelpout (Zoarces viviparus) in marine environmental monitoring. *Marine Pollution Bulletin*, 62(10):2015–2029, 2011.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall series in statistics. 1998. ISBN 9780138341947.

Erik Kristiansson, Anders Sjögren, Mats Rudemo, and Olle Nerman. Weighted Analysis of Paired Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1–47, 2005.

M.K. Sellin Jeffries, A.C. Mehinto, B.J. Carter, N.D. Denslow, and A.S. Kolok. Taking microarrays to the field: Differential hepatic gene expression of caged fathead minnows from Nebraska watersheds. *Environmental Science and Technology*, 46(3):1877–1885, 2012.

A. Sjögren, E. Kristiansson, M. Rudemo, and O. Nerman. Weighted analysis of general microarray experiments. *BMC Bioinformatics*, 8, 2007.

G. K. Smyth. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 23. New York, 2005.

G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

GK Smyth, NP Thorne, and J Wettenhall. LIMMA: Linear Models for Microarray Data User's Guide, 2003. *URL http://www. bioconductor. org*, 2003.

J.J. Whyte, R.E. Jung, C.J. Schmitt, and D.E. Tillitt. Ethoxyresorufin-O-deethylase (EROD) activity in fish as a biomarker of chemical exposure. *Critical Reviews in Toxicology*, 30(4):347–570, 2000.

T.D. Williams, I.M. Davies, H. Wu, A.M. Diab, L. Webster, M.R. Viant, J.K. Chipman, M.J. Leaver, S.G. George, C.F. Moffat, and C.D. Robinson. Molecular responses of European flounder (Platichthys flesus) chronically exposed to contaminated estuarine sediments. *Chemosphere*, 108:152–158, 2014.

# A

# Appendix

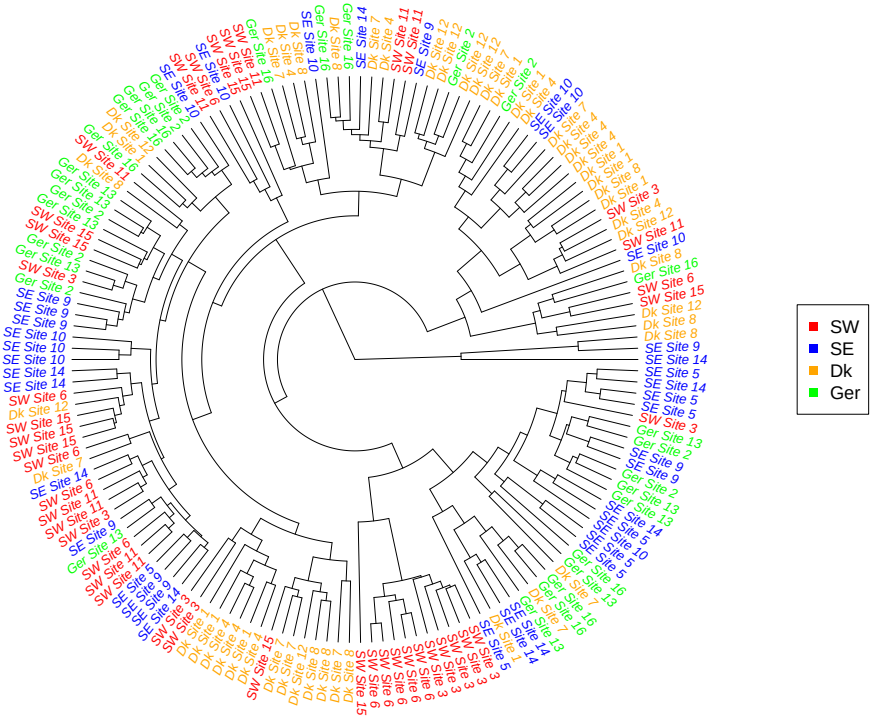


**Figure A.1:** Hierarchical clustering of the samples based on correlation distance metric and complete linkage using the 10% least varying probes. Dendrogram coloured by region and sites numbered in alphabetic order showing less grouping than was seen for clustering using the 10% most varying probes.

