Biomarker Discovery Using Statistical and Machine Learning Approaches on Gene Expression Data

Xiaokang Zhang

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2020

UNIVERSITY OF BERGEN

Biomarker Discovery Using Statistical and Machine Learning Approaches on Gene Expression Data

Xiaokang Zhang

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 30.10.2020

Year: 2020

Title: Biomarker Discovery Using Statistical and Machine Learning Approaches on Gene Expression Data

Name: Xiaokang Zhang

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

The work presented in the thesis was funded by University of Bergen and carried out at the Computational Biology Unit (CBU) at the Department of Informatics of University of Bergen. Half a year of my PhD was carried out in the Systems Biology Research Group at the Department of Bioengineering of University of California, San Diego.

This thesis was supervised by Professor Inge Jonassen employed at the CBU and Professor Anders Goksøyr employed at the Department of Biological Sciences of University of Bergen.

I have been affiliated with the National Research School in Bioinformatics, Biostatistics and Systems Biology (NORBIS), the Digital Life Norway Research School (DLNRS), and the Molecular and Computational Biology Research School (MCB).

dCod 1.0
decoding the systems toxicology of atlantic cod

The Research Council of Norway

CBU
COMPUTATIONAL BIOLOGY UNIT

UNIVERSITAS BERGENSIS

UNIVERSITY OF CALIFORNIA
LET THERE BE LIGHT
SAN DIEGO

NORBIS

CENTRE FOR DIGITAL LIFE NORWAY
Research School

MCB
Molecular and Computational Biology Research School

Acknowledgements

I am the type of person who does not always have a plan and it is very easy for me to change my mind even if I got one. This PhD was totally out of my plan. Just as Forrest Gump's mom said, "life was like a box of chocolates. You never know what you're gonna get".

Now I feel very thankful for my decision four years ago. It has been a wonderful time in my life!

My PhD is affiliated with project dCod 1.0: decoding the systems toxicology of Atlantic cod (*Gadus morhua*), which is funded by the Research Council of Norway (grant number 248840). dCod is a big project which involves people from different departments, universities and countries. Collaboration with people from diverse backgrounds was difficult at the beginning but more and more fun came after. I want to thank all the dCoders in Bergen, and special thanks go to Marta Eide, Eileen M. Hanna, and Fekadu Yadetie.

Most of my PhD courses are actually from the research schools, NORBIS, DLNRS and MCB. Without those courses (the knowledge and the credits), I could not finish my PhD. Especially NORBIS of which I am also the student representative, it provides me with enormous resources that I have learnt a lot and made lots of friends.

I feel so lucky to be part of the CBU, with a friendly atmosphere, helpful colleagues, and broad research topics. With experts in different fields, I can always find help whatever research or technical problems I am faced with. Special thanks go to Fatemeh Z. Ghavidel. The Department of Informatics and University of Bergen are also very important to my PhD in providing me such a perfect research environment.

I would like to shout to my supervisor, Inge Jonassen, thank you so much! You are the best! With many titles on the shoulder, he is very busy with many different projects. But he would always ensure enough time for me, following and helping on my research, caring and solving my problems whether research related or not. And my thanks also go to my co-supervisor, Anders Gokøøyr, who is also the leader of dCod

1.0 project. Thanks to him, the project went well and so did my PhD.

I am grateful to get to know so many amazing friends in the past four years in Norway and San Diego, US. You made my life wonderful and joyful. And I also want to thank the old friends who have accompanied me through different stages of my life. Especially those few best friends, you make me believe that no matter where I am, no matter what difficulty I am facing, you are always there for me. Only with that feeling of safety, can I explore my life more bravely.

Last but not least, I would like to thank my family for their generous love. Thanks to my sister taking care of our mom, so that I can study and work far away from home without worry. Special thanks to my dad, who had always wanted me to do a PhD, but also respected my choice when I told him that I did not want to. Even though it is impossible, but I wish he would know that I later on changed my mind and actually did the PhD, part of the reason for him. (最后我要感谢我家人给予我的无私的爱。因为有姐姐照看母亲，我才能在千里之外安心学习工作。特殊的感谢致于我的父亲。他一直以来都想让我读博士，但是当我提出并不想继续读博士时，他完全尊重我的个人选择。虽已成空，但我多希望他能知道我后来改变了主意，如他所愿读了博士，其中部分原因来自于对他的爱。)

Xiaokang Zhang
Bergen, July 2020

Summary

My PhD is affiliated with the dCod 1.0 project (https://www.uib.no/en/dcod): decoding the systems toxicology of Atlantic cod (*Gadus morhua*), which aims to better understand how cods adapt and react to the stressors in the environment. One of the research topics is to discover the biomarkers which discriminate the fish under normal biological status and the ones that are exposed to toxicants.

A biomarker, or biological marker, is an indicator of a biological state in response to an intervention, which can be for example toxic exposure (in toxicology), disease (for example cancer), or drug response (in precision medicine). Biomarker discovery is a very important research topic in toxicology, cancer research, and so on. A good set of biomarkers can give insight into the disease / toxicant response mechanisms and be useful to find if the person has the disease / the fish has been exposed to the toxicant.

On the molecular level, a biomarker could be "genotype" - for instance a single nucleotide variant linked with a particular disease or susceptibility; another biomarker could be the level of expression of a gene or a set of genes. In this thesis we focus on the latter one, aiming to find out the informative genes that can help to distinguish samples from different groups from the gene expression profiling. Several transcriptomics technologies can be used to generate the necessary data, and among them, DNA microarray and RNA sequencing (RNA-Seq) have become the most useful methods for whole transcriptome gene expression profiling. Especially RNA-Seq has become an attractive alternative to microarrays since it was introduced.

Prior to analysis of gene expression, the RNA-Seq data needs to go through a series of processing steps, so a workflow which can automate the process is highly required. Even though many workflows have been proposed to facilitate this process, their application is usually limited to such as model organisms, high-performance computers, computer fluent users, and so on. To fill these gaps, we developed a maximally general RNA-Seq analysis workflow: **R**NA-Seq **A**nalysis **S**nakemake Work**flow** (RASflow), which is applicable to a wide range of applications and requires little programming skills. It takes the sequencing data as input, and maps them to either transcriptome or genome for quantification, and after that the gene expression profile can be achieved which afterwards goes through normalization and statistical tests to find out the differentially expressed genes. This work was presented in **Paper I** and **Paper II**.

Differential expression analysis used in RASflow, together with other univariate methods are widely used in biomarker discovery for their simplicity and interpretability. But they rely on a hypothesis that variables are independent, so they can only identify variables that possess significant information by themselves. However, biological processes usually involve many variables that have complex interactions. Multivariate methods which take the interactions between variables into consideration are therefore also popular for biomarker discovery. To study whether there is a significant advantage of one over the other, we conducted a comparative study of various methods from these two categories and evaluated these methods on two aspects: stability and prediction accuracy, we found that a method's performance is quite data-dependent. This work was presented in **Paper III**.

Since the biomarker discovery methods perform quite differently on different datasets, then how to choose the most appropriate one for a particular dataset? One solution is to use the function perturbation strategy to combine the outputs from multiple methods. Function perturbation is capable of maintaining prediction accuracy compared with the original individual methods, but its stability is not satisfactory enough. On the other hand, data perturbation uses a similar ensemble learning logic: it firstly generates multiple datasets by resampling the original dataset and then combines the results from those datasets. Data perturbation has been proven to improve the stability of the biomarker discovery method. We therefore proposed a framework which combines function perturbation with data perturbation: **E**nsemble **F**eature **S**election **I**ntegrating **S**tability (EFSIS) which achieves both high prediction accuracy and stability. This work was presented in **Paper IV**.

List of Publications

Publications included in the thesis

(I) Yadetie, F., **Zhang, X.**, Hanna, E. M., Aranguren-Abadía, L., Eide, M., Blaser, N., Brun, M., Jonassen, I., Goksøyr, A., & Karlsen, O. A. (2018). RNA-Seq analysis of transcriptome responses in Atlantic cod (*Gadus morhua*) precision-cut liver slices exposed to benzo[a]pyrene and 17α-ethynylestradiol. *Aquatic Toxicology*, 201, 174-186.

(II) **Zhang, X.**, & Jonassen, I. (2020). RASflow: an RNA-Seq analysis workflow with Snakemake. *BMC Bioinformatics*, 21(1), 1-9.

(III) **Zhang, X.**, & Jonassen, I. (2019). A Comparative Analysis of Feature Selection Methods for Biomarker Discovery in Study of Toxicant-Treated Atlantic Cod (*Gadus morhua*) Liver. In *Symposium of the Norwegian AI Society, Communications in Computer and Information Science* (pp. 114-123). Springer, Cham.

(IV) **Zhang, X.**, & Jonassen, I. (2019). An Ensemble Feature Selection Framework Integrating Stability. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2792-2798). IEEE.

Other publications not included in the thesis

- Hanna, E. M.†, **Zhang, X.**†, Eide, M., Fallahi, S., Furmanek, T., Yadetie, F., Zielinski, D. C., Goksøyr, A., & Jonassen, I. (2020). ReCodLiver0.9: Overcoming challenges in genome-scale metabolic reconstruction of a non-model species. *BioRxiv*. doi: https://doi.org/10.1101/2020.06.23.162792

- Khan, E. A., **Zhang , X.**, Hanna, E. M., Bartosova , Z., Yadetie, F., Jonassen , I., Goksøyr, A., & Arukwe, A. (2020). Quantitative transcriptomics, and lipidomics in evaluating ovarian developmental effects in Atlantic cod (*Gadus morhua*) caged at a capped marine waste disposal site. *Environmental Research*, 109906.

- Dale, K., Yadetie, F., Müller, M. B., Pampanin, D. M., Gilabert, A., **Zhang, X.**, Tairova, Z., Haarr, A., Lille-Langøy, R., Lyche, J. L., Porte, C., Karlsen, O. A., & Goksøyr, A. (2020). Proteomics and lipidomics analyses reveal modulation of lipid metabolism by perfluoroalkyl substances in liver of Atlantic cod (*Gadus morhua*). *Aquatic Toxicology*, 105590.

- Khan, E. A., **Zhang , X.**, Hanna, E. M., Yadetie, F., Jonassen , I., Goksøyr, A., & Arukwe, A. Application of quantitative transcriptomics in evaluating the *ex vivo* effects of per- and polyfluoroalkyl substances on Atlantic cod (*Gadus morhua*) ovarian physiology. (Under review).

- Eide, M.†, **Zhang, X.**†, Karlsen, O. A., Goldstone, J. V., Stegeman, J., Jonassen, I., & Goksøyr, A. The chemical defensome of model fish species. (Manuscript in preparation).

- Eide, M., Goksøyr, A., Yadetie, F., Gilabert, A., Bartosova, Z., Frøysa, H., Fallahi, S., **Zhang, X.**, Blaser, N., Jonassen, I., Bruheim, P., Alendal, G., Brun, M., Porte, C., & Karlsen, O. A. A multi-level omics approach to study lipid metabolism regulation by PPARa and PPARb agonists in Atlantic cod (*Gadus morhua*). (Manuscript in preparation).

† These authors contributed equally to this work.

Contents

Chapter 1

Introduction

1.1 Genomics

Genome is the genetic information stored in the DNA (RNA for some virus) of an organism. And genomics is the subject which studies the whole genome of an organism and then makes use of its information. It differs from genetics which focuses on individual genes. Instead, genomics focuses on the whole genome and studies its structure, function, evolution, mapping and editing.

Genomics harnesses the availability of complete genome sequences. A genome sequence is a list of the nucleotides (A, C, G, and T for DNA genomes and U for RNA genomes). Genome sequencing reveals the order of the nucleotides present in the genome. The effort traced back to 1976 when Walter Fiers at the University of Ghent established the complete genome sequence of a viral RNA genome (bacteriophage MS2), and the next year Fred Sanger completed the first DNA genome sequence (bacteriophage ΦX174) [1]. Later on, distinguished from the earlier methods like Sanger sequencing, second-generation sequencing or next generation sequencing was developed and led to increasingly faster, low-cost, and high-throughput genome sequencing, and has been dominating the genome sequencing field since its development [2]. Recently, third generation sequencing was introduced which can produce longer reads than second-generation sequencing [3].

Officially launched in 1990, the Human Genome Project aimed to obtain a highly accurate sequence of the vast majority of the euchromatic portion of the human genome [4]. The drafts of the human genome were published by *Celera Genomics* [5] and the *International Human Genome Sequencing Consortium* [6] scientists in 2001, using whole-genome shotgun sequencing method and hierarchical shotgun sequencing method respectively. A more complete draft was published in 2003 [7]. In this process, the development of new technologies for large-scale, high-throughput generation of biological data at low cost ensured the completion of the project [7]. Since then, more and more sequencing technologies and machines which are capable of generating high-quality sequencing data have been developed to correct errors in the human genome sequence and to sequence the genomes of other species.

Lots of efforts have been devoted to sequencing the genome of Atlantic cod (*Gadus morhua*). The first Atlantic cod genome assembly (gadMor1) was published back in 2011, obtained by 454 sequencing [8] of shotgun and paired-end libraries and 22,154 genes were identified by automated annotation [9]. An improved genome assembly (gadMor2) was generated by combining data from Illumina, 454 and the longer PacBio sequencing technologies, as well as integrating the results of multiple assembly programs in 2017 [10]. The recently released gadMor3 assembly (GenBank assembly accession: GCA_902167405) was developed based on long-read sequencing technology. The genome assembly used in the dCod 1.0 project was also updated as the new release came out. gadMor3 was used in the late stage of dCod 1.0 project due to its better quality than the previous two versions [11]. However, despite the efforts and quality improvement of Atlantic cod genome assembly, as a non-model organism, Atlantic cod is still less annotated and less resources are available compared with the model organisms, such as human (*Homo sapiens*), mouse (*Mus musculus*), and zebrafish (*Danio rerio*). This has been a very challenging issue for the dCod project. The genome annotation leads us to the next subsection.

1.2 Functional genomics

As the mapping and sequencing (structural genomics) phase of the Human Genome Project came to an end, a new era of functional genomics focusing on the study of gene function came to shape [12]. Unlike structural genomics, functional genomics focuses more on dynamics of gene expression and regulation of it, involving genomics, transcriptomics, proteomics, metabolomics and their interactions [13].

1.2.1 Gene expression

Gene expression is the process by which the genetic information stored in DNA is used to direct the synthesis of functional gene products. Gene expression is summarized in the central dogma of molecular biology firstly presented by Francis Crick in 1970 [14]. Central dogma states that DNA, as the repository of genetic information, can replicate itself (DNA replication) and can also pass the genetic information to (messenger) RNA which occurs in the process called transcription. The messenger RNA (mRNA) then serves as a template to direct the synthesis of protein which is called translation. The central dogma was later expanded by adding RNA replication [15] and reverse transcription [16] (dotted arrows in Figure 1.1).

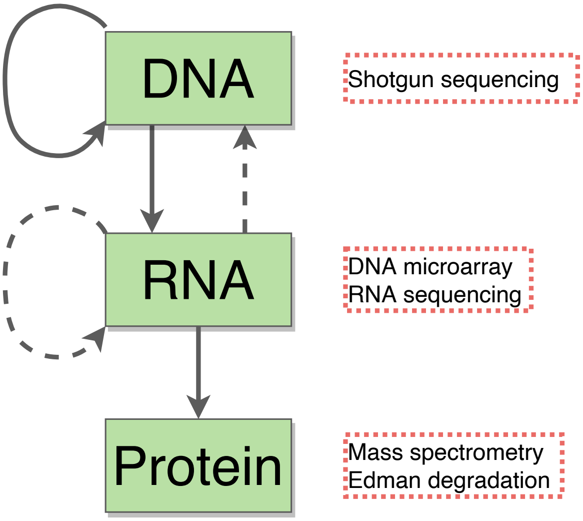


Figure 1.1: Central dogma of molecular biology with some important technologies in each level.

In Figure 1.1, next to each level, some important technologies are given. Shotgun sequencing method was widely used in the Human Genome Project as mentioned in the previous section. DNA microarray and RNA sequencing (RNA-Seq) are the most useful methods for whole transcriptome gene expression profiling [17], and RNA-Seq will be discussed in more details in the next subsection. The two major methods of protein sequencing are mass spectrometry and Edman degradation [18].

1.2.2 RNA sequencing analysis

RNA sequencing (RNA-Seq) has overcome many limitations of DNA microarray and has become an attractive alternative since it was introduced over a decade ago [19–28]. Lots of studies have been conducted using RNA-Seq and most of the generated datasets are shared in public repositories such as the Gene Expression Omnibus (GEO) [29] and ArrayExpress [30]. The underlying sequencing reads are typically archived on the Sequence Read Archive (SRA) [31]. The growth of deposited RNA-Seq samples on SRA is shown in Figure 1.2. There are currently more than 1.9 million samples (https://www.ncbi.nlm.nih.gov/sra/?term=RNA-Seq, accessed on 30 June 2020).

The RNA-Seq datasets obtained from the public repositories mentioned above or

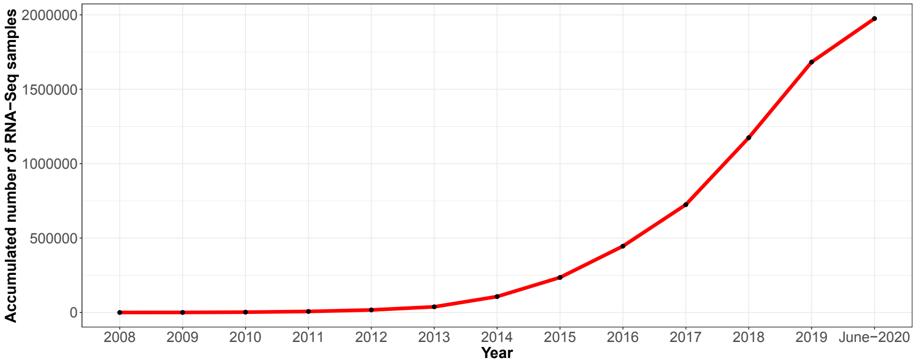


Figure 1.2: Growth of RNA-Seq sample entries on Sequence Read Archive (SRA) over years.

from a sequencing center are usually raw reads, which need to go through a series of processing steps. The raw sequence information is usually saved in FASTQ file format. There are four line types in the FASTQ format describing one read / sequence at a time. First comes a "@" title line with a record identifier. Second is the sequence itself and white space such as spaces or tabs is not allowed. Third comes the "+" line which is a signal of the end of the sequence and optionally with a full repeat of the title line. Finally is the quality line which must be equal in length to the sequence string [32].

There is no optimal workflow for various different applications and analysis scenarios in which RNA-Seq can be used. We focus on the standard and typical RNA-Seq analysis workflow including quality control of raw reads, trimming, quantification of transcripts or genes, differential expression analysis, and visualization.

An overview of the steps and some popular tools for each step are given in Figure 1.3 (adapted from [33]). Some detailed introduction will be given in the following subsections.