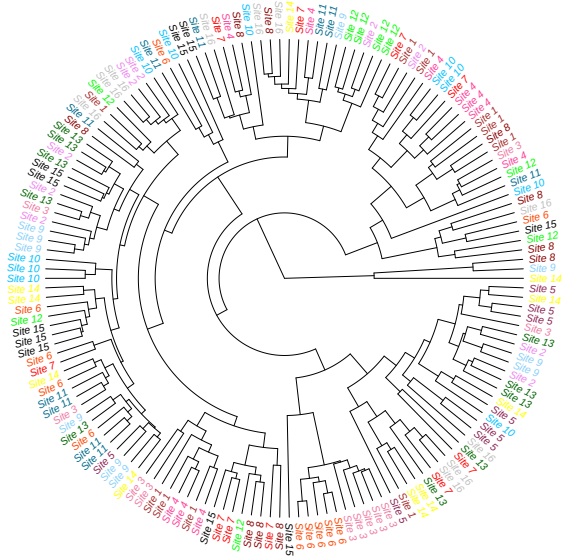

**Figure A.2:** Hierarchical clustering of the samples based on correlation distance metric and complete linkage using the 10% least varying probes. Dendrogram coloured by site and numbered in alphabetic order. Less grouping was seen here than for the clustering using the 10% most varying probes.

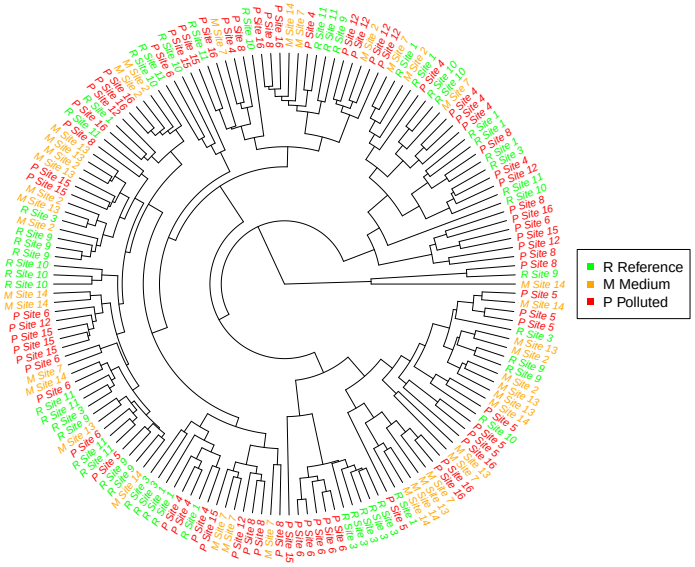


**Figure A.3:** Hierarchical clustering of the samples based on correlation distance metric and complete linkage using the 10% least varying probes. Dendrogram coloured by pollution classification and sites numbered in alphabetic order. Neither this nor the clustering using the 10% most varying probes gave any division based on pollution classification.

# B

# Appendix

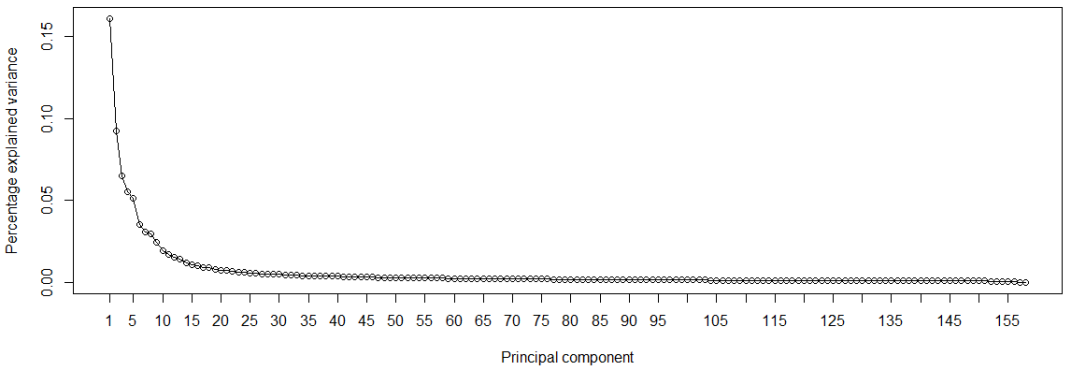


**Figure B.1:** The percentage of variance that was explained by the principal components in the unscaled principal component analysis performed on samples from all regions.