Quality control of raw reads and trimming

Quality control of the raw reads includes the analysis of sequence quality, GC content, the presence of adaptors, and so on [34]. FastQC is a popular open-source software for quality control and will generate a quality control report including investigation on: (1) per base sequence quality, (2) per sequence quality scores, (3) per base sequence content, (4) per base GC content, (5) per base N content, (6) sequence length distribution, (7) duplication level, (8) overrepresented sequences, (9) adapter content, (10) kmer content, and (11) per tile sequence quality (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Generally speaking, read quality decreases towards the 3’ end, and if the quality of some bases become

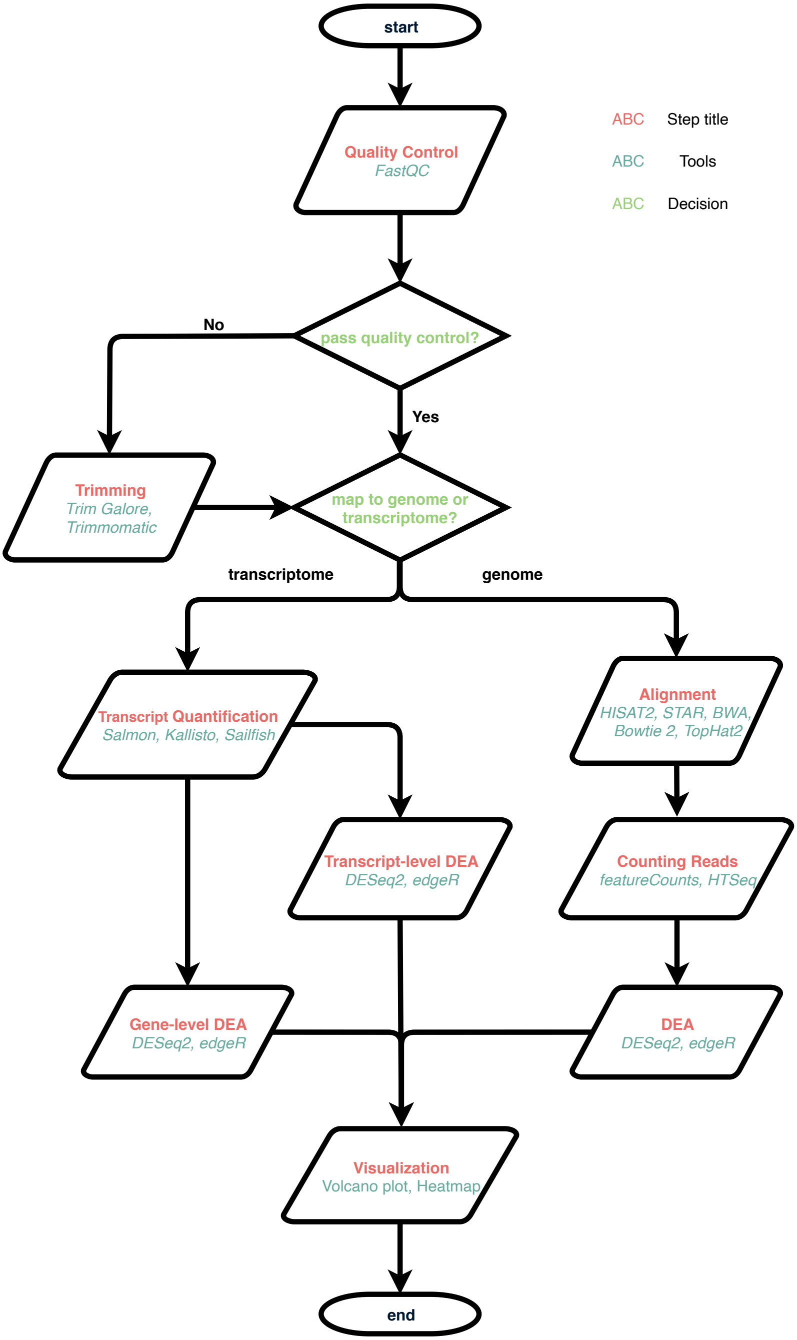


Figure 1.3: Overview of the steps performed in a typical RNA-Seq analysis workflow and some popular tools used in each step. DEA: Differential Expression Analysis. Adapted from Fig. 1 of ***Paper II****.*

too low, they should be removed [34]. Tools such as Trimmomatic [35] and Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) can be used to discard low-quality reads, trim adaptor sequences and eliminate poor-quality bases.

Quantification of transcripts or genes

The high-quality reads can then be mapped to either a transcriptome or a genome. The traditional way is to map the reads to the genome and this process is usually called alignment. Many tools such as HISAT2 [36], STAR [37], BWA [38], Bowtie 2 [39], and TopHat2 [40] are popular aligners. Alignment is followed by counting reads associated with genes which can be done by featureCounts [41] or HTSeq [42]. They take both the output from alignment (BAM files) and General Feature Format (GFF) / General Transfer Format (GTF) file as input.

Recently, several transcriptome-based novel tools introduced alignment-free transcript quantification utilizing k-mer-based counting algorithms becoming more and more popular such as Salmon [43], Kallisto [44], and Sailfish [45]. Mapping to a transcriptome is generally faster than to a genome but it does not allow *de novo* transcript discovery [34].

Differential expression analysis

The quantification of transcripts or genes is followed by Differential Expression Analysis (DEA) where the purpose is to identify genes that are expressed at different levels between two classes of samples (e.g. healthy, disease) [46].

The raw read counts can not be used directly for statistical tests, because of some systematic variations including between-sample differences such as library size or sequencing depth [25], within-sample differences such as gene / transcript length [47], and technical effects such as library preparations [48]. Normalization is therefore necessary to remove those unwanted variations. Many different ways of normalization have been proposed so far [49]. Normalization by library size and gene (transcript) length includes Reads Per Kilobase per Million (RPKM, for single-end sequencing) [25], Fragments Per Kilobase per Million (FPKM, for pair-end sequencing) [50], Transcripts Per kilobase Million (TPM) [51], and so on. Normalization by distribution includes Trimmed Mean of the M-values (TMM) [52] which is used in edgeR [53], median-of-ratios method which is used in DESeq [54], and so on.

With the raw reads normalized, statistical tests can be done afterwards to find out the differentially expressed genes.

We give a typical experimental design example here: there are fish exposed to seawater and chemical solution and these two groups are considered as control and treatment groups respectively. We have measured the expression values of a particular gene of all samples and we would like to know whether this gene expresses differently between these two groups.

The Fold Change measuring the difference of the gene expression values between control and treatment groups can be calculated as Equation 1.1.

$$FC = \frac{\overline{X}_2}{\overline{X}_1} \tag{1.1}$$

where $\overline{X}_1$ is the mean expression values of control samples, and $\overline{X}_2$ is the mean expression values of treated samples.

Usually a Log2 Fold Change is used so that a positive value indicates that the gene is up-regulated and a negative value indicates that the gene is down-regulated.

The Fold Change can measure the magnitude of difference between groups, but it ignores the variance within each group, so it fails to find out the genes of high reproducibility with comparably low differentiality [55]. A statistical test is usually applied to measure the significance of how the gene is differentially expressed between these two groups. A basic Student's t-test calculates a t-score which is the ratio of difference between groups' mean values and the variability within groups. The t-score can be calculated as Equation 1.2.

$$t\text{-}score = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{1.2}$$

where s_1 and s_2 are the standard deviations of two groups, and n_1 and n_2 are the sample sizes of two groups. Referring to a T-Distribution table, a corresponding P-value can be got. The P-value is the probability of obtaining an experimental result at least as extreme as observed under the null hypothesis, that there is no difference in expression between the experimental conditions.

Another typical experimental design for biomarker discovery is to design the experiment in a pair-wise way. In medical study, it can be instances of the same patient being tested repeatedly - before and after receiving a particular treatment; in toxicology study, it can be tissue slices of the same fish being exposed to seawater and chemical solution. An example is given in Table 1.1. There are 6 samples which are liver slices cut from 3 fish: A, B and C. S1 and S4 are from fish A, S2 and S5 are from fish B, and S3 and S6 are from fish C. S1, S2, and S3 are exposed to seawater (control group), and S4, S5, and S6 are exposed to chemical solution (treatment group). The purpose

Sample (liver slice)	Group	Fish
S1	Control	A
S2	Control	B
S3	Control	C
S4	Treat	A
S5	Treat	B
S6	Treat	C

Table 1.1: An example of experimental design in a pair-wise way.

is to find out the differentially expressed genes comparing treatment group and control group. In this case, a paired t-test is preferred. And the corresponding t-score is calculated by Equation 1.3.

$$t\text{-}score_{pair} = \frac{\overline{X}_D}{S_D/\sqrt{n}} \tag{1.3}$$

where $\overline{X}_D$ and S_D are the mean and standard deviation of the differences of all pairs and n is the number of pairs.

We have only talked about one gene or one comparison test above, but in the gene expression data, there can easily be over ten thousand genes. The P-value needs to be adjusted to account for the multiple testing issue. The simplest way to adjust the P-values is to use the conservative Bonferroni correction method which multiplies the raw P-values by the number of tests [56].

A statistical test usually requires specific distributional assumptions, for example, the basic Student's t-test requires a normal distribution [57], Fisher's exact test and likelihood ratio test (applied by R package DEGseq) require a Poisson distribution [58], Generalized Linear Model methods (GLMs) (applied by R packages DESeq, DESeq2, and edgeR) require a negative binomial distribution [53; 54; 59; 60].

Due to the application of different normalization methods and statistical tests, and some other more detailed aspects, different DEA tools can give very different results for the same dataset [61]. Merely based on the citation, DESeq2 [59] and edgeR [53] are the most popular tools for DEA.

Visualization

The results of DEA can be visualized in several ways and two popular ones are presented here: Volcano plot and Heatmap.

Volcano plot is a scatter-plot that summarizes both statistical significance and the magnitude of the change. It plots the negative log of P-value (usually base 10) on the y axis, so the genes (or transcripts) towards the top are the ones showing statistical significance. On the x axis is the log of fold change (usually base 2), so the points with large magnitude of change are either to the left (down-regulated) or to the right (up-regulated). The interesting genes are therefore at the top-left and top-right corners as shown in Figure 1.4.

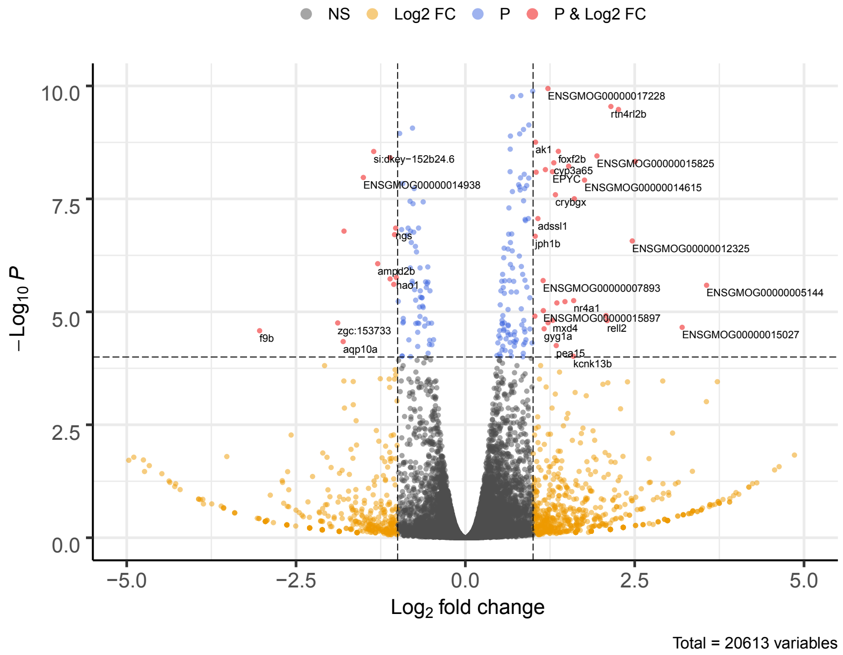


Figure 1.4: Fig. 3a of ***Paper II****. Volcano plot of genes presented by* Log_2 *Fold Change and* $-\mathrm{Log}_{10}$ *P-value.*

Heatmap is useful for visualizing the expression of genes across samples from different conditions, and specifically, the cluster heatmap can also indicate how well the samples from the same condition are grouped together by the expression pattern of the genes selected (usually top differentially expressed genes from the results of DEA). Figure 1.5 shows that the genes in this cluster Heatmap express very differently in the samples from those two groups (a control group and another group exposed to low-dose oil).

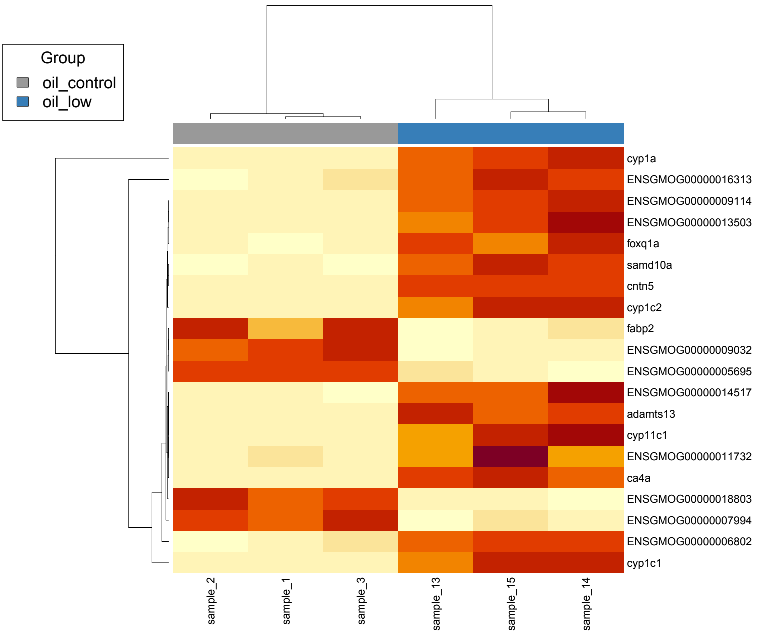


Figure 1.5: Fig. 3b of ***Paper II****. Heatmap of samples from two conditions using the top differentially expressed genes.*

1.3 Toxicogenomics

Toxicology is a multidisciplinary subject that studies the harmful interactions between chemicals and biological systems [62]. Toxicogenomics combines toxicology with genomics or other high-throughput molecular profiling technologies to study how genomes respond to toxicant exposure [63].

Toxicogenomics aims to understand and predict toxicity in order to understand how organisms respond to toxicant exposure or compound treatment using omics data, especially gene expression data due to its rapidly increasing amount in recent years [64] thanks to the new sequencing technologies. It utilizes the comprehensive gene expression data to identify gene expression signatures that highly relate to genetic toxicity [65], which are also referred as biomarkers. A biomarker is an indicator of a biological state in response to intervention, which is toxic exposure in this case, and can also be a disease (such as cancer), or drug response in precision medicine. The use of these approaches has a long history and rapidly developed in the past decade due to development in gene expression technologies such as DNA microarray, RNA-Seq.

In a toxicant exposure study, the differentially expressed genes can be regarded as the biomarkers of the particular toxicant. A further study of those genes can shed some light on how the organism reacts to the interruption of that toxicant.

To be noted, the significance of a differentially expressed gene is usually defined by the P-value or Q-value calculated from statistical tests [66] which is univariate because each gene is treated independently, ignoring the reality that genes interact with each other. Hence, some multivariate methods should also be introduced in the study of biomarker discovery.

1.4 Machine learning

Machine learning applies mathematical approaches to train the machine to learn from data for some particular tasks. It is often divided into supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning is fed with input and the corresponding labeled output. The task is to find the best function that can map input with output. When the output is categorical, it is a problem of classification; and when the output is real-valued, it is then a problem of regression [67]. In the real world, the outputs are not necessarily available for all inputs, for example when labeling the samples is very expensive, so only a small part of the input is labeled. Semi-supervised technique is then required in this situation to make the best out of the available data [68]. When the samples are not labeled at all, unsupervised learning is applied to find out the inherent pattern of the data [69]. Reinforcement learning continuously takes into new observations and adjusts the current model to maximize the reward or minimize the risk [70].

In this thesis, we focus on supervised classification where a model (or classifier) of distribution of class labels in terms of predictor features is built. The resulting classifier is then used to assign class labels to unknown instances [71]. For example, a tumor classifier can be trained from the gene expression profiles of some patients diagnosed with benign and malignant tumor, where the genes are features and "benign" or "malignant" tumor is class label. When the gene expression profile of an unknown patient is provided, the trained classifier can be used to predict whether the patient is carrying the malignant tumor.

The trained classifier is usually evaluated before it is applied to real cases based on prediction accuracy (some common performance metrics will be introduced later). There are at least three techniques to calculate a classifier's accuracy. One technique is to split the samples into a training set for model training and a testing set for performance evaluation. Another technique called cross-validation is to divide the samples into mutually exclusive and equal-sized subsets and for each subset as testing set the classifier is trained on the union of all the other subsets. The average prediction accuracy of each subset is therefore the overall estimate of the prediction accuracy of the classifier. Leave-one-out cross-validation is a special case of cross-validation, where all testing sets consist of only one sample. This type of testing scheme is of course more computationally expensive, but useful when the number of samples is quite limited or the most accurate estimate of a classifier's prediction accuracy is required [71].

1.5 Feature selection

In the real world, the collected data for training the classification model usually comes with lots of noise. The reasons causing the noise are many and the major two reasons are the imperfection of the technologies collecting the data and the data source itself [72]. For example, DNA microarray experiments suffer from noise from sample preparation steps and the subsequent readout processes [73]. The RNA-Seq technology overcomes those problems and is capable of detecting genes with low expression [74], but those low-expression genes are again a problem in differential expression analysis [75].

Feature selection can be applied to remove the noise before training the classifiers. By doing that, feature selection can improve the prediction performance of the classifier, contribute to faster and more cost-effective prediction, and provide a better understanding of the input data [76].

There are many different algorithms for feature selection. Take one for example called minimal-redundancy-maximal-relevance criterion (mRMR) [77]. Information theory is used in this method. Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x,y)$:

$$I(x;y) = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{1.4}$$

Apparently, mRMR includes two parts: maximal relevance and minimal redundancy. Maximal relevance is to search a feature subset F with m features $(f_1, f_2, \dots, f_m)$ satisfying Equation 1.5 which maximizes the mutual information between the features in subset F and class label c.

$$\max D(F,c),\ D = \frac{1}{m} \sum\nolimits_{f_i \in F} I(f_i;c) \tag{1.5}$$

where $i \in \{1,2,...,m\}$.

It can be ensured that the features selected above are highly relevant to the class label, but the information they store may be redundant. Therefore, the class-discriminative power would not change much if one of them was removed. So minimal redundancy is added to select mutually exclusive features:

$$\min R(F),\ R = \frac{1}{m^2} \sum\nolimits_{f_i, f_j \in F} I(f_i; f_j) \tag{1.6}$$

where $i, j \in \{1, 2, ..., m\}$.

The criterion combing these two constraints and satisfying Equation 1.7 is then called "minimal-redundancy-maximal-relevance" (mRMR).

$$\max \Phi(D, R),\ \Phi = D - R \tag{1.7}$$

Since feature selection is capable of selecting the features highly relevant to the class labels, it can be applied to a toxicogenomics study to select the biomarkers indicating the toxicity. In our toxicant exposure study, the cod fish from both control and treated groups are the input samples, and the gene expression profiles are the features. But quite often, the feature dimension is very high but the sample size is very small in omics data [78; 79]. For example, in one of our exposure experiments, there are 6 to 8 samples from each group but there are 19,000 genes left even after filtering out the low-expression ones [80]. Too many features can make the computation complexity very high, and can also lead to overfitting, meaning that the trained model performs well on the known samples, but very badly on unseen new samples. So feature selection is often applied before classification to get rid of the unimportant features.

To be noted, feature selection is different from feature extraction even though both of them can reduce the dimensionality of the feature list. Feature selection picks out the important features that are a subset of the original feature set, but feature extraction generates new features. Two popular examples are unsupervised technique Principle Component Analysis (PCA) [81] and supervised technique Partial Least Square Regression (PLS-R) [82]. PCA is applied without considering the correlation between the dependent and the independent variables (or to say features and class labels), while PLS-R is applied based on the correlation. But both of them generate new features which are linear combinations of the original features which can not be directly used as biomarkers.

Feature selection can mainly be divided into three families: filter methods which focus on the correlation between variables and targets; wrapper methods which use an objective function (such as classification accuracy) to evaluate the importance of features; embedded methods where the feature selection procedure is embedded with classification procedure and the features are selected automatically during the classification process [83]. Filter methods are independent of classification procedure, but not the other two families. Biomarkers should be independent of the classification algorithm, so we only focus on filter methods in this thesis.

1.6 Evaluation of feature selection

Many feature selection methods have been proposed by researchers so far, and more are being and will be proposed. Quite often the feature selection methods are proposed for a specific research question or some type of datasets. Then how to select the best feature selection method for a given context?

1.6.1 Prediction accuracy

Feature selection is very often used upstream of classification problems. The purpose of applying feature selection prior to classification is to improve the prediction accuracy. So a good feature selection method should be able to improve the prediction accuracy compared to using the whole original feature set.

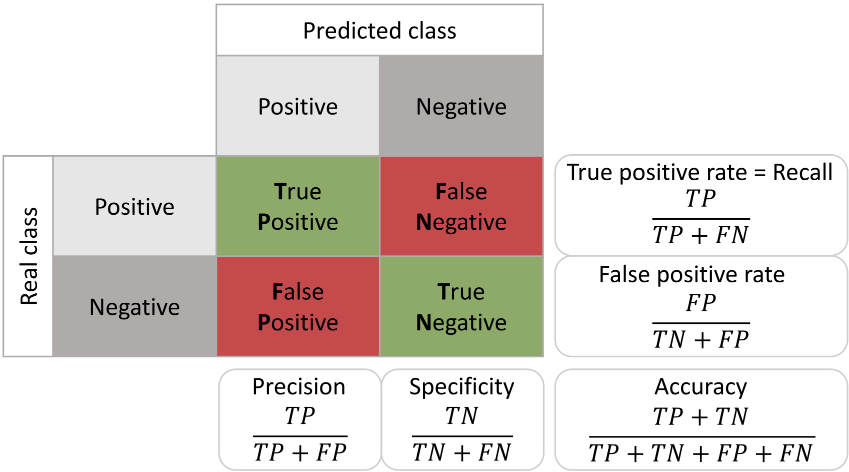
A Receiver Operating Characteristics (ROC) graph is a technique for visualizing the performance of classifiers, and the area under an ROC curve (AUC) is usually used as a scalar to evaluate a classifier's performance [84]. Some relevant concepts will be introduced in this subsection.

Figure 1.6 shows the confusion matrix and some common performance metrics that can be calculated from it. The two green cells along the diagonal, True Positive (TP) and True Negative (TN), represent the correct predictions. The two red cells, False Positive (FP) and False Negative (FN), represent the errors. Next to the confusion matrix, some common performance metrics are calculated. Among them, the two metrics related to ROC graph are True positive rate (also called Recall, Equation 1.8) and False positive rate (Equation 1.9).

$$True\ positive\ rate = \frac{Positives\ correctly\ classified}{Total\ positives} \tag{1.8}$$

$$False\ positive\ rate = \frac{Negatives\ incorrectly\ classified}{Total\ negatives} \tag{1.9}$$

ROC graph is a two-dimensional graph in which False positive rate is plotted on the X axis and True positive rate is plotted on the Y axis as Figure 1.7 shows. For



		Predicted class		
		Positive	Negative	
Real class	Positive	**T**rue **P**ositive	**F**alse **N**egative	True positive rate = Recall $\frac{TP}{TP+FN}$
	Negative	**F**alse **P**ositive	**T**rue **N**egative	False positive rate $\frac{FP}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	Specificity $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

Figure 1.6: Confusion matrix and some performance metrics calculated from it.

a classifier, when its True positive rate against its False positive rate is plotted in the ROC graph along a curve as shown in Figure 1.7, the area under it (the grey shadow in the figure) is called Area Under the Curve (AUC). The AUC value indicates a tradeoff between benefits (true positive) and costs (false positive). The dotted line shows the worst case where AUC is 0.5 and it means that the model has no discrimination capacity to distinguish between positive class and negative class.

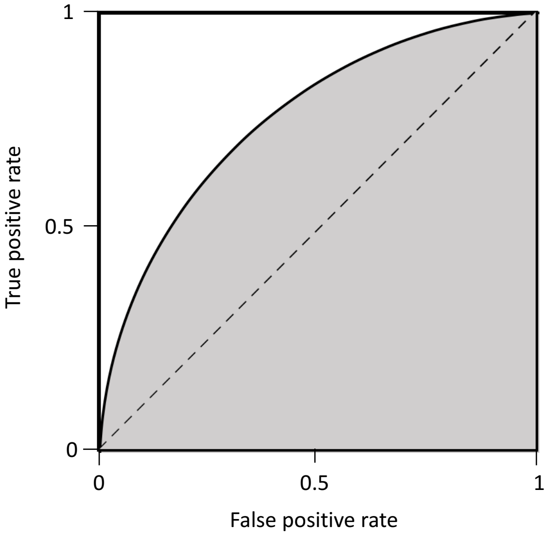


Figure 1.7: A Receiver Operating Characteristics (ROC) graph and the Area Under the Curve (AUC).

1.6.2 Stability

Stability shows the ability of a feature selection method to give a consistent set of features when the training data changes [85].

One application of feature selection in medical science is biomarker discovery. The stability of biomarker discovery is a challenging task since the sources of instability in biomarker discovery are many:

- Algorithm design without considering stability
  - Only aiming to find a feature subset to construct a classifier of the best prediction accuracy
- Existence of multiple sets of true biomarkers
  - Highly correlated features, different ones may be selected under different settings
  - No redundant features, but existence of multiple non-correlated sets of real biomarkers is also possible
- Small sample size vs. High dimensional features
  - In analysis of gene expression data and proteomics data, there are typically hundreds or even less than one hundred samples but thousands of features

Figure 1.8 shows an example of stability of feature selection methods. In the study of the same disease, the gene expression profiles of some patients from two hospitals are collected which can be used as training data for biomarker discovery. Two feature selection methods are applied to these two datasets. Method 1 gives two identical gene lists but method 2 gives two different gene lists. Theoretically the biomarkers for a disease should be independent of the training data. In this example, we say that feature selection method 1 is more stable than method 2.

Then we need a metric to evaluate the stability of a feature selection method. Let $\mathcal{A}$ be the training set and the samples are presented by M features $(f_1, f_2, \dots, f_M)$. N resampling steps generate N training subsets and after applying the feature selection method to them, N selected feature subsets are obtained: $(F_1, F_2, \dots, F_N)$.

$$\mathcal{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M,1} & x_{M,2} & \cdots & x_{M,N} \end{pmatrix} \tag{1.10}$$

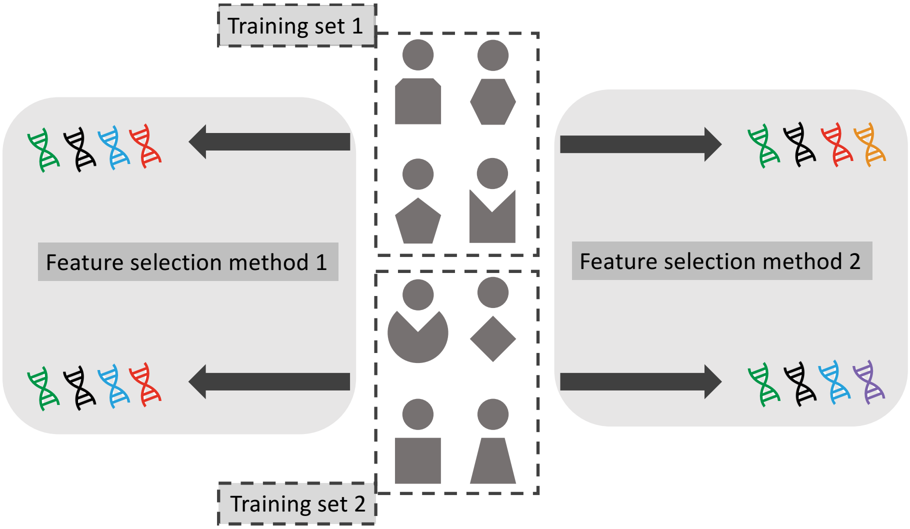


Figure 1.8: Illustration of stability of feature selection. With two training sets, method 1 always selects the same gene list as biomarkers, but method 2 gives two different lists.

where the element $x_{m,n}$ ($m \in \{1,2,\ldots,M\}, n \in \{1,2,\ldots,N\}$) in matrix $\mathcal{X} : \{0,1\}^{M\times N}$ indicates whether feature f_m is included in selected feature subset F_n. So we have:

$$\mathcal{F} = (F_1, F_2, \ldots, F_N) = (f_1, f_2, \ldots, f_M) \times \mathcal{X} \tag{1.11}$$

Let $k_n = |F_n|$ be the cardinality of feature subset F_n. For two feature subsets F_i and F_j ($\{i,j\} \in \{1,2,\ldots,N\}$), $r_{i,j} = |F_i \cap F_j|$ is the cardinality of the intersection.

Kalousis et al. [86] introduced the similarity index of two feature subsets, F_i and F_j, as:

$$S_{Kalousis}(F_i, F_j) = \frac{|F_i \cap F_j|}{|F_i \cup F_j|} = \frac{r_{i,j}}{k_i + k_j - r_{i,j}} \tag{1.12}$$

Kuncheva [87] pointed out that given a fixed k for two feature subsets ($k_i = k_j = k$), the expected value of r is $E(r) = \frac{k^2}{M}$ and $max(r) = k$. In this case, the above similarity index becomes:

$$S'_{Kalousis}(F_i, F_j) = \frac{\frac{k^2}{M}}{2k - \frac{k^2}{M}} = \frac{k}{2M - k} \tag{1.13}$$

which has a tendency to increase with increasing k. So they proposed to use the expected value as a modified index:

$$S_{Kuncheva}(F_i, F_j) = \frac{r_{i,j} - E(r_{i,j})}{max(r_{i,j}) - E(r_{i,j})} = \frac{r_{i,j} - \frac{k^2}{M}}{k - \frac{k^2}{M}} \tag{1.14}$$

Both of the metrics mentioned above are used to measure the similarity between two feature subsets. If there is a sequence of feature subsets, $(F_1, F_2, \ldots, F_N)$, the stability index can be calculated by averaging all pairwise similarity indices [87]:

$$S'_{Kuncheva}(\mathcal{F}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_{Kuncheva}(F_i, F_j) \tag{1.15}$$

Davis et al. [88] proposed a more straightforward and flexible metric to measure the stability regarding multiple feature subsets and various number of features in them (various k in those feature subsets):

$$S_{Davis}(\mathcal{F}) = \frac{\sum_{f \in F} (\omega(f)/N)}{|F|} \tag{1.16}$$

where F is the set of features that appear in at least one of the N subsets ($F = F_1 \cup F_2 \cup \ldots F_N$); $|F|$ indicates the cardinality of F; $\omega(f)$ is the frequency of feature $f \in F$ that appears in those N subsets.

1.7 Ensemble feature selection

Ensemble feature selection includes two strategies: function perturbation and data perturbation. To be noted, many feature selection methods calculate a score for each feature which indicates its importance. The features then can be ranked based on that score. So this type of feature selection methods are also called rankers.

1.7.1 Function perturbation

With so many different feature selection methods which are often data-dependent, picking out the best one for a specific research topic or data can be very challenging. With the idea of ensemble learning, we can combine the outputs from multiple feature selection methods and this strategy is called function perturbation [89–92]. Figure 1.9 illustrates function perturbation. Many different rankers are applied to the same training data and their ranking results of the features are then combined into one ranked list which is used as the final result of function perturbation.

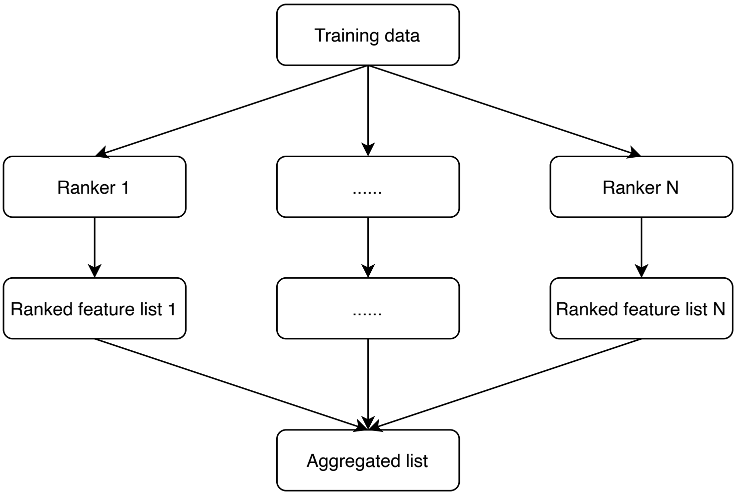


Figure 1.9: Function perturbation.

1.7.2 Data perturbation

As mentioned above, when the sample size is small and feature dimensionality is high, the stability of an algorithm is subject to being low. Data perturbation was introduced which can increase the stability [88; 92–95]. Figure 1.10 illustrates data perturbation. Only one ranker is used in data perturbation, but many resampled training datasets are generated from the original training data and the results from all those resampled datasets are combined into the final result.

1.7.3 Aggregation strategy

Similar to the notations introduced in section 1.6.2, assume that the samples in the training set $\mathcal{A}$ are presented by M features $(f_1, f_2, \ldots, f_M)$. Either in the framework of function perturbation or data perturbation, there are N ranked feature lists, $\mathcal{L} = (L_1, L2, \ldots, L_N)$ to be aggregated.

$$\mathcal{R} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,N} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M,1} & r_{M,2} & \cdots & r_{M,N} \end{pmatrix} \tag{1.17}$$

where $r_{m,n}$ $(m \in \{1,2,\ldots,M\}, n \in \{1,2,\ldots,N\})$ in matrix $\mathcal{R} : \{1,2,\ldots,M\}^{M\times N}$ indicates the ranking position of f_m in ranking list L_n.

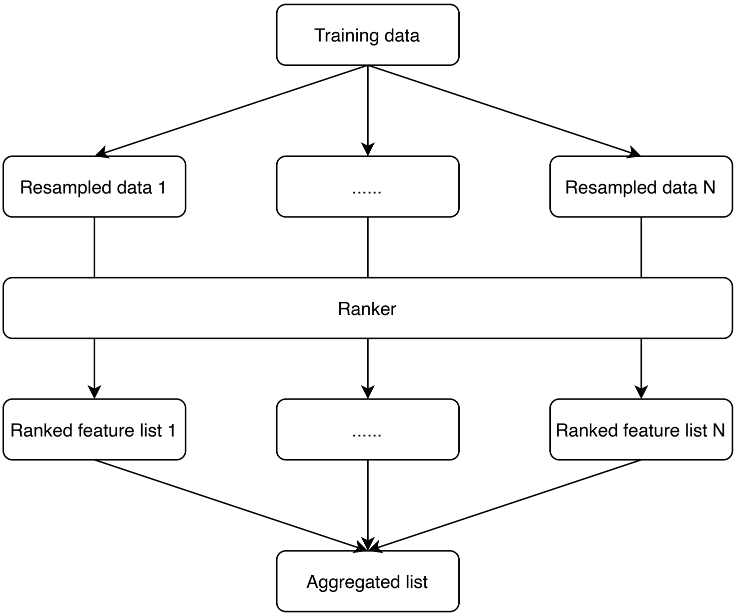


Figure 1.10: Data perturbation.

To aggregate the ranked feature lists, Chiew et al. proposed to use the intersection and union operations to aggregate the lists [96]. With a pre-defined k ($1 < k < M$, the number of selected features), we have:

$$x_{m,n} = \begin{cases} 0 & \text{if } r_{m,n} > k \\ 1 & \text{if } r_{m,n} \leq k \end{cases} \tag{1.18}$$

In this case, $\mathcal{L}$ is transformed into $\mathcal{F}$. The aggregated feature lists using the two strategies proposed by Chiew et al. [96] can be illustrated as:

$$F_{intersection} = F_1 \cap F_2 \cap \ldots F_N \tag{1.19}$$

$$F_{union} = F_1 \cup F2 \cup \ldots F_N \tag{1.20}$$

But if there are very few common features shared by those N feature subsets, $|F_{intersection}|$ will be very small and $|F_{union}|$ will be very large. Another shortcoming of these two strategies is that the original ranking information of f_m in L_n is not used while aggregating the feature lists in $\mathcal{F}$.

Breitling et al. proposed a more robust and informative strategy: Rank Products (RP), using the ranking product to score each feature which avoids the problems mentioned above [97]. RP aggregates the ranked feature lists in $\mathcal{L}$ into one ranked list L in which the ranking position of a feature f_m is calculated as:

$$r_m = (\prod_{n=1}^{N} r_{m,n})^{1/N} \tag{1.21}$$

With a given k, only the top k features in the aggregated ranked feature list L are kept as selected features, meaning that f_m is kept only if $r_m \leq k$.

Chapter 2

Aim of the study

As the fast advances of high-throughput transcriptomics technologies, increasing amounts of RNA sequencing (RNA-Seq) data are being generated, which provides enormous resources for biomarker study. The dCod 1.0 project has been using RNA-Seq data to identify the biomarker genes for some pollutants of particular concern, using Atlantic cod (*Gadus morhua*) as the study species which is of great importance to the North Atlantic fisheries.

The processing of RNA-Seq data involves many steps. To automate that work, many workflows have been developed. But most of them are designed for model species and do not support paired test in the differential expression analysis step, and some of them are hard to be scaled up for large dataset. So the first aim of this study was to develop a maximally generalized RNA-Seq analysis workflow which can solve all the issues mentioned above and can be applied to a wide range of applications.

In the differential expression analysis step, statistical tests are utilized to find out the genes that express differently in different groups. Those genes can be considered as biomarker gene candidates. But similar to other univariate methods, statistical tests have the issue that they treat each gene independently, but genes actually have complex interactions. The multivariate methods take those interactions into consideration. So the second aim was to study the difference of those methods for biomarker discovery.

Our study showed that a method performs quite differently on different datasets along the two evaluating dimensions: stability and prediction accuracy. The strategy of *function perturbation* has been proposed to combine different biomarker discovery methods, or feature selection methods. But the stability of *function perturbation* is still in question. It has been claimed that *data perturbation* can improve a method's stability. Our final aim was to design a framework combining these two strategies that possesses the advantages of both, so that the aggregated biomarker discovery method is not data-dependent any more and also has a high stability and prediction accuracy.