# C

# Appendix

**Table C.1:** Table showing the within-site Pearson correlation coefficients for the correlation between log transformed EROD levels and log transformed gene expression of *CYP1A*. There were clear positive correlations for almost all sites and strongest correlations were seen for Swedish sites.

| | Probe 13544 | Probe 13545 | Probe 13546 | Probe 13547 | Mean | Sd |
|---|---|---|---|---|---|---|
| Agersö | 0.291 | 0.235 | 0.438 | 0.244 | 0.302 | 0.094 |
| Eggers Wiek | -0.258 | -0.003 | -0.195 | 0.085 | -0.093 | 0.161 |
| Fjällbacka | 0.776 | 0.808 | 0.825 | 0.783 | 0.798 | 0.022 |
| Fredriksvaerk | -0.031 | -0.009 | 0.129 | -0.002 | 0.022 | 0.073 |
| Gåsö | -0.100 | -0.219 | -0.018 | -0.312 | -0.162 | 0.130 |
| Göteborg harbour | 0.579 | 0.565 | 0.534 | 0.794 | 0.618 | 0.119 |
| Isefjord | 0.471 | 0.305 | 0.450 | 0.573 | 0.450 | 0.111 |
| Karrebaek Fjord | -0.167 | 0.204 | -0.017 | 0.050 | 0.017 | 0.154 |
| Kvädöfjärden | 0.525 | 0.717 | 0.643 | 0.736 | 0.655 | 0.096 |
| Marsö | 0.640 | 0.656 | 0.663 | 0.583 | 0.636 | 0.036 |
| Ringhals | -0.073 | 0.244 | 0.042 | 0.354 | 0.142 | 0.193 |
| Roskilde Fjord | 0.538 | 0.318 | 0.661 | 0.384 | 0.475 | 0.155 |
| Salzhaff/Kroy | 0.674 | 0.367 | 0.492 | 0.460 | 0.498 | 0.129 |
| Slakmöre | 0.105 | 0.236 | 0.315 | 0.466 | 0.280 | 0.151 |
| Stenungsund | 0.262 | 0.277 | 0.097 | 0.388 | 0.256 | 0.120 |
| Wesmar/Wendorf | 0.393 | 0.436 | 0.217 | 0.421 | 0.367 | 0.101 |
| Mean | 0.289 | 0.321 | 0.330 | 0.375 | | |
| Sd | 0.335 | 0.274 | 0.302 | 0.305 | | |

# D

# Appendix

**Table D.1:** The within-site correlation was estimated for each probe under the assumption of no between-site correlation. The probes representing known genes and had largest estimated within-site correlations, $\widehat{\rho}$, are listed below.

| Rank | $\widehat{\rho}$ | Annotation |
|---|---|---|
| 1 | 0.84 | tryptase-2 precursor |
| 5 | 0.83 | tryptase-2 precursor |
| 8 | 0.81 | tryptase-2 precursor |
| 9 | 0.81 | tryptase-2 precursor |
| 15 | 0.73 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 22 | 0.72 | ANP2_ANALU Type-3 ice-structuring protein |
| 23 | 0.71 | ANP2_ANALU Type-3 ice-structuring protein |
| 26 | 0.70 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 28 | 0.70 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 31 | 0.69 | type III antifreeze protein |
| 35 | 0.68 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 36 | 0.68 | type III antifreeze protein |
| 39 | 0.678 | ANP2_ANALU Type-3 ice-structuring protein |
| 42 | 0.67 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 45 | 0.66 | cytochrome P450, family 2, subfamily J, polypeptide 2 |
| 46 | 0.66 | ANP2_ANALU Type-3 ice-structuring protein |
| 48 | 0.65 | cytochrome P450, family 2, subfamily J, polypeptide 2 |
| 49 | 0.65 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 50 | 0.65 | cytochrome P450, family 2, subfamily J, polypeptide 2 |
| 51 | 0.65 | type III antifreeze protein |
| 54 | 0.65 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 56 | 0.64 | type III antifreeze protein |
| 58 | 0.64 | protein FAM110B-like |
| 59 | 0.64 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 63 | 0.63 | ATP-binding cassette, sub-family B (MDR/TAP), member 6a |
| 65 | 0.62 | cytochrome P450, family 2, subfamily J, polypeptide 2 |
| 70 | 0.61 | type III antifreeze protein |
| 71 | 0.61 | PAX3 and PAX7 binding protein |
| 72 | 0.61 | type III antifreeze protein |
| 73 | 0.61 | type III antifreeze protein |
| 75 | 0.60 | PAX3 and PAX7 binding protein |