Chapter 3

Results

3.1 RNA-Seq analysis and differential expression analysis

Prior to analysis of gene expression data, the RNA sequencing (RNA-Seq) data needs to be processed through a series of procedures resulting in quantification of transcript abundance and gene expression.

In the dCod 1.0 project, lots of RNA-Seq data were generated. The need of a workflow which can automate the whole process was crucial. After some review of the published workflows and several failed attempts to use some of them, we decided to develop our own workflow which can satisfy all our requirements. The workflow was published on GitHub (https://github.com/zhxiaokang/RNA-Seq-Snakemake) and applied in the attached **Paper I**.

Paper I studied the toxicogenomics of Atlantic cod (*Gadus morhua*) by mapping its transcriptomics changes in response to exposure of benzo[a]pyrene (BaP) and 17α-ethynylestradiol (EE2) which are model compounds in environmental toxicology. In the experiments, slices from the same liver sample were assigned to each of the exposure groups in a paired-sample design. In the study of differential expression analysis of these exposure experiments, we found that paired test [98] which makes use of the information that some of the slices in different groups are from the same fish, can highly improve the statistical test strength, resulting in more significantly differentially expressed genes. But very few current RNA-Seq analysis workflows provide this option.

That experience made us aware that a generalized workflow which could be applied to various contexts was still needed. We therefore developed a maximally general RNA-Seq analysis workflow, RASflow (an RNA-Seq Analysis Snakemake Workflow)

(**Paper II**, source codes available at https://github.com/zhxiaokang/RASflow). An overview of the steps performed in RASflow can be found in Figure 1.3.

Before RASflow, some other workflows making the similar efforts had already been published. We reviewed seven workflows published between 2017 and 2019 [99–105]. Compared with them, some characteristic features of RASflow are listed as follows:

- RASflow provides quality control of the sequencing data both before and after trimming. It also provides quality control of alignment.
- RASflow can be applied to any organism, no matter if it is a model or non-model organism.
- Both genome and transcriptome can be used as mapping reference.
- Differential expression analysis can be done on both transcript- and gene-level, and options of both single- and paired- test are provided.
- It has relatively modest memory requirements (4.3GB for the human genome).
- Using Conda [106], RASflow is very easy to install and has no version conflicts problems.
- Using Snakemake [107], RASflow is highly modular, so that replacing tools used in the workflow can easily be done.
- Very little programming skills are required from the users.
- RASflow can be run on all the mainstream operating systems: Linux, macOS, Windows.

Differential expression analysis can be done on both transcript- and gene-level if transcriptome is used as mapping reference. Two most popular tools for RNA-Seq differential expression analysis, edgeR [53; 60] and DESeq2 [59], are provided in RASflow. RASflow was evaluated on a benchmarking dataset (SRA accession: SRP082682) [61] for which we assume that the biomarkers (significantly differentially expressed genes) are already known. The results show that edgeR has a higher precision and DESeq2 has a higher recall, meaning that edgeR is more conservative in reporting a gene as differentially expressed.

Besides the benchmarking dataset mentioned above, RASflow was also tested on three other real datasets of three organisms: human (ArrayExpress accession: E-MTAB-567) [108], mouse (GEO accession: GSE141199), and Atlantic cod (GEO accession: GSE106968) [80]. The sequencing data were mapped to both the genome and the transcriptome and the job was run on both a High Performance Computing (HPC) machine (1TB RAM 60 cores Dell PowerEdge R910) and an ordinary desktop computer (8GB RAM 4 cores Intel Core 2). The runtime of the alignment step which is the most time-consuming part of the workflow was recorded. As expected, mapping to a genome takes much longer than to a transcriptome, especially when the raw data is large or the job is run on an ordinary computer.

3.2 Comparative study of feature selection methods for biomarker discovery

Differential expression analysis corresponds to performing univariate statistical tests. An adjusted P-value is calculated for each gene or transcript, then the genes or transcripts with the top lowest adjusted P-values are picked out as biomarkers. The problem for univariate methods is that they treat features as independent which is not necessarily true especially in the context of genomics study, since genes usually interact with each other. Multivariate methods which take the interactions between variables into consideration do not have that problem and are therefore also popular for biomarker discovery. To study whether one performs significantly better than the other, we conducted a comparative study of various methods from these two categories and evaluated those methods on two aspects: stability and prediction accuracy.

Significance Analysis of Microarrays (SAM) was picked as the representative of univariate methods. SAM was originally designed for detecting differentially expressed genes in DNA microarray data but was later widely used in differential expression analysis and biomarker discovery [61; 109–111]. For multivariate methods, minimum Redundancy Maximum Relevance (mRMR) and Characteristic Direction (GeoDE) were used. mRMR is based on information theory and it finds out the features that are least redundant and most relevant with the class labels [77]. Inspired by graphical perspective, GeoDE defines a separating hyperplane using linear discriminant analysis and uses the orientation of the hyperplane to identify the differentially expressed genes [112].

The stability is calculated using Equation 1.16. To evaluate the prediction accuracy, several classification algorithms were applied and they are: Support Vector Machine (SVM) [113], Random Forest (RF) [114], and extended two-class logistic regression (RIDGE and LASSO) [115].

We evaluated the methods on the datasets from two experiments of toxicant-treated Atlantic cod liver. In each experiment, the samples were exposed to different toxicant doses. We found that the methods perform quite differently (both on stability and prediction accuracy) in the datasets from these two experiments and different doses.

No method always outperforms the others across all circumstances. GeoDE performs better in stability than SAM and mRMR. Regarding the ability to improve a classifier' s prediction accuracy, mRMR performs the best in high-dose condition, but in low-dose condition, GeoDE outperforms the other two methods. So the performance of a biomarker discovery method quite depends on the dataset that it is applied to. The work was presented in **Paper III**.

3.3 An ensemble feature selection framework integrating stability

As we have discussed in the previous subsection, biomarker discovery methods (feature selection methods) perform very differently on different datasets, so choosing the most appropriate method for their datasets becomes a challenging problem for researchers. One solution is to apply function perturbation, which is a strategy to combine the outputs from several methods. And it has been approved to maintain or improve the prediction accuracy. But we found that function perturbation can hardly achieve satisfactory stability. With a similar logic as function perturbation, data perturbation applies the feature selection method on several datasets that are generated from the original dataset and then combines the outputs from them. Data perturbation is capable of improving the stability of that feature selection method. Considering the characteristics of these two strategies, we proposed an Ensemble Feature Selection Integrating Stability (EFSIS) framework combining both strategies and using stability of each individual method as their weight to make use of the advantages of both.

We included four varied feature selection methods in EFSIS but of course it is not limited to four. To demonstrate the generality of EFSIS, the four methods are based on very different sets of assumptions. They include SAM and GeoDE which were mentioned in the previous subsection, together with Information Gain which applies the entropy concept of information theory to evaluate the features [116], and ReliefF [117] which is a more robust version of Relief algorithm [118; 119] which evaluates a feature by how well it distinguishes the samples that are near to each other.

EFSIS, together with those individual methods and basic function perturbation of them, were evaluated by stability and prediction accuracy using six gene expression datasets produced using DNA microarray. The stability is calculated by Equation 1.16 and prediction accuracy is evaluated by SVM. Across all the 54 experiments, the ensemble methods, basic function perturbation and EFSIS, are slightly better than the individual methods in prediction accuracy. EFSIS performs much better than basic function perturbation in stability. The work was presented in **Paper IV** and the source codes are available on GitHub (https://github.com/zhxiaokang/EFSIS).

Chapter 4

Discussion

4.1 Challenges of working with a non-model species

The dCod 1.0 project focuses on Atlantic cod (*Gadus morhua*). Atlantic cod is one of the most commercially important species for North Atlantic fisheries and is also commonly used in marine pollution monitoring and environmental toxicology studies [80; 120–125]. Furthermore, its genome has been continuously sequenced and annotated [9; 10] which makes it possible to study its systems toxicology using the omics approaches.

But as a non-model species, the effort and work devoted to the study of Atlantic cod and the available resources are still limited, compared with the model species, such as human (*Homo sapiens*), mouse (*Mus musculus*), and zebrafish (*Danio rerio*). In **Paper II**, we studied 7 RNA-Seq analysis workflows in the past three years, and three of them only support the two model species: human and mouse. In a study to reconstruct the metabolic pathway model for Atlantic cod [11], we found that the potential options of computational tools that can help automate the reconstruction process are also highly limited.

As more work is being done aiming for a better sequenced and annotated Atlantic cod genome, since 2011 when the first version of Atlantic cod genome (gadMor1) was published [9], two new versions have been published in 2017 (gadMor2) [10] and 2019 (gadMor3) respectively. At the early stage of the dCod 1.0 project, we mainly used gadMor1 (**Paper I**). Later on, we also generated a *de novo* assembly integrating gadMor1, gadMor2 and lots of RNA-Seq data from different developmental stages and tissues of Atlantic cod [11]. After gadMor3 was released, we started using gadMor3 [11]. Considering that the genome of such non-model species is being updated rapidly, we designed the workflow (RASflow in **Paper II**) in a modular way so that each functional part of the whole workflow can easily be changed, and the application of Conda [106] and Snakemake [107] makes the workflow highly reproducible.

4.2 Biomarker discovery using gene expression data

An important data analysis of gene expression data is biomarker discovery, and is typically performed on a set of gene expression profiles of samples from different groups. The mainstream is Differential Expression Analysis (DEA) which finds out the significantly differentially expressed genes across different groups using statistical tests. It has been popular for its simplicity and interpretability. But there are also some drawbacks of such methods. Fold Change can measure the magnitude of difference between groups, but it ignores the variance within each group, so it fails to find out the genes of high reproducibility with comparably low differentiality. The statistical test such as Student's t-test requires a specific distribution of samples which can not always be satisfied [55]. Recent debates include alleged misuse of P-value [126–131]. The thresholds for Fold Change and P-value also significantly alter the gene expression data interpretations [132].

In recent years, machine learning techniques have been gaining popularity in gene expression data analysis for biomarker discovery. The two relevant machine learning techniques are feature selection and classification.

Lyons-Weiler et al. proposed to combine statistical tests with classification [55]. Targeting at achieving the highest classification accuracy, they choose the threshold for P-value (or Fold Change). Van IJzendoorn et al. proposed to combine statistical tests with feature selection [133]. They apply Random Forest to the significantly differentially expressed genes (adjusted P-value < 0.05) to pick out the most important genes.

Some other researchers compared statistical tests and feature selection to see which one performs better. Blanco et al. studied the application of both classical statistical approaches and machine learning methods on RNA-Seq data for cancer research, and found that there is a big overlap between the biomarker genes identified by Random Forest and classical statistical approach (edgeR), but GLMNET is different in terms of the choice of genes [134]. Clark et al. claimed that their multivariate approach (Characteristic Direction method) outperformed all the tested univariate statistical test approaches, including Significance Analysis of Microarrays (SAM) and Linear Models for Microarray Data (limma) [135] for DNA microarray data, and DESeq and edgeR for RNA-Seq data [112]. In the study of human preimplantation development using single-cell RNA-Seq data, Liang et al. found that the F-score algorithm (from Support Vector

Machine) achieves the highest prediction accuracy with the least genes compared with classical statistical tests provided by DESeq, edgeR, and limma, and the functional enrichment analysis showed that the F-score algorithm can obtain key signaling pathways related to embryo development [136].

4.3 Evaluation of biomarker discovery methods

Regarding the evaluation of performance while comparing statistical tests and feature selection, no standard way has been broadly approved. Many evaluation metrics have been proposed so far, and they can be summarized into three categories which are described as follows.

The most straightforward way is to compare the selected genes to an already known list of true biomarkers. But the difficult part for this is to establish such a "gold standard". Williams et al. produced a test dataset generated using highly purified human classical and nonclassical monocyte subsets from a clinical cohort [61]. They applied two approaches, SAM and limma, to pick out the differentially expressed genes. The genes at the intersection of these two methods are used as the "gold standard" to evaluate the performance of other differential expression analysis tools. Clark et al. [112] used the genes associated with binding regions which are differentially bound by STAT3 as the "gold standard" of differentially expressed genes supported by a relevant finding by Hardee et al. [137]. Di Camillo et al. generated a simulated dataset so that the biomarkers are already known which can then be used as the "gold standard" [138].

Another strategy is to generate multiple biomarker lists and simply compare between them. The comparison is carried out along two dimensions: compare the sets of genes selected as biomarkers by the different methods [134]; apply one method to the resampled data subsets generated from the original dataset and compare the results from those subsets which is widely used to evaluate the method's stability. Most researchers calculate the stability of a method by looking at the overlap of gene sets that it selects from different data subsets [86–88; 139–141]. Dessì et al. proposed to compare the gene lists in functional terms based on the molecular function GO annotations [142].

As defined, a biomarker is an indicator of biological states. So a good set of biomarker genes should be discriminative genes that can characterize samples from different biological states. Therefore, using the biomarker genes, a classifier should obtain high prediction accuracy. This is widely used to evaluate the performance of a biomarker discovery method [136; 139–142].

However, each of these approaches have challenges. For the first strategy, it is difficult to establish a convincing "gold standard". Comparing the gene sets from different methods can only indicate that some methods tend to select some common genes, but it can not tell which method is better. There are many ways proposed for calculating the stability, but some of them tend to give a higher stability when there are more genes in the lists [87], which is unfair for the methods that are more strict in selecting redundant genes. The evaluation of the ability of a biomarker discovery method to improve a classifier's prediction accuracy is influenced by the choice of classification algorithm. For example, the SVM would be expected to work better with a feature selection method implemented in its own package [136]. Or to be more general, the choice of the classification algorithm can affect the evaluation of the biomarker discovery methods [140].

Chapter 5

Conclusions and future prospects

In this thesis, we have mainly focused on biomarker discovery using gene expression data. Two main approaches are statistical tests and machine learning or rather feature selection. Prior to applying statistical tests to RNA-Seq data to select the differentially expressed genes as biomarkers, the sequencing data needs to be pre-processed through a series of procedures. We developed a maximally generalized and modular RNA-Seq analysis workflow, which can be applied to a wide range of applications and can easily be extended to new functions. On the feature selection aspect, we found that the method performs quite differently depending on the particular datasets. To address that issue, we proposed an ensemble framework combining several individual feature selection methods. The results showed that our proposed method achieves both high stability and prediction accuracy.

For the future prospects, the proposed ensemble framework may be applied to statistical tests, instead of feature selection. Since there has been a worry about misuse of P-value and the choice of its threshold, the P-value can be used as an intermediate parameter instead of the final evaluation metric of whether a gene is significantly differentially expressed. But in the end, it will again come to the issue of how to evaluate the performance of the method or the quality of the selected genes. So more efforts should be devoted to that direction.

In this thesis, we separate the steps of biomarker discovery and classification to select the biomarkers independent of the classifier, so that the selected biomarkers are as generalized as possible and can be applied to more circumstances and in different research topics. But when it comes to disease diagnosis, classification of disease types is of the main interest [143; 144]. So in this case, the biomarker discovery and the classification steps should be evaluated together: to achieve a high and stable prediction accuracy. The stability here means that the prediction accuracy is stably high regardless of the change of training samples, instead of the stability of biomarker discovery methods which has been discussed in the thesis. Deep learning is not used in this thesis for its low interpretability, but with the research focus switched, deep learning becomes a good choice for its robust and high prediction performance [145–148].

The whole thesis has been focusing on using gene expression data itself to discover biomarkers, but some prior knowledge and other data may also be incorporated to indicate the importance of genes. Systems biology models can be used to identify the mechanistically essential genes in distinguishing different biological states or the genes involved in the core metabolic pathways of a toxicant [149–151]. The known chemical defensome genes can also be used as prior knowledge in identifying biomarker genes in response to a chemical [152; 153]. Integration of multi-omics data can also strengthen the predictive power of identifying biomarkers at a more systematic level [154–156].

Regarding the challenges of working with non-model species, we have gone through many problems in dCod 1.0 project, and have also put our efforts in solving them, such as developing RNA-Seq analysis workflow applicable to non-model species [33], and generating a draft liver reconstruction model of Atlantic cod (*Gadus morhua*) [11]. There will be more and more novel genomes being analyzed motivated by initiatives such as the Earth BioGenome Project [157]. We anticipate that more research on non-model species will be conducted. Some efforts can be devoted to but not limited to the following directions: annotating the novel genomes because a high-quality annotated genome is a basis of many other studies, and the use of updated Atlantic cod assembly in dCod 1.0 project is one example; developing methods applicable to non-model species such as RASflow; and so on.

Bibliography

[1] Heather J M and Chain B 2016 *Genomics* **107** 1–8 URL http://dx.doi.org/10.1016/j.ygeno.2015.11.003 1.1

[2] Goodwin S, McPherson J D and McCombie W R 2016 *Nature Reviews. Genetics* **17** 333–351 URL http://dx.doi.org/10.1038/nrg.2016.49 1.1

[3] van Dijk E L, Jaszczyszyn Y, Naquin D and Thermes C 2018 *Trends in Genetics* **34** 666–681 ISSN 01689525 URL https://linkinghub.elsevier.com/retrieve/pii/S0168952518300969 1.1

[4] Consortium I H G S 2004 *Nature* **431** 931–945 ISSN 1476-4687 URL http://dx.doi.org/10.1038/nature03001 1.1

[5] Venter J C, Adams M D, Myers E W, Li P W, Mural R J, Sutton G G, Smith H O, Yandell M, Evans C A, Holt R A, Gocayne J D, Amanatides P, Ballew R M, Huson D H, Wortman J R, Zhang Q, Kodira C D, Zheng X H, Chen L, Skupski M, Subramanian G, Thomas P D, Zhang J, Gabor Miklos G L, Nelson C, Broder S, Clark A G, Nadeau J, McKusick V A, Zinder N, Levine A J, Roberts R J, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian A E, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman T J, Higgins M E, Ji R R, Ke Z, Ketchum K A, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov G V, Milshina N, Moore H M, Naik A K, Narayan V A, Neelam B, Nusskern D, Rusch D B, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng M L, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun

S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers Y H, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint N N, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril J F, Guigó R, Campbell M J, Sjolander K V, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang Y H, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A and Zhu X 2001 *Science* **291** 1304–1351 URL http://dx.doi.org/10.1126/science.1058040 1.1

[6] Lander E S, Linton L M, Birren B, Nusbaum C, Zody M C, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov J P, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin J C, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston R H, Wilson R K, Hillier L W, McPherson J D, Marra M A, Mardis E R, Fulton L A, Chinwalla A T, Pepin K H, Gish W R, Chissoe S L, Wendl M C, Delehaunty K D, Miner T L, Delehaunty A, Kramer J B, Cook L L, Fulton R S, Johnson D L, Minx P J, Clifton S W, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J F, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs R A, Muzny D M, Scherer S E, Bouck J B, Sodergren E J, Worley K C, Rives C M, Gorrell J H, Metzker M L, Naylor S L, Kucherlapati R S, Nelson D L, Weinstock G M, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith D R, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee

H M, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis R W, Federspiel N A, Abola A P, Proctor M J, Myers R M, Schmutz J, Dickson M, Grimwood J, Cox D R, Olson M V, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans G A, Athanasiou M, Schultz R, Roe B A, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie W R, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey J A, Bateman A, Batzoglou S, Birney E, Bork P, Brown D G, Burge C B, Cerutti L, Chen H C, Church D, Clamp M, Copley R R, Doerks T, Eddy S R, Eichler E E, Furey T S, Galagan J, Gilbert J G, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson L S, Jones T A, Kasif S, Kaspryzk A, Kennedy S, Kent W J, Kitts P, Koonin E V, Korf I, Kulp D, Lancet D, Lowe T M, McLysaght A, Mikkelsen T, Moran J V, Mulder N, Pollara V J, Ponting C P, Schuler G, Schultz J, Slater G, Smit A F, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf Y I, Wolfe K H, Yang S P, Yeh R F, Collins F, Guyer M S, Peterson J, Felsenfeld A, Wetterstrand K A, Patrinos A, Morgan M J, de Jong P, Catanese J J, Osoegawa K, Shizuya H, Choi S, Chen Y J, Szustakowki J and Consortium I H G S 2001 *Nature* **409** 860–921 URL http://dx.doi.org/10.1038/35057062 1.1

[7] Collins F S, Morgan M and Patrinos A 2003 *Science* **300** 286–290 URL http://dx.doi.org/10.1126/science.1084564 1.1

[8] Rothberg J M and Leamon J H 2008 *Nature Biotechnology* **26** 1117–1124 URL http://dx.doi.org/10.1038/nbt1485 1.1

[9] Star B, Nederbragt A J, Jentoft S, Grimholt U, Malmstrøm M, Gregers T F, Rounge T B, Paulsen J, Solbakken M H, Sharma A, Wetten O F, Lanzén A, Winer R, Knight J, Vogel J H, Aken B, Andersen O, Lagesen K, Tooming-Klunderud A, Edvardsen R B, Tina K G, Espelund M, Nepal C, Previti C, Karlsen B O, Moum T, Skage M, Berg P R, Gjøen T, Kuhl H, Thorsen J, Malde K, Reinhardt R, Du L, Johansen S D, Searle S, Lien S, Nilsen F, Jonassen I, Omholt S W, Stenseth N C and Jakobsen K S 2011 *Nature* **477** 207–210 ISSN 1476-4687 URL http://dx.doi.org/10.1038/nature10342 1.1, 4.1

[10] Tørresen O K, Star B, Jentoft S, Reinar W B, Grove H, Miller J R, Walenz B P, Knight J, Ekholm J M, Peluso P, Edvardsen R B, Tooming-Klunderud A, Skage M, Lien S, Jakobsen K S and Nederbragt A J 2017 *BMC Genomics* **18** 95 URL http://dx.doi.org/10.1186/s12864-016-3448-x 1.1, 4.1

[11] Hanna E M, Zhang X, Eide M, Fallahi S, Furmanek T, Yadetie F, Zielinski D C, Goksøyr A and Jonassen I 2020 *BioRxiv* URL http://biorxiv.org/lookup/doi/10.1101/2020.06.23.162792 1.1, 4.1, 5

[12] Hieter P and Boguski M 1997 *Science* **278** 601–602 URL http://dx.doi.org/10.1126/science.278.5338.601 1.2

[13] Bunnik E M and Le Roch K G 2013 *Advances in wound care* **2** 490–498 URL http://dx.doi.org/10.1089/wound.2012.0379 1.2

[14] Crick F 1970 *Nature* **227** 561–563 URL http://dx.doi.org/10.1038/227561a0 1.2.1

[15] Koonin E V, Gorbalenya A E and Chumakov K M 1989 *FEBS Letters* **252** 42–46 URL http://dx.doi.org/10.1016/0014-5793(89)80886-5 1.2.1

[16] Varmus H 1988 *Science* **240** 1427–1435 URL http://dx.doi.org/10.1126/science.3287617 1.2.1

[17] Mantione K J, Kream R M, Kuzelova H, Ptacek R, Raboch J, Samuel J M and Stefano G B 2014 *Medical science monitor basic research* **20** 138–142 URL http://dx.doi.org/10.12659/MSMBR.892101 1.2.1

[18] Steen H and Mann M 2004 *Nature Reviews. Molecular Cell Biology* **5** 699–711 URL http://dx.doi.org/10.1038/nrm1468 1.2.1

[19] Barbazuk W B, Emrich S J, Chen H D, Li L and Schnable P S 2007 *The Plant Journal: for Cell and Molecular Biology* **51** 910–918 URL http://dx.doi.org/10.1111/j.1365-313X.2007.03193.x 1.2.2

[20] Emrich S J, Barbazuk W B, Li L and Schnable P S 2007 *Genome Research* **17** 69–73 URL http://dx.doi.org/10.1101/gr.5145806 1.2.2

[21] Cloonan N, Forrest A R R, Kolle G, Gardiner B B A, Faulkner G J, Brown M K, Taylor D F, Steptoe A L, Wani S, Bethel G, Robertson A J, Perkins A C, Bruce S J, Lee C C, Ranade S S, Peckham H E, Manning J M, McKernan K J and Grimmond S M 2008 *Nature Methods* **5** 613–619 URL http://dx.doi.org/10.1038/nmeth.1223 1.2.2

[22] Lister R, O'Malley R C, Tonti-Filippini J, Gregory B D, Berry C C, Millar A H and Ecker J R 2008 *Cell* **133** 523–536 ISSN 1097-4172 URL http://dx.doi.org/10.1016/j.cell.2008.03.029 1.2.2

[23] Marioni J C, Mason C E, Mane S M, Stephens M and Gilad Y 2008 *Genome Research* **18** 1509–1517 URL http://dx.doi.org/10.1101/gr.079558.108 1.2.2

[24] Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S and Marra M 2008 *Biotechniques* **45** 81–94 URL http://dx.doi.org/10.2144/000112900 1.2.2

[25] Mortazavi A, Williams B A, McCue K, Schaeffer L and Wold B 2008 *Nature Methods* **5** 621–628 ISSN 1548-7105 URL http://dx.doi.org/10.1038/nmeth.1226 1.2.2, 1.2.2

[26] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M and Snyder M 2008 *Science* **320** 1344–1349 ISSN 1095-9203 URL http://dx.doi.org/10.1126/science.1158441 1.2.2

[27] Vera J C, Wheat C W, Fescemyer H W, Frilander M J, Crawford D L, Hanski I and Marden J H 2008 *Molecular Ecology* **17** 1636–1647 ISSN 1365-294X URL http://dx.doi.org/10.1111/j.1365-294X.2008.03666.x 1.2.2

[28] Wilhelm B T, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett C J, Rogers J and Bähler J 2008 *Nature* **453** 1239–1243 URL http://dx.doi.org/10.1038/nature07002 1.2.2

[29] Barrett T, Wilhite S E, Ledoux P, Evangelista C, Kim I F, Tomashevsky M, Marshall K A, Phillippy K H, Sherman P M, Holko M, Yefanov A, Lee H, Zhang N, Robertson C L, Serova N, Davis S and Soboleva A 2013 *Nucleic Acids Research* **41** D991–5 URL http://dx.doi.org/10.1093/nar/gks1193 1.2.2

[30] Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca N A, Petryszak R, Papatheodorou I, Sarkans U and Brazma A 2019 *Nucleic Acids Research* **47** D711–D715 URL http://dx.doi.org/10.1093/nar/gky964 1.2.2

[31] Leinonen R, Sugawara H, Shumway M and Collaboration I N S D 2011 *Nucleic Acids Research* **39** D19–21 URL http://dx.doi.org/10.1093/nar/gkq1019 1.2.2

[32] Cock P J A, Fields C J, Goto N, Heuer M L and Rice P M 2010 *Nucleic Acids Research* **38** 1767–1771 URL http://dx.doi.org/10.1093/nar/gkp1137 1.2.2

[33] Zhang X and Jonassen I 2020 *BMC Bioinformatics* **21** 110 URL http://dx.doi.org/10.1186/s12859-020-3433-x 1.2.2, 5

[34] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczeniak M W, Gaffney D J, Elo L L, Zhang X and Mortazavi A 2016 *Genome Biology* **17** 13 ISSN 1474-760X URL http://genomebiology.com/2016/17/1/13 1.2.2, 1.2.2

[35] Bolger A M, Lohse M and Usadel B 2014 *Bioinformatics* **30** 2114–2120 URL http://dx.doi.org/10.1093/bioinformatics/btu170 1.2.2

[36] Kim D, Langmead B and Salzberg S L 2015 *Nature Methods* **12** 357–360 URL http://dx.doi.org/10.1038/nmeth.3317 1.2.2

[37] Dobin A, Davis C A, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M and Gingeras T R 2013 *Bioinformatics* **29** 15–21 URL http://dx.doi.org/10.1093/bioinformatics/bts635 1.2.2

[38] Li H and Durbin R 2009 *Bioinformatics* **25** 1754–1760 URL http://dx.doi.org/10.1093/bioinformatics/btp324 1.2.2

[39] Langmead B and Salzberg S L 2012 *Nature Methods* **9** 357–359 URL http://dx.doi.org/10.1038/nmeth.1923 1.2.2

[40] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R and Salzberg S 2013 *Genome Biology* **14** R36 URL http://dx.doi.org/10.1186/gb-2013-14-4-r36 1.2.2

[41] Liao Y, Smyth G K and Shi W 2014 *Bioinformatics* **30** 923–930 URL http://dx.doi.org/10.1093/bioinformatics/btt656 1.2.2

[42] Anders S, Pyl P T and Huber W 2015 *Bioinformatics* **31** 166–169 URL http://dx.doi.org/10.1093/bioinformatics/btu638 1.2.2

[43] Patro R, Duggal G, Love M I, Irizarry R A and Kingsford C 2017 *Nature Methods* **14** 417–419 URL http://dx.doi.org/10.1038/nmeth.4197 1.2.2

[44] Bray N L, Pimentel H, Melsted P and Pachter L 2016 *Nature Biotechnology* **34** 525–527 ISSN 1087-0156 URL http://www.nature.com/doifinder/10.1038/nbt.3519 1.2.2

[45] Patro R, Mount S M and Kingsford C 2014 *Nature Biotechnology* **32** 462–464 URL http://dx.doi.org/10.1038/nbt.2862 1.2.2

[46] Stark R, Grzelak M and Hadfield J 2019 *Nature Reviews. Genetics* **20** 631–656 ISSN 1471-0056 URL http://www.nature.com/articles/s41576-019-0150-2 1.2.2

[47] Oshlack A and Wakefield M J 2009 *Biology Direct* **4** 14 URL http://dx.doi.org/10.1186/1745-6150-4-14 1.2.2

[48] Bullard J H, Purdom E, Hansen K D and Dudoit S 2010 *BMC Bioinformatics* **11** 94 URL http://dx.doi.org/10.1186/1471-2105-11-94 1.2.2

[49] Evans C, Hardin J and Stoebel D M 2018 *Briefings in Bioinformatics* **19** 776–792 URL http://dx.doi.org/10.1093/bib/bbx008 1.2.2

[50] Trapnell C, Williams B A, Pertea G, Mortazavi A, Kwan G, van Baren M J, Salzberg S L, Wold B J and Pachter L 2010 *Nature Biotechnology* **28** 511–515 URL http://dx.doi.org/10.1038/nbt.1621 1.2.2

[51] Wagner G P, Kin K and Lynch V J 2012 *Theory in Biosciences = Theorie in Den Biowissenschaften* **131** 281–285 URL http://dx.doi.org/10.1007/s12064-012-0162-3 1.2.2

[52] Robinson M D and Oshlack A 2010 *Genome Biology* **11** R25 URL http://dx.doi.org/10.1186/gb-2010-11-3-r25 1.2.2

[53] Robinson M D, McCarthy D J and Smyth G K 2010 *Bioinformatics* **26** 139–140 URL http://dx.doi.org/10.1093/bioinformatics/btp616 1.2.2, 1.2.2, 3.1

[54] Anders S and Huber W 2010 *Genome Biology* **11** R106 ISSN 1465-6914 URL http://dx.doi.org/10.1186/gb-2010-11-10-r106 1.2.2, 1.2.2

[55] Lyons-Weiler J, Patel S and Bhattacharya S 2003 *Genome Research* **13** 503–512 URL http://dx.doi.org/10.1101/gr.104003 1.2.2, 4.2

[56] Jafari M and Ansari-Pour N 2019 *Cell journal* **20** 604–607 URL http://dx.doi.org/10.22074/cellj.2019.5992 1.2.2

[57] Kim T K 2015 *Korean Journal of Anesthesiology* **68** 540–546 URL http://dx.doi.org/10.4097/kjae.2015.68.6.540 1.2.2

[58] Wang L, Feng Z, Wang X, Wang X and Zhang X 2010 *Bioinformatics* **26** 136–138 URL http://dx.doi.org/10.1093/bioinformatics/btp612 1.2.2

[59] Love M I, Huber W and Anders S 2014 *Genome Biology* **15** 550 URL http://dx.doi.org/10.1186/s13059-014-0550-8 1.2.2, 3.1

[60] McCarthy D J, Chen Y and Smyth G K 2012 *Nucleic Acids Research* **40** 4288–4297 URL http://dx.doi.org/10.1093/nar/gks042 1.2.2, 3.1

[61] Williams C R, Baccarella A, Parrish J Z and Kim C C 2017 *BMC Bioinformatics* **18** 38 URL http://dx.doi.org/10.1186/s12859-016-1457-z 1.2.2, 3.1, 3.2, 4.3

[62] Timbrell J 2019 *Introduction to Toxicology* (CRC Press) ISBN 9781000026962 URL https://www.taylorfrancis.com/books/9781000026962 1.3

[63] on Communicating Toxicogenomics Information to Nonexperts N R C U C 2005 *Communicating toxicogenomics information to nonexperts: A workshop*

summary The National Academies Collection: Reports funded by National Institutes of Health (Washington (DC): National Academies Press (US)) ISBN 0309095387 URL http://dx.doi.org/10.17226/11179 1.3

[64] Alexander-Dann B, Pruteanu L L, Oerton E, Sharma N, Berindan-Neagoe I, Módos D and Bender A 2018 *Molecular Omics* **14** 218–236 ISSN 2515-4184 URL http://xlink.rsc.org/?DOI=C8MO00042E 1.3

[65] Beedanagari S, Vulimiri S, Bhatia S and Mahadevan B 2014 Genotoxicity biomarkers *Biomarkers in Toxicology* (Elsevier) pp 729–742 ISBN 9780124046306 URL https://linkinghub.elsevier.com/retrieve/pii/B9780124046306000439 1.3

[66] Storey J D and Tibshirani R 2003 *Proceedings of the National Academy of Sciences of the United States of America* **100** 9440–9445 ISSN 0027-8424 URL http://dx.doi.org/10.1073/pnas.1530509100 1.3

[67] Dhall D, Kaur R and Juneja M 2020 Machine learning: A review of the algorithms and its applications *Proceedings of ICRIC 2019: recent innovations in computing* (*Lecture notes in electrical engineering* vol 597) ed Singh P K, Kar A K, Singh Y, Kolekar M H and Tanwar S (Cham: Springer International Publishing) pp 47–63 ISBN 978-3-030-29406-9 URL http://link.springer.com/10.1007/978-3-030-29407-6_5 1.4

[68] Chapelle O, Scholkopf B and Zien Eds A 2009 *IEEE Transactions on Neural Networks* **20** 542–542 ISSN 1045-9227 URL http://ieeexplore.ieee.org/document/4787647/ 1.4

[69] Jain A K, Murty M N and Flynn P J 1999 *ACM Computing Surveys* **31** 264–323 ISSN 03600300 URL http://portal.acm.org/citation.cfm?doid=331499.331504 1.4

[70] Sutton R and Barto A 1998 *IEEE Transactions on Neural Networks* **9** 1054–1054 ISSN 1045-9227 URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=712192 1.4

[71] Kotsiantis S, Zaharakis I and Pintelas P 2007 *Emerging artificial intelligence applications in computer engineering* **160** 3–24 URL https://cmapspublic2.ihmc.us/rid=1MYWPSVZJ-16FZV2W-30S4/Reference_Classification_Comparison.pdf 1.4

[72] Tang J, Alelyani S and Liu H 2014 *Data classification: Algorithms and applications* URL https://www.cc.gatech.edu/~hic/{CS7616}/Papers/Tang-et-al-2014.pdf 1.5

[73] Tu Y, Stolovitzky G and Klein U 2002 *Proceedings of the National Academy of Sciences of the United States of America* **99** 14031–14036 URL http://dx.doi.org/10.1073/pnas.222164199 1.5

[74] Zhao S, Fung-Leung W P, Bittner A, Ngo K and Liu X 2014 *Plos One* **9** e78644 URL http://dx.doi.org/10.1371/journal.pone.0078644 1.5

[75] Sha Y, Phan J H and Wang M D 2015 *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2015** 6461–6464 URL http://dx.doi.org/10.1109/EMBC.2015.7319872 1.5

[76] Guyon I and Elisseeff A 2003 *Journal of machine learning research* **3** 1157–1182 URL http://www.jmlr.org/papers/v3/guyon03a.html 1.5

[77] Peng H, Long F and Ding C 2005 *IEEE transactions on pattern analysis and machine intelligence* **27** 1226–1238 URL http://dx.doi.org/10.1109/TPAMI.2005.159 1.5, 3.2

[78] Kirpich A, Ainsworth E A, Wedow J M, Newman J R B, Michailidis G and McIntyre L M 2018 *Plos One* **13** e0197910 URL http://dx.doi.org/10.1371/journal.pone.0197910 1.5

[79] Zang C, Wang T, Deng K, Li B, Hu S, Qin Q, Xiao T, Zhang S, Meyer C A, He H H, Brown M, Liu J S, Xie Y and Liu X S 2016 *Nature Communications* **7** 11305 URL http://dx.doi.org/10.1038/ncomms11305 1.5

[80] Yadetie F, Zhang X, Hanna E M, Aranguren-Abadía L, Eide M, Blaser N, Brun M, Jonassen I, Goksøyr A and Karlsen O A 2018 *Aquatic Toxicology* **201** 174–186 URL http://dx.doi.org/10.1016/j.aquatox.2018.06.003 1.5, 3.1, 4.1

[81] Ringnér M 2008 *Nature Biotechnology* **26** 303–304 URL http://dx.doi.org/10.1038/nbt0308-303 1.5

[82] Abdi H 2003 *Encyclopedia for research methods for the social sciences* **6** 792–795 URL https://d1wqtxts1xzle7.cloudfront.net/34923345/Abdi-PLSR2007-pretty.pdf?1411988001=&response-content-disposition=inline%3B+filename%3DPartial_Least_Square_Regression_PLS-Regr.pdf&Expires=1593621010&Signature=DeeeGN8x1ZXPxsW-bqXiT7nNwiYC~J9Rlz1wYXk6qdGYaIkJJ8qrjt98LTk0~92NkqKzP8eR8_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA 1.5

[83] He Z and Yu W 2010 *Computational biology and chemistry* **34** 215–225 URL http://dx.doi.org/10.1016/j.compbiolchem.2010.07.002 1.5

[84] Fawcett T 2006 *Pattern recognition letters* **27** 861–874 ISSN 01678655 URL http://linkinghub.elsevier.com/retrieve/pii/S016786550500303X 1.6.1

[85] Kalousis A, Prados J and Hilario M 2007 *Knowledge and information systems* **12** 95–116 ISSN 0219-1377 URL http://link.springer.com/10.1007/s10115-006-0040-8 1.6.2

[86] Kalousis A, Prados J and Hilario M 2005 Stability of feature selection algorithms *Fifth IEEE International Conference on Data Mining (ICDM'05)* (IEEE) pp 218–225 ISBN 0-7695-2278-5 URL http://ieeexplore.ieee.org/document/1565682/ 1.6.2, 4.3

[87] Kuncheva L 2007 *Artificial intelligence and applications* 421 URL https://pdfs.semanticscholar.org/9c4f/006d5547cddecef588e66ab6b33e6845b33a.pdf 1.6.2, 1.6.2, 4.3

[88] Davis C A, Gerick F, Hintermair V, Friedel C C, Fundel K, Küffner R and Zimmer R 2006 *Bioinformatics* **22** 2356–2363 URL http://dx.doi.org/10.1093/bioinformatics/btl400 1.6.2, 1.7.2, 4.3

[89] Tan N C, Fisher W G, Rosenblatt K P and Garner H R 2009 *BMC Bioinformatics* **10** 144 URL http://dx.doi.org/10.1186/1471-2105-10-144 1.7.1

[90] Ben Brahim A and Limam M 2013 Robust ensemble feature selection for high dimensional data sets *2013 International Conference on High Performance Computing & Simulation (HPCS)* (IEEE) pp 151–157 ISBN 978-1-4799-0838-7 URL http://ieeexplore.ieee.org/document/6641406/ 1.7.1

[91] Ahmed S, Zhang M and Peng L 2014 *Connection science* **26** 215–243 ISSN 0954-0091 URL http://www.tandfonline.com/doi/abs/10.1080/09540091.2014.906388 1.7.1

[92] Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V and Alonso-Betanzos A 2017 *Knowledge-Based Systems* **118** 124–139 ISSN 09507051 URL http://linkinghub.elsevier.com/retrieve/pii/S0950705116304749 1.7.1, 1.7.2

[93] Bach F R 2008 Bolasso *Proceedings of the 25th international conference on Machine learning - ICML '08* (New York, New York, USA: ACM Press) pp 33–40 URL http://portal.acm.org/citation.cfm?doid=1390156.1390161 1.7.2

[94] Abeel T, Helleputte T, Van de Peer Y, Dupont P and Saeys Y 2010 *Bioinformatics* **26** 392–398 URL http://dx.doi.org/10.1093/bioinformatics/btp630 1.7.2

[95] Pes B, Dessì N and Angioni M 2017 *Information Fusion* **35** 132–147 ISSN 15662535 URL http://linkinghub.elsevier.com/retrieve/pii/S1566253516300847 1.7.2

[96] Chiew K L, Tan C L, Wong K, Yong K S and Tiong W K 2019 *Information sciences* **484** 153–166 ISSN 00200255 URL https://linkinghub.elsevier.com/retrieve/pii/S0020025519300763 1.7.3, 1.7.3

[97] Breitling R, Armengaud P, Amtmann A and Herzyk P 2004 *FEBS Letters* **573** 83–92 ISSN 0014-5793 URL http://dx.doi.org/10.1016/j.febslet.2004.07.055 1.7.3

[98] McDonald J 2009 *Handbook of biological statistics* vol 2 (Baltimore MD: sparky house publishing) URL http://www.uni-koeln.de/math-nat-fak/genetik/groups/Langer/HandbookBioStatSecond.pdf 3.1

[99] Alonso A, Lasseigne B N, Williams K, Nielsen J, Ramaker R C, Hardigan A A, Johnston B, Roberts B S, Cooper S J, Marsal S and Myers R M 2017 *Bioinformatics* **33** 1727–1729 URL http://dx.doi.org/10.1093/bioinformatics/btx023 3.1

[100] Sahraeian S M E, Mohiyuddin M, Sebra R, Tilgner H, Afshar P T, Au K F, Bani Asadi N, Gerstein M B, Wong W H, Snyder M P, Schadt E and Lam H Y K 2017 *Nature Communications* **8** 59 URL http://dx.doi.org/10.1038/s41467-017-00050-4 3.1

[101] Torre D, Lachmann A and Ma'ayan A 2018 *Cell Systems* **7** 556–561.e3 ISSN 24054712 URL https://linkinghub.elsevier.com/retrieve/pii/S2405471218304320 3.1

[102] Wang D 2018 *Briefings in Bioinformatics* **19** 622–626 URL http://dx.doi.org/10.1093/bib/bbw143 3.1

[103] Cornwell M, Vangala M, Taing L, Herbert Z, Köster J, Li B, Sun H, Li T, Zhang J, Qiu X, Pun M, Jeselsohn R, Brown M, Liu X S and Long H W 2018 *BMC Bioinformatics* **19** 135 ISSN 1471-2105 URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2139-9 3.1

[104] Kohen R, Barlev J, Hornung G, Stelzer G, Feldmesser E, Kogan K, Safran M and Leshkowitz D 2019 *BMC Bioinformatics* **20** 154 ISSN 1471-2105 URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2728-2 3.1

[105] Orjuela S, Huang R, Hembach K M, Robinson M D and Soneson C 2019 *G3 (Bethesda, Md.)* **9** 2089–2096 URL http://dx.doi.org/10.1534/g3.119.400185 3.1

[106] Analytics C 2016 Anaconda software distribution URL https://www.anaconda.com/ 3.1, 4.1

[107] Köster J and Rahmann S 2012 *Bioinformatics* **28** 2520–2522 URL http://dx.doi.org/10.1093/bioinformatics/bts480 3.1, 4.1

[108] Ren S, Peng Z, Mao J H, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W, Chen C, Wang F, Guo X, Lu J, Yang J, Wei M, Tian Z, Guan Y, Tang L, Xu C, Wang L, Gao X, Tian W, Wang J, Yang H, Wang J and Sun Y 2012 *Cell Research* **22** 806–821 URL http://dx.doi.org/10.1038/cr.2012.30 3.1

[109] Tusher V G, Tibshirani R and Chu G 2001 *Proceedings of the National Academy of Sciences of the United States of America* **98** 5116–5121 URL http://dx.doi.org/10.1073/pnas.091062498 3.2

[110] Yadetie F, Bjørneklett S, Garberg H K, Oveland E, Berven F, Goksøyr A and Karlsen O A 2016 *BMC Genomics* **17** 554 URL http://dx.doi.org/10.1186/s12864-016-2864-2 3.2

[111] Eckert M A, Coscia F, Chryplewicz A, Chang J W, Hernandez K M, Pan S, Tienda S M, Nahotko D A, Li G, Blaenovi I, Lastra R R, Curtis M, Yamada S D, Perets R, McGregor S M, Andrade J, Fiehn O, Moellering R E, Mann M and Lengyel E 2019 *Nature* **569** 723–728 ISSN 0028-0836 URL http://www.nature.com/articles/s41586-019-1173-8 3.2

[112] Clark N R, Hu K S, Feldmann A S, Kou Y, Chen E Y, Duan Q and Ma'ayan A 2014 *BMC Bioinformatics* **15** 79 URL http://dx.doi.org/10.1186/1471-2105-15-79 3.2, 4.2, 4.3

[113] Cortes C and Vapnik V 1995 *Machine learning* **20** 273–297 ISSN 0885-6125 URL http://link.springer.com/10.1007/BF00994018 3.2

[114] Breiman L 2001 *Machine learning* **45** 5–32 URL https://link.springer.com/article/10.1023/a:1010933404324 3.2

[115] Friedman J, Hastie T and Tibshirani R 2010 *Journal of statistical software* **33** 1–22 URL http://dx.doi.org/10.18637/jss.v033.i01 3.2

[116] Quinlan J R 1986 *Machine learning* **1** 81–106 ISSN 0885-6125 URL http://link.springer.com/10.1007/BF00116251 3.3

[117] Kononenko I 1994 Estimating attributes: Analysis and extensions of RELIEF *Machine Learning: ECML-94* (*Lecture notes in computer science* vol 784) ed Bergadano F, Raedt L, Carbonell J G, Siekmann J, Goos G and Hartmanis

J (Berlin, Heidelberg: Springer Berlin Heidelberg) pp 171–182 ISBN 978-3-540-48365-6 URL http://link.springer.com/10.1007/3-540-57868-4_57 3.3

[118] Kira K and Rendell L A 1992 A practical approach to feature selection *Machine learning proceedings 1992* (Elsevier) pp 249–256 ISBN 9781558602472 URL https://linkinghub.elsevier.com/retrieve/pii/B9781558602472500371 3.3

[119] Kira K and Rendell L 1992 *Aaai* **2** 129 URL https://www.aaai.org/Library/AAAI/1992/aaai92-020.php 3.3

[120] Goksøyr A, Beyer J, Husøy A M, Larsen H E, Westrheim K, Wilhelmsen S and Klungsøyr J 1994 *Aquatic Toxicology* **29** 21–35 ISSN 0166445X URL https://linkinghub.elsevier.com/retrieve/pii/0166445X94900450 4.1

[121] Balk L, Hylland K, Hansson T, Berntssen M H G, Beyer J, Jonsson G, Melbye A, Grung M, Torstensen B E, Børseth J F, Skarphedinsdottir H and Klungsøyr J 2011 *Plos One* **6** e19735 URL http://dx.doi.org/10.1371/journal.pone.0019735 4.1

[122] Berg K, Puntervoll P, Klungsøyr J and Goksøyr A 2011 *Aquatic Toxicology* **105** 206–217 URL http://dx.doi.org/10.1016/j.aquatox.2011.06.010 4.1

[123] Yadetie F, Karlsen O A, Lanzén A, Berg K, Olsvik P, Hogstrand C and Goksøyr A 2013 *Aquatic Toxicology* **126** 314–325 URL http://dx.doi.org/10.1016/j.aquatox.2012.09.013 4.1

[124] Eide M, Karlsen O A, Kryvi H, Olsvik P A and Goksøyr A 2014 *Aquatic Toxicology* **153** 110–115 URL http://dx.doi.org/10.1016/j.aquatox.2013.10.027 4.1

[125] Yadetie F, Oveland E, Døskeland A, Berven F, Goksøyr A and Karlsen O A 2017 *Aquatic Toxicology* **185** 19–28 URL http://dx.doi.org/10.1016/j.aquatox.2017.01.014 4.1

[126] Halsey L G, Curran-Everett D, Vowler S L and Drummond G B 2015 *Nature Methods* **12** 179–185 URL http://dx.doi.org/10.1038/nmeth.3288 4.2

[127] Ioannidis J P A 2018 *The Journal of the American Medical Association* **319** 1429–1430 ISSN 0098-7484 URL http://doi.org/10.1001/jama.2018.1536 4.2

[128] Amrhein V, Greenland S and McShane B 2019 *Nature* **567** 305–307 ISSN 0028-0836 URL http://www.nature.com/articles/d41586-019-00857-9 4.2

[129] Kraemer H C 2019 *JAMA psychiatry* URL http://dx.doi.org/10.1001/jamapsychiatry.2019.1965 4.2

[130] Krueger J I and Heck P R 2019 *The American Statistician* **73** 122–128 ISSN 0003-1305 URL https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1470033 4.2

[131] Wasserstein R, Schirm A and Lazar N 2019 *The American statistician* **73** 1–19 ISSN 0003-1305 URL https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913 4.2

[132] Dalman M R, Deeter A, Nimishakavi G and Duan Z H 2012 *BMC Bioinformatics* **13 Suppl 2** S11 URL http://dx.doi.org/10.1186/1471-2105-13-S2-S11 4.2

[133] van IJzendoorn D G P, Szuhai K, Briaire-de Bruijn I H, Kostine M, Kuijjer M L and Bovée J V M G 2019 *PLoS Computational Biology* **15** e1006826 ISSN 1553-7358 URL http://dx.plos.org/10.1371/journal.pcbi.1006826 4.2

[134] Liñares Blanco J, Gestal M, Dorado J and Fernandez-Lozano C 2019 Differential gene expression analysis of RNA-seq data using machine learning for cancer research *Machine learning paradigms: applications of learning and analytics in intelligent systems* (*Learning and analytics in intelligent systems* vol 1) ed Tsihrintzis G A, Virvou M, Sakkopoulos E and Jain L C (Cham: Springer International Publishing) pp 27–65 ISBN 978-3-030-15627-5 URL http://link.springer.com/10.1007/978-3-030-15628-2_3 4.2, 4.3

[135] Smyth G K 2005 Limma: linear models for microarray data *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* ed Gentleman R, Carey V J, Huber W, Irizarry R A and Dudoit S (New York: Springer) pp 397–420 ISBN 0-387-25146-4 URL http://link.springer.com/10.1007/0-387-29362-0_23 4.2

[136] Liang P, Yang W, Chen X, Long C, Zheng L, Li H and Zuo Y 2020 *Molecular therapy. Nucleic acids* **20** 155–163 ISSN 21622531 URL https://linkinghub.elsevier.com/retrieve/pii/S2162253120300767 4.2, 4.3

[137] Hardee J, Ouyang Z, Zhang Y, Kundaje A, Lacroute P and Snyder M 2013 *G3 (Bethesda, Md.)* **3** 2173–2185 URL http://dx.doi.org/10.1534/g3.113.007674 4.3

[138] Di Camillo B, Sanavia T, Martini M, Jurman G, Sambo F, Barla A, Squillario M, Furlanello C, Toffolo G and Cobelli C 2012 *Plos One* **7** e32200 URL http://dx.doi.org/10.1371/journal.pone.0032200 4.3

[139] Soufan O, Kleftogiannis D, Kalnis P and Bajic V B 2015 *Plos One* **10** e0117988 URL http://dx.doi.org/10.1371/journal.pone.0117988 4.3

[140] Zhang X and Jonassen I 2019 A comparative analysis of feature selection methods for biomarker discovery in study of toxicant-treated atlantic cod (gadus morhua) liver *Nordic artificial intelligence research and development: third symposium of the norwegian AI society, NAIS 2019, trondheim, norway, may 27–28, 2019, proceedings* (*Communications in computer and information science* vol 1056) ed Bach K and Ruocco M (Cham: Springer International Publishing) pp 114–123 ISBN 978-3-030-35663-7 URL http://link.springer.com/10.1007/978-3-030-35664-4_11 4.3

[141] Zhang X and Jonassen I 2019 An ensemble feature selection framework integrating stability *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE) pp 2792–2798 ISBN 978-1-7281-1867-3 URL https://ieeexplore.ieee.org/document/8983310/ 4.3

[142] Dessì N, Pascariello E and Pes B 2013 *BioMed research international* **2013** 387673 URL http://dx.doi.org/10.1155/2013/387673 4.3

[143] Moteghaed N Y, Maghooli K, Pirhadi S and Garshasbi M 2015 *Journal of medical signals and sensors* **5** 88–96 URL https://www.ncbi.nlm.nih.gov/pubmed/26120567 5

[144] Zafeiris D, Rutella S and Ball G R 2018 *Computational and structural biotechnology journal* **16** 77–87 ISSN 20010370 URL http://linkinghub.elsevier.com/retrieve/pii/S2001037017300843 5

[145] Fakoor R, Ladhak F, Nazi A and Huber M 2013 Using deep learning to enhance cancer diagnosis and classification *Proceedings of the international conference on machine learning* vol 28 (ACM New York, USA) URL https://www.researchgate.net/profile/Rasool_Fakoor/publication/281857285_Using_deep_learning_to_enhance_cancer_diagnosis_and_classification/links/5982f029458515a60df82098/Using-deep-learning-to-enhance-cancer-diagnosis-and-classification.pdf 5

[146] Vandenberghe M E, Scott M L J, Scorer P W, Söderberg M, Balcerzak D and Barker C 2017 *Scientific Reports* **7** 45938 URL http://dx.doi.org/10.1038/srep45938 5

[147] Xiao Y, Wu J, Lin Z and Zhao X 2018 *Computer Methods and Programs in Biomedicine* **153** 1–9 URL http://dx.doi.org/10.1016/j.cmpb.2017.09.005 5

[148] Lyu B and Haque A 2018 Deep learning based tumor type classification using gene expression data *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18* (New York, New York, USA: ACM Press) pp 89–96 ISBN 9781450357944 URL http://dl.acm.org/citation.cfm?doid=3233547.3233588 5

[149] Abu-Asab M S, Chaouchi M, Alesci S, Galli S, Laassri M, Cheema A K, Atouf F, VanMeter J and Amri H 2011 *Omics : a journal of integrative biology* **15** 105–112 URL http://dx.doi.org/10.1089/omi.2010.0023 5

[150] Pisitkun T, Gandolfo M T, Das S, Knepper M A and Bagnasco S M 2012 *Proteomics. Clinical Applications* **6** 268–278 URL http://dx.doi.org/10.1002/prca.201100108 5

[151] Villoslada P and Baranzini S 2012 *Journal of Neuroimmunology* **248** 58–65 URL http://dx.doi.org/10.1016/j.jneuroim.2012.01.001 5

[152] Goldstone J V, Hamdoun A, Cole B J, Howard-Ashby M, Nebert D W, Scally M, Dean M, Epel D, Hahn M E and Stegeman J J 2006 *Developmental Biology* **300** 366–384 URL http://dx.doi.org/10.1016/j.ydbio.2006.08.066 5

[153] Zhang X, Eide M, Karlsen O A, Hanna E M, Brun M, Blaser N, Jonassen I and Goksøyr A 2019 The chemical defensome of atlantic cod (gadus morhua): how does it differ from defensome networks in other teleost species? URL https://sites.google.com/icloud.com/primo20/home 5

[154] Gibbs D L, Gralinski L, Baric R S and McWeeney S K 2014 *Frontiers in genetics* **4** 309 URL http://dx.doi.org/10.3389/fgene.2013.00309 5

[155] Ge S, Wang Y, Song M, Li X, Yu X, Wang H, Wang J, Zeng Q and Wang W 2018 *Omics : a journal of integrative biology* **22** 514–523 URL http://dx.doi.org/10.1089/omi.2018.0053 5

[156] Fan Z, Zhou Y and Ressom H W 2020 *Metabolites* **10** URL http://dx.doi.org/10.3390/metabo10040144 5

[157] Lewin H A, Robinson G E, Kress W J, Baker W J, Coddington J, Crandall K A, Durbin R, Edwards S V, Forest F, Gilbert M T P, Goldstein M M, Grigoriev I V, Hackett K J, Haussler D, Jarvis E D, Johnson W E, Patrinos A, Richards S, Castilla-Rubio J C, van Sluys M A, Soltis P S, Xu X, Yang H and Zhang G 2018 *Proceedings of the National Academy of Sciences of the United States of America* **115** 4325–4333 URL http://dx.doi.org/10.1073/pnas.1720115115 5

Appendices

Paper I

RNA-Seq analysis of transcriptome responses in Atlantic cod (*Gadus morhua*) precision-cut liver slices exposed to benzo[a]pyrene and 17α- ethynylestradiol

Fekadu Yadetiea, Xiaokang Zhang, Eileen Marie Hanna, Libe Aranguren-Abadía, Marta Eide, Nello Blaser, Morten Brun, Inge Jonassen, Anders Goksøyr, Odd André Karlsen
Aquatic Toxicology, **174/186** (2018)

Aquatic Toxicology 201 (2018) 174–186

Contents lists available at ScienceDirect

Aquatic Toxicology

journal homepage: www.elsevier.com/locate/aqtox

RNA-Seq analysis of transcriptome responses in Atlantic cod (*Gadus morhua*) precision-cut liver slices exposed to benzo[*a*]pyrene and 17α-ethynylestradiol

Fekadu Yadetie[a,*], Xiaokang Zhang[b], Eileen Marie Hanna[b], Libe Aranguren-Abadía[a], Marta Eide[a], Nello Blaser[c], Morten Brun[c], Inge Jonassen[b], Anders Goksøyr[a], Odd André Karlsen[a]

[a] *Department of Biological Sciences, University of Bergen, Bergen, Norway*
[b] *Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway*
[c] *Department of Mathematics, University of Bergen, Bergen, Norway*

ARTICLE INFO

Keywords:
RNA-Seq
Ahr
Esr1
Atlantic cod
Liver

ABSTRACT

Polycyclic aromatic hydrocarbons such as benzo[*a*]pyrene (BaP) that activate the aryl hydrocarbon receptor (Ahr) pathway, and endocrine disruptors acting through the estrogen receptor pathway are among environmental pollutants of major concern. In this work, we exposed Atlantic cod (*Gadus morhua*) precision-cut liver slices (PCLS) to BaP (10 nM and 1000 nM), ethynylestradiol (EE2) (10 nM and 1000 nM), and equimolar mixtures of BaP and EE2 (10 nM and 1000 nM) for 48 h, and performed RNA-Seq based transcriptome mapping followed by systematic bioinformatics analyses. Our gene expression analysis showed that several genes were differentially expressed in response to BaP and EE2 treatments in PCLS. Strong up-regulation of genes coding for the cytochrome P450 1a (Cyp1a) enzyme and the Ahr repressor (Ahrrb) was observed in BaP treated PCLS. EE2 treatment of liver slices strongly up-regulated genes coding for precursors of vitellogenin (Vtg) and eggshell zona pellucida (Zp) proteins. As expected, pathway enrichment and network analysis showed that the Ahr and estrogen receptor pathways are among the top affected by BaP and EE2 treatments, respectively. Interestingly, two genes coding for fibroblast growth factor 3 (Fgf3) and fibroblast growth factor 4 (Fgf4) were up-regulated by EE2 in this study. To our knowledge, the *fgf3* and *fgf4* genes have not previously been described in relation to estrogen signaling in fish liver, and these results suggest the modulation of the FGF signaling pathway by estrogens in fish. The signature expression profiles of top differentially expressed genes in response to the single compound (BaP or EE2) treatment were generally maintained in the expression responses to the equimolar binary mixtures. However, in the mixture-treated groups, BaP appeared to have anti-estrogenic effects as observed by lower number of differentially expressed putative EE2 responsive genes. Our in-depth quantitative analysis of changes in liver transcriptome in response to BaP and EE2, using PCLS tissue culture provides further mechanistic insights into effects of the compounds. Moreover, the analyses demonstrate the usefulness of PCLS in cod for omics experiments.

1. Introduction

Among environmental pollutants of particular concern are carcinogenic polycyclic aromatic hydrocarbons (PAHs), dioxins and dioxin-like polychlorinated biphenyls (PCB), as well as estrogenic endocrine disruptors. PAHs such as benzo[*a*]pyrene (BaP), dioxins and dioxin-like compounds activate the aryl hydrocarbon receptor (Ahr) pathway (Hahn, 2002), and estrogen mimicking endocrine disruptors such as ethynylestradiol (EE2), act through the estrogen receptor pathway (Goksøyr, 2006b). In fish, Cyp1a enzymes induced by exposure to pollutants such as dioxin and PAHs have long been in use as biomarkers (Goksøyr and Förlin, 1992; Stegeman and Lech, 1991). The increased synthesis of mRNA and protein levels of genes encoding vitellogenins and eggshell zona pellucida/ radiata proteins in response to estrogenic compounds has led to their use as biomarkers of endocrine disruptors in male and juvenile fish (Arukwe and Goksøyr, 2003; Sumpter and Jobling, 1995).

In recent years, the wide availability of omics techniques has

[*] Corresponding author at: Department of Biological Sciences, University of Bergen, PO Box 7803, N-5020, Bergen, Norway.
E-mail addresses: Fekadu.Yadetie@uib.no (F. Yadetie), Xiaokang.Zhang@uib.no (X. Zhang), Eileen.Hanna@uib.no (E.M. Hanna), Libe.Aranguren@uib.no (L. Aranguren-Abadía), Marta.Eide@uib.no (M. Eide), Nello.Blaser@uib.no (N. Blaser), Morten.Brun@uib.no (M. Brun), Inge.Jonassen@uib.no (I. Jonassen), Anders.Goksoyr@uib.no (A. Goksøyr), Odd.Karlsen@uib.no (O.A. Karlsen).

https://doi.org/10.1016/j.aquatox.2018.06.003
Received 31 October 2017; Received in revised form 1 June 2018; Accepted 6 June 2018
Available online 07 June 2018

opened for deeper investigations into effects of pollutants at the genome scale, promising a more complete mapping of effects (Denslow et al., 2007; Hahn, 2011; Martyniuk et al., 2011). Omics techniques are increasingly recognized as useful tools for identifying perturbed pathways and discovery of new potential biomarkers of exposure. Using omics methods, studies in many fish species have characterized gene expression in response to various environmental pollutants (such as activators of the transcription factors Ahr and estrogen receptors) in the liver of fish or primary hepatocytes (Baker et al., 2013; Colli-Dula et al., 2014; Hahn et al., 2014; Williams et al., 2013).

For many animals such as large fish species, toxicological studies using *in vivo* methods are often cumbersome, expensive and low throughput, and can be ethically challenging. Therefore, there is an increasing need for efficient *in vitro* methods, computational models and systems biology approaches that can replace or minimize the use of animal models (Krewski et al., 2010). Precision-cut liver slice (PCLS) culture is an efficient method for *in vitro* or *ex vivo* toxicological studies (Eide et al., 2014; Miller et al., 1993; Singh et al., 1996). PCLS methods can be better alternatives to primary hepatocyte cultures methods for many studies because the former method maintains intact cellular architecture similar to *in vivo* condition. It has also been shown that compared to primary hepatocyte cultures, gene expression patterns in liver slices are more similar to *in vivo* liver gene expression patterns (Boess et al., 2003). PCLS in conjunction with omics technologies can be used to generate high throughput data for a large number of chemical exposures. Such data may lead to further mechanistic studies and generate computational models employing systems biology approaches and the adverse outcome pathway (AOP) framework to facilitate chemical risk assessment (Ankley et al., 2010; Brockmeier et al., 2017).

The objective of this study is to map transcriptome changes in Atlantic cod liver tissue in response to BaP and EE2 exposure using PCLS. We used BaP and EE2 as model compounds that activate the Ahr and estrogen receptor pathways, which are among the most important pathways in environmental toxicology. BaP is a ubiquitous environmental pollutant and activator of the Ahr (Hahn, 2002). EE2 is a potent estrogen receptor-activating pharmaceutical estrogen used in contraceptive pills that is commonly detected in aquatic environments contaminated with domestic sewage (Larsson et al., 1999). The Atlantic cod is one of the most important fish species in North Atlantic fisheries that is also used in environmental monitoring programs (Balk et al., 2011; Hylland et al., 2008). Atlantic cod is a well-studied species and the availability of a sequenced and annotated genome (Star et al., 2011; Torresen et al., 2017) has facilitated use of omics approaches in toxicological studies (Bratberg et al., 2013; Karlsen et al., 2011; Yadetie et al., 2013, 2017).

2. Materials and methods

2.1. Experimental design

Fish sex was determined by dissection and inspection of the gonads. All fish were juveniles with immature gonads. A total of 8 (3 females and 5 males) juvenile cod were used for liver slicing for RNA-Seq experiment. The seven treatment groups of PCLS consist of DMSO (vehicle control), 10 nM (2.52 μg/L) BaP, 1000 nM (252.31 μg/L), BaP, 10 nM (2.96 μg/L) EE2, 1000 nM (296.40 μg/L) EE2, 10 nM Mix (BaP + EE2, 10 nM each) and 1000 nM Mix (BaP + EE2, 1000 nM each). Concentrations and exposure time used in these experiments were based on preliminary experiments (performed with different concentrations of BaP) and data from a previous study (Eide et al., 2014). Accordingly, to be able to map gene expression responses without causing significant cytotoxicity, two concentrations (low and high) were used for each compound in 48-hour exposure experiments. Slices from the same liver sample were assigned to each of the seven groups in a paired sample design and received corresponding treatments. Each group contained slices from 6 to 8 fish, which are biological replicates (n = 6–8). In total of 47 RNA samples were sequenced at a depth of approximately 50 million 75 bp paired-end reads per sample.

An additional experiment was performed to evaluate anti-estrogenic effects of BaP at different concentrations and BaP:EE2 M ratios using 4 (3 females and 1 male) juvenile Atlantic cod for liver slicing and qPCR assay. In this experiment, there were four treatment groups of liver slices that consisted of DMSO (vehicle control), 50 nM EE2, 50 nM EE2 + 1 μM BaP, and 50 nM EE2 + 10 μM BaP. Slices from each replicate liver sample were assigned to each of the four groups in a paired sample design and received the corresponding treatments.

2.2. Fish

The fish were obtained from the Institute of Marine Research (Austevoll station, Norway) and maintained at the Industrial and Aquatic Laboratory (Bergen, Norway). Juvenile Atlantic cod (*G. morhua*) approximately 1.5 years old used for the experiment were kept in 500 L tanks in 10 °C seawater with a 12 h light/l2 h dark cycle. The fish were fed with a commercial diet (Harmony Nature 500, EWOS, Bergen, Norway). The mean body weight of the fish was 498 g (standard deviation = 163 g). The fish were sexually immature, as confirmed during dissection. The fish were maintained and treated in accordance with the guidelines of the Norwegian Board of Biological Experiments with Living Animals.

2.3. Chemicals

DMSO (CAS No: 67-68-5), Benzo[*a*]pyrene (BaP) (CAS No: 50-32-8), 17α-ethynylestradiol (EE2) (CAS No: 57-63-6) and Thiazolyl Blue Tetrazolium Bromide (MTT) (CAS No: 298-93-1) were purchased from Sigma-Aldrich (Sigma-Aldrich, Oslo, Norway).

2.4. Preparation of PCLS

PCLS preparation was performed as previously described (Eide et al., 2014), with the main modification being agarose embedding of the liver tissue as described below. Briefly, the fish was killed by a blow to the head and the liver was dissected out and kept in ice-cold PCLS buffer containing NaCl (122 mM), KCl (4.8 mM), MgSO4 (1.2 mM), Na2HPO4 (11 mM) and NaHCO3 (3.7 mM), pH 8.4. Cylindrical cores (8 mm diameter) were excised from the central part of the liver and kept in the culture medium, which is Leibowitz-15 medium (Life Technologies™ Gibco®, Paisley, UK) supplemented with 10% charcoal-stripped and heat-inactivated fetal bovine serum and 1% penicillin-–streptomycin–amphotericin (10,000 U/mL potassium penicillin, 10,000 μg/mL streptomycin and 25 μg/mL amphotericin B; Sigma-Aldrich). Agarose gel embedding of the cylindrical core was performed as follows. The core was placed at the bottom of an inverted 15 mL falcon tube (with the lid on) cut at the 12 mL mark. The tube (placed on ice) was filled with 3% ultra-low gelling temperature agarose gel (CAS Number 9012-36-6, Sigma-Aldrich, Oslo, Norway) at a maximum temperature of 25 °C, to completely cover the core. After about 10 min gelling on ice, the resulting cylindrical block (about 2 cm long) was glued to the sample holder and cut into 250 μm slices using Leica vibrating blade microtome VT1200 (Leica, Wetzlar, Germany) at a speed of 0.9 mm/s and amplitude 3 mm in ice-cold PCLS buffer. The slices were kept in the culture medium at 4 °C, then pre-incubated at 10 °C for 2 h in the culture medium before exposure.

2.5. PCLS culture and exposure assays

Cod liver slices were cultured in 24-well plates (Costar, Corning, New York, USA) in 1 mL of the culture medium per slice in an incubator at 10 °C with shaking at 50 rpm. After 2 h of pre-incubation, the growth medium was replaced by medium containing either DMSO vehicle, 10 nM (2.52 μg/L) BaP, 1000 nM (252.31 μg/L) BaP, 10 nM (2.96 μg/L)

EE2, 1000 nM (296.40 μg/L) EE2, 10 nM mixture (BaP + EE2, 10 nM each) or 1000 nM mixture (BaP + EE2, 1000 nM each). The concentration of DMSO solvent in each group was 0.01%. After 48 h in culture, parallel slices were either collected for viability assay (MTT assay) or snap frozen (two slices per sample, about 20 mg) in liquid nitrogen and stored at −80 °C for RNA extraction.

2.6. PCLS viability test using MTT assay

Individual slices were rinsed in a 24-well plate (1 slice per well) containing 1 mL PBS per well, at 4 °C. The PBS was replaced by ice-cold MTT solution (dissolved in L-15 medium at 2 mg/mL) and incubated for 90 min at 10 °C with shaking at 50 rpm. Then the MTT solution was removed and the slice in each well was washed by 1 mL PBS. After removing the PBS, 1 mL DMSO was added to each well and the plate was incubated at room temperature for 20 min with shaking at 50 rpm. Absorbance was measured in 100 μL of the DMSO solution in each well, in triplicates at 590 nm with an EnSpire plate reader (Perkin Elmer).

2.7. RNA extraction

Total RNA was isolated from frozen slices (two slices pooled per treatment for each liver sample, n = 6–8 per group) using mirVana™ miRNA Isolation Kit with phenol (Cat# AM1560, Ambion, Austin, TX, USA). To remove any DNA contamination, total RNA was treated using TURBO DNase (TURBO DNA-free kit, Ambion) and further purified using RNA Clean & Concentrator-5 (Zymo Research Corp, Irvine, CA, USA) and eluted with RNase-free water. The concentration of total RNA was measured using NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). The RNA integrity was assessed using agarose gel electrophoresis. The RNA samples were submitted to the Genomics Core Facility at the University of Bergen for RNA sequencing.

2.8. RNA sequencing

RNA samples were submitted for sequencing to the Genomics Core Facility at the University of Bergen. RNA quality was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Each RNA sample (0.4 μg) was processed and sequenced using Illumina® TruSeq® Stranded mRNA Sample Preparation Kits according to Illumina TruSeq® Stranded mRNA Sample Preparation Guide (October 2013) on Illumina HiSeq 4000 (Illumina, Inc., San Diego, CA, USA). Poly(A)+ RNA was purified, fragmented and converted to first strand and second cDNAs. The second strand cDNA was amplified using PCR (15 cycles) to create the final cDNA library, which was sequenced to generate 50 million 75 bp paired-end reads per sample.

2.9. RNA-Seq read mapping and analysis of differential expression

The quality of the sequenced RNA samples was verified using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). In total, we had 47 high-quality samples, which were aligned to the published CDS of Atlantic cod (ftp://ftp.ensembl.org/pub/release-90/fasta/gadus_morhua/cds/Gadus_morhua.gad) using HISAT2 v2.1.0 (Kim et al., 2015). Counts were generated from the alignments using SAMtools v1.4.1 (Li et al., 2009). The RNA-Seq reads mapped to about 22,100 unique Atlantic cod CDSs in Ensembl (http://www.ensembl.org/) of which, about 19,000 genes were obtained after filtering out genes with zero counts in all samples.

Differential expression analysis was performed using edgeR v3.18.1 (McCarthy et al., 2012) between control group and each treated group using “TMM” normalization method and paired test, after removing the genes with zero counts across all samples of the two compared groups. Differentially expressed genes were defined by p-value < 0.05 after adjustment using the Benjamini-Hochberg multiple testing correction. Only genes with average counts per million/cpm > 1 in at least one of the control or treated groups and with fold changes > 1.5 (for up-regulated) or < 0.67 (for down regulated) were included in the final list of differentially expressed genes. A Snakemake (Koster and Rahmann, 2012) workflow was implemented and is available at https://github.com/zhxiaokang/RNA-Seq-Snakemake. This allows the whole analysis to be reproduced and the workflow to be applied in similar studies. The RNA-Seq data has been deposited in GEO (accession: GSE106968).

2.10. Pathway analysis

For pathway analysis, human and zebrafish (*Danio rerio*) orthologs of the Atlantic cod genes were obtained using the BioMart tool in Ensembl (https://www.ensembl.org). Pathway enrichment and network analyses were performed in DAVID (Database for Annotation, Visualization and Integrated Discovery) (Huang et al., 2009), Enrichr (Kuleshov et al., 2016), STRING and STITCH protein–protein and chemical-protein interaction network analysis tools (Szklarczyk et al., 2015, 2016) and MetaCore (https://portal.genego.com/). For the interaction network construction in STRING and STITCH databases, zebrafish orthologs of the differentially expressed genes were used with the low-confidence interaction option. In MetaCore (which does not allow pathway analysis for fish), the human orthologs of the differentially expressed genes were used. Visualization and analysis of networks generated using STRING and annotated with expression data was performed using Cytoscape (Cline et al., 2007). For all pathway enrichment analyses, a false discovery rate (FDR) < 0.05 was considered as significant enrichment. For Gene Set Enrichment Analysis (GSEA), Atlantic cod genes reliably quantified by RNA-Seq (with average normalized CPM > 1 in either control or treated group) in PCLS (treated with 1000 nM BaP and DMSO) that could be mapped to human orthologs in Ensembl (about 13,000 genes) were used (Subramanian et al., 2005). GSEA software and gene sets KEGG and Hallmarks in the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) at Broad Institute (http://www.broadinstitute.org/gsea/index.jsp) were used. GSEA was performed with 1000 permutations of phenotypes with default settings. Gene sets enriched with FDR < 0.25 were considered significant as recommended (Subramanian et al., 2005).

2.11. Hierarchical clustering

Hierarchical clustering analysis (Euclidian metric, complete linkage) was performed in Qlucore Omics Explorer, using log2-transformed expression ratio values (library-size normalized cpm in treated/DMSO control). Genes differentially expressed in Multi Group Comparison paired-test analysis (Qlucore Omics Explorer, q-value < 0.05) were used for hierarchical clustering. Paired test was performed to account for the fact that slices from the same fish were used in each of the seven treatment groups. Possible sex-related effects were also evaluated using Qlucore Omics Explorer. There were no significant differences in gene expression responses to EE2 or BaP between the liver slices from male and female juvenile fish (data not shown).

2.12. Quantitative polymerase chain reaction (qPCR)

For qPCR, cDNA was prepared from 1.0 μg of each total RNA sample in 20 μL reactions using iScript cDNA Synthesis Kit (Bio-Rad, Hercules, CA, USA). The reverse transcription reaction mix containing 1 μL of the reverse transcriptase (RT) provided in the kit was pre-incubated at 25 °C for 5 min (priming), at 46 °C for 30 min (reverse transcription) and at 95 °C for 1 min (RT inactivation). The cDNA was diluted (1:20) and 5 μL was used in a 20 μL PCR reaction mix containing 0.5 μM of each of the forward and reverse primers. PCR was performed on BioRad CFX96 real-time PCR detection system (Bio-Rad Laboratories). The reaction conditions were as follows: an initial incubation at 95 °C for 10 min, followed by 40 cycles of denaturation at 95 °C for 10 s, annealing and elongation at 60 °C for 30 s. Negative controls with no reverse transcriptase enzyme were run for each primer pair. For each primer pair, serial dilutions of cDNA prepared from pooled RNA samples from the same experiment were used to construct a standard curve to estimate amplification efficiency. qPCR assays were performed in triplicates. The *actb* gene was used as a reference for normalization. To check primer specificity, melting curve analysis was performed and PCR products were also analyzed by agarose gel electrophoresis. Expression levels were compared between controls and treated samples using the ΔΔCq method (Schmittgen and Livak, 2008) and further statistical analysis was performed using the log2-transformed fold-changes in expression.

2.13. Statistical analysis

Statistical analyses of MTT data and log2-transformed fold-changes (treated/control) from qPCR were performed with GraphPad Prism Software version 7 (GraphPad Prism, La Jolla, CA, USA). Data were checked for normality using Kolmogorov–Smirnov test. Repeated measures one-way ANOVA with Greenhouse–Geisser correction followed by Dunnett's test (for normally distributed data), and Friedman test followed by Dunn's test (for non-normally distributed data) were used. Data are reported as mean ± standard deviation (SD), and $p < 0.05$ was considered significant. Further statistical analysis and hierarchical clustering analysis of RNA-Seq data was performed using Qlucore Omics Explorer version 3.2 (Qlucore AB, Lund, Sweden). Slices from the same individual fish were used in each of the seven treatment groups and paired test was performed in Qlucore Omics Explorer. Possible sex-related effects on gene expression were also evaluated using Qlucore Omics Explorer.

3. Results

3.1. Genes differentially expressed in BaP treated PCLS and enriched pathways

The read counts generated from RNA-Seq mapped to about 19,000 unique Atlantic cod genes. Based on differential expression analysis using edgeR (McCarthy et al., 2012), 13 unique genes (13 genes by the 1000 nM and 1 gene by the 10 nM BaP concentrations) were significantly differentially expressed (paired test, FDR < 0.05) (Table 1, Supplementary Table S1). Out of these 13 genes, all except two (*vtg1-1* and *znf366*) were up-regulated. The Ahr pathway and other related or interacting pathways were among the top pathways significantly enriched (Table 2). The top enriched gene ontology (GO) biological process (BP) terms and pathways are those related to xenobiotic and steroid metabolism, attributed to Cyp1a, Cyp1b1 and Ahrr encoding genes up-regulated here (Table 2A–C). Other pathways affected by BaP treatment include immune response (attributed to genes such as *il1b* and *ccr9a*), steroid metabolism, chemical carcinogenesis and oxidative stress (Tables 1, 2A–C). A heatmap generated at a more relaxed cutoff (unadjusted p < 0.01) shows more differentially expressed genes (Supplementary Fig. S1). For example, the heatmap shows three genes encoding vitellogenin are slightly down regulated in BaP-treated slices. To supplement the pathway analysis performed with the small number of genes differentially expressed by BaP, GSEA (with input of all quantified genes) was also performed. GSEA showed significant enrichment of gene sets and pathways related to those identified for the 13 differentially expressed genes (Table 2A–C, Supplementary Fig. S2A-F, Supplementary Table S8A and B). The heatmap from GSEA analysis shows expression profiles of the top ranked genes (Supplementary Fig. S3), and as expected, many of the differentially regulated genes in Table 1 are seen in the top ranks.

Table 1
Genes differentially expressed in BaP treated PCLS.

Cod gene ID	Zebrafish symbol	fold-change	FDR
ENSGMOG00000000855	*avpr1ab*	53.0	1.43E-23
ENSGMOG00000009114	*ahrrb*	17.1	5.15E-22
ENSGMOG00000000318	*cyp1a*	59.2	5.01E-16
ENSGMOG00000006842	*cyp1b1*	39.8	5.44E-08
ENSGMOG00000020520	*ccr9a*	4.0	3.31E-05
ENSGMOG00000001034	*tll1*	4.3	4.97E-04
ENSGMOG00000001139	*slc43a3b*	2.6	1.01E-03
ENSGMOG00000002589	*ctss1*	3.5	7.86E-03
ENSGMOG00000005676	*dcstamp*	2.8	1.08E-02
ENSGMOG00000016347	*si:dkey-4c23.3 (vtg1-1)*	−3.1	1.21E-02
ENSGMOG00000000345	*il1b*	2.7	1.21E-02
ENSGMOG00000001247	*tpte*	2.4	3.74E-02
ENSGMOG00000007538	*znf366*	−2.2	4.47E-02

3.2. Genes differentially expressed in EE2 treated PCLS and enriched pathways

Fig. 1 shows a Venn diagram of genes differentially expressed in EE2-treated PCLS. Out of about 19,000 unique Atlantic cod genes, the total number of differentially expressed genes (FDR < 0.05 and minimum 1.5 fold-change) by EE2 was 79. The 10 nM and 1000 nM EE2 concentrations resulted in 17 and 72 differentially expressed genes, respectively, with 11 genes differentially expressed by both concentrations (Fig. 1). Genes differentially expressed in PCLS treated with the different concentrations of EE2 and BaP and EE2 mixtures are listed in Supplementary Tables S2-4. All except 7 EE2 modulated genes were up-regulated (Supplementary Tables S2 and S3). Many well-known oogenesis related estrogen responsive genes such as those encoding estrogen receptor alpha (*esr1*), vitellogenins (*vtg*), and genes coding for eggshell *zona pellicuda* (ZP)/ *zona radiata* proteins were up-regulated in the PCLS treated with EE2 and mixture of EE2 and BaP (Supplementary Tables S2-5, Figs. 1–3). For pathway analysis in DAVID and STRING, zebrafish or human orthologs of the differentially expressed genes were used. For pathway analysis in MetaCore, the human orthologs of the differentially expressed genes were used. Functional enrichment analysis in DAVID using zebrafish orthologs of the differentially expressed genes showed that the top significant pathways are related to vitellogenins, eggshell zona pellucida proteins and cellular response to estrogen (Table 3). Protein domains for Zp and Vtg were also highly enriched (Table 3). Many *vtg* and *zp* genes were up-regulated in the EE2 or mixture groups (Table 3, Supplementary Tables S2-5, Figs. 1–3). Pathway analysis of their human orthologs of the EE2-responsive cod genes in MetaCore also showed enrichment of many pathways, particularly pathways and processes that appear to be related to growth promoting effects of estrogens in some mammalian tissue (Supplementary Tables S6 and S7). Significantly enriched (FDR < 0.05) Metacore Map Folders include estrogen signaling, reproductive tissue neoplasms, mitogenic signaling, as well as tissue remodeling and wound repair (Supplementary Table S6).

For network analysis in STRING, the zebrafish orthologs of EE2-modulated genes were used because of its richer annotation compared to cod. Network analysis in STRING and visualization using Cytoscape revealed subnetworks of the core estrogen receptor pathway genes consisting of *vtgs* and *zps*, within the *esr1* hub (Fig. 4). Expression levels (log2cpm) and fold-changes of expression of the genes were also indicated by color gradients in the network (Fig. 4).

3.3. Comparison of genes differentially expressed in BaP, EE2, and mixture treated PCLS

The total number of significantly differentially expressed genes in the low and high concentration mixture group was 31 and 36, respectively, and 13 genes were differentially expressed in both low and high mixture groups (Supplementary Tables S4 and S5). The differentialy regulated genes in the BaP-, EE2- and mixture-treated PCLS were

Table 2
Top pathways enriched in genes differentially expressed in BaP treated PCLS.[a]

A) KEGG Term	Adjusted P-value	Genes
Tryptophan metabolism_Homo sapiens_hsa00380	0.007	CYP1A1; CYP1B1
Ovarian steroidogenesis_Homo sapiens_hsa04913	0.007	CYP1A1; CYP1B1
Steroid hormone biosynthesis_Homo sapiens_hsa00140	0.007	CYP1A1; CYP1B1
Metabolism of xenobiotics by cytochrome P450_Homo sapiens_hsa00980	0.008	CYP1A1; CYP1B1
Chemical carcinogenesis_Homo sapiens_hsa05204	0.008	CYP1A1; CYP1B1
Cytokine-cytokine receptor interaction_Homo sapiens_hsa04060	0.065	IL1B; CCR9

B) Wikipathway Term	Adjusted P-value	Genes
Aryl Hydrocarbon Receptor Pathway_Homo sapiens_WP2873	4.3E-07	IL1B; CYP1A1; AHRR; CYP1B1
Aryl Hydrocarbon Receptor_Homo sapiens_WP2586	5.0E-05	CYP1A1; AHRR; CYP1B1
Benzo(a)pyrene metabolism_Homo sapiens_WP696	1.8E-04	CYP1A1; CYP1B1
Estrogen metabolism_Mus musculus_WP1264	2.0E-04	CYP1A1; CYP1B1
Estrogen Receptor Pathway_Homo sapiens_WP2881	2.3E-04	CYP1A1; CYP1B1
Estrogen metabolism_Homo sapiens_WP697	3.8E-04	CYP1A1; CYP1B1
Tamoxifen metabolism_Homo sapiens_WP691	4.4E-04	CYP1A1; CYP1B1
Melatonin metabolism and effects_Homo sapiens_WP3298	7.0E-04	CYP1A1; CYP1B1
Oxidation by Cytochrome P450_Mus musculus_WP1274	1.2E-03	CYP1A1; CYP1B1
Tryptophan metabolism_Mus musculus_WP79	1.3E-03	CYP1A1; CYP1B1
Tryptophan metabolism_Homo sapiens_WP465	1.4E-03	CYP1A1; CYP1B1
Oxidation by Cytochrome P450_Homo sapiens_WP43	2.3E-03	CYP1A1; CYP1B1
Peptide GPCRs_Mus musculus_WP234	2.4E-03	CCR9; AVPR1A
Peptide GPCRs_Homo sapiens_WP24	2.8E-03	CCR9; AVPR1A
Metapathway biotransformation_Mus musculus_WP1251	8.2E-03	CYP1A1; CYP1B1
GPCRs, Class A Rhodopsin-like_Mus musculus_WP189	1.2E-02	CCR9; AVPR1A
Metapathway biotransformation_Homo sapiens_WP702	1.2E-02	CYP1A1; CYP1B1
Non-odorant GPCRs_Mus musculus_WP1396	2.1E-02	CCR9; AVPR1A
GPCRs, Class A Rhodopsin-like_Homo sapiens_WP455	2.1E-02	CCR9; AVPR1A
Oxidative Stress_Homo sapiens_WP408	3.1E-02	CYP1A1

C) GO BP Term	Adjusted P-value	Genes
cellular response to organic cyclic compound (GO:0071407)	2.6E-06	IL1B; CYP1A1; CYP1B1
omega-hydroxylase P450 pathway (GO:0097267)	4.7E-04	CYP1A1; CYP1B1
epoxygenase P450 pathway (GO:0019373)	1.8E-03	CYP1A1; CYP1B1
positive regulation of vascular endothelial growth factor production (GO:0010575)	1.8E-03	IL1B; CYP1B1
steroid metabolic process (GO:0008202)	2.1E-03	CYP1A1; CYP1B1
positive regulation of angiogenesis (GO:0045766)	1.3E-02	IL1B; CYP1B1
xenobiotic metabolic process (GO:0006805)	1.3E-02	AHRR; CYP1B1
oxidation-reduction process (GO:0055114)	1.8E-02	CYP1A1; CYP1B1
positive regulation of granulocyte macrophage colony-stimulating factor production (GO:0032725)	1.8E-02	IL1B
cellular response to organic substance (GO:0071310)	1.8E-02	IL1B
positive regulation of heterotypic cell-cell adhesion (GO:0034116)	1.8E-02	IL1B
retinal metabolic process (GO:0042574)	1.8E-02	CYP1B1
positive regulation of JAK-STAT cascade (GO:0046427)	1.8E-02	CYP1B1
positive regulation of protein export from nucleus (GO:0046827)	1.8E-02	IL1B
positive regulation of histone acetylation (GO:0035066)	1.8E-02	IL1B
estrogen metabolic process (GO:0008210)	1.8E-02	CYP1B1
nitric oxide biosynthetic process (GO:0006809)	1.8E-02	CYP1B1
positive regulation of monocyte differentiation (GO:0045657)	1.8E-02	DCSTAMP
positive regulation of histone phosphorylation (GO:0033129)	1.8E-02	IL1B
positive regulation of bone resorption (GO:0045780)	1.8E-02	DCSTAMP

[a] For B and C, only the top 20 terms significantly enriched (Adjusted P-value < 0.05) are shown.

compared in a Venn diagram (Fig. 2). More genes were differentially expressed in EE2 treated PCLS compared to BaP. Only one gene (si:dkey-4c23.3/*vtg1-1*) was differentially expressed in all three groups. Interestingly, the *vtg1-1* gene was down regulated by BaP and up-regulated in the EE2 and Mix groups (Supplementary Tables S1-5). Only 6 of 13 BaP-regulated genes are common with the mixture group, whereas a higher number of EE2-responsive genes (35 of 77) were found in the mixture group. Closer inspection of Fig. 2 showed that the 14 genes exclusively differentially expressed in the mixture group are likely EE2-responsive since they show similar changes (but not significant; FDR > 0.05) in the EE2 groups and not in the BaP-treated groups (data not shown). Thus the number of differentially regulated EE2-responsive genes was 77 and 49 (36% reduction) in the EE2-treated and mixture group, respectively. The corresponding reduction is about 44% when we compare differentially expressed EE2-responsive genes only in the high concentration (1000 nM) BaP, EE2, and mixture group (Supplementary Tables, S1, S3 and S5).

A two-way hierarchical clustering analysis of the samples and the differentially expressed genes shows two major clusters largely corresponding to BaP-responsive and EE2-responsive genes (Fig. 3). The top up-regulated genes in BaP alone treated group (*cyp1a*, *cyp1b1*, *ahrrb*, *tll1* and *ccr9a*) show similar changes in the mixture treated group

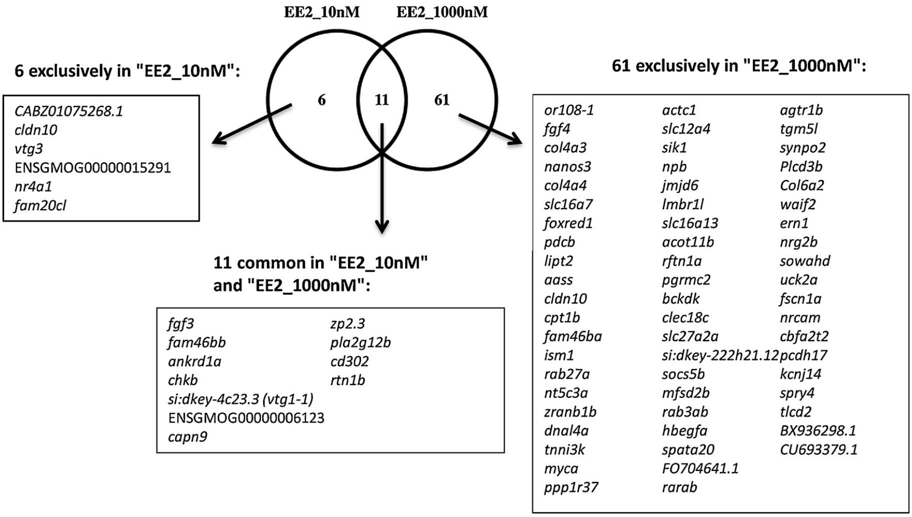


Fig. 1. Venn diagram showing genes differentially expressed in 10 nM and 1000 nM EE2 treated PCLS. Zebrafish or human orthologs of the differentially expressed cod genes (edgeR, FDR $<$ 0.05 and minimum fold-change 1.5) are presented in Venn diagram. Gene with no zebrafish or human orthologs are listed with their Ensembl cod gene ids. In case of more than one zebrafish ortholog, only one is listed here.

(Table 1, Fig. 3, A). The other genes affected by BaP in Table 1 also show changes in the same direction (not shown), except *il1b* (up-regulated by BaP and no change in mixture groups) and *si:dkey-4c23.3/vtg1-1* (down-regulated by BaP and up-regulated in mixture groups). The top differentially expressed vitellogenesis-related genes were indicated as core estrogen receptor pathway genes (*zp*, *vtg* and *esr1* genes) with similar expression profiles in both EE2-treated and mixture groups (Fig. 3, B).

To see possible interaction of the two chemicals (BaP and EE2) and the genes differentially expressed in liver slices exposed to the binary mixture, we used the STITCH database that enables combined view of protein-chemical and protein-protein interactions (Fig. 5). Enrichment analysis in STITCH showed that Interpro protein domains for Vtg, Zp and Cyp1 were significantly enriched in the genes differentially expressed in the mixture groups (FDR $<$ 0.05) (not shown). Distinct estrogen receptor pathway and the Ahr pathway modules connected by the *esr1* hub can be seen in the network (Fig. 5).

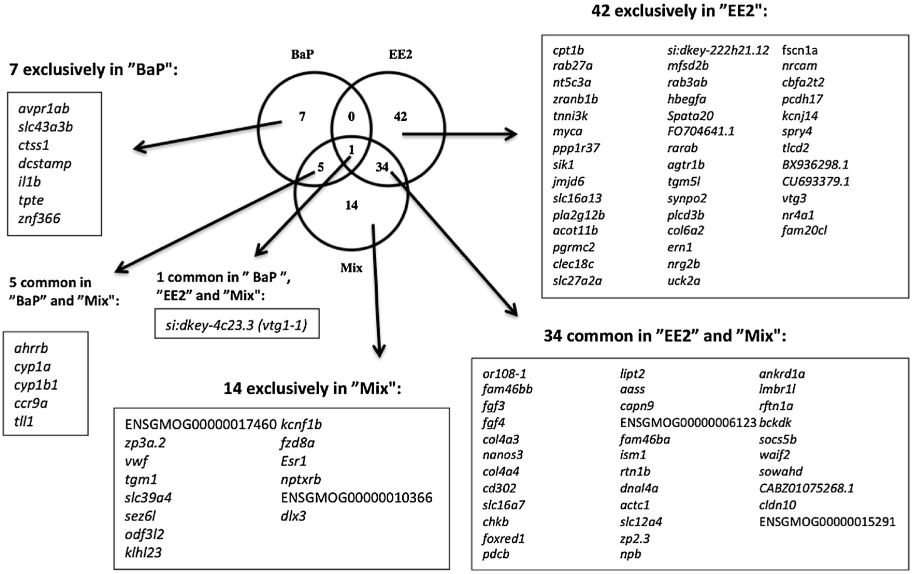


Fig. 2. Venn diagram comparing differentialy expressed genes in BaP, EE2 and Mix (mixture) treatments. Genes significantly responding to treatments of PCLS by BaP (10 nM and 1000 nM), EE2 (10 nM and 1000 nM) and mixture (10 nM and 1000 nM, each of BaP and EE2) (edgeR, FDR $<$ 0.05) were compared. The number of differentially expressed genes indicated for each of BaP, EE2 and Mix represents combined unique genes from both low and high concentration treatments.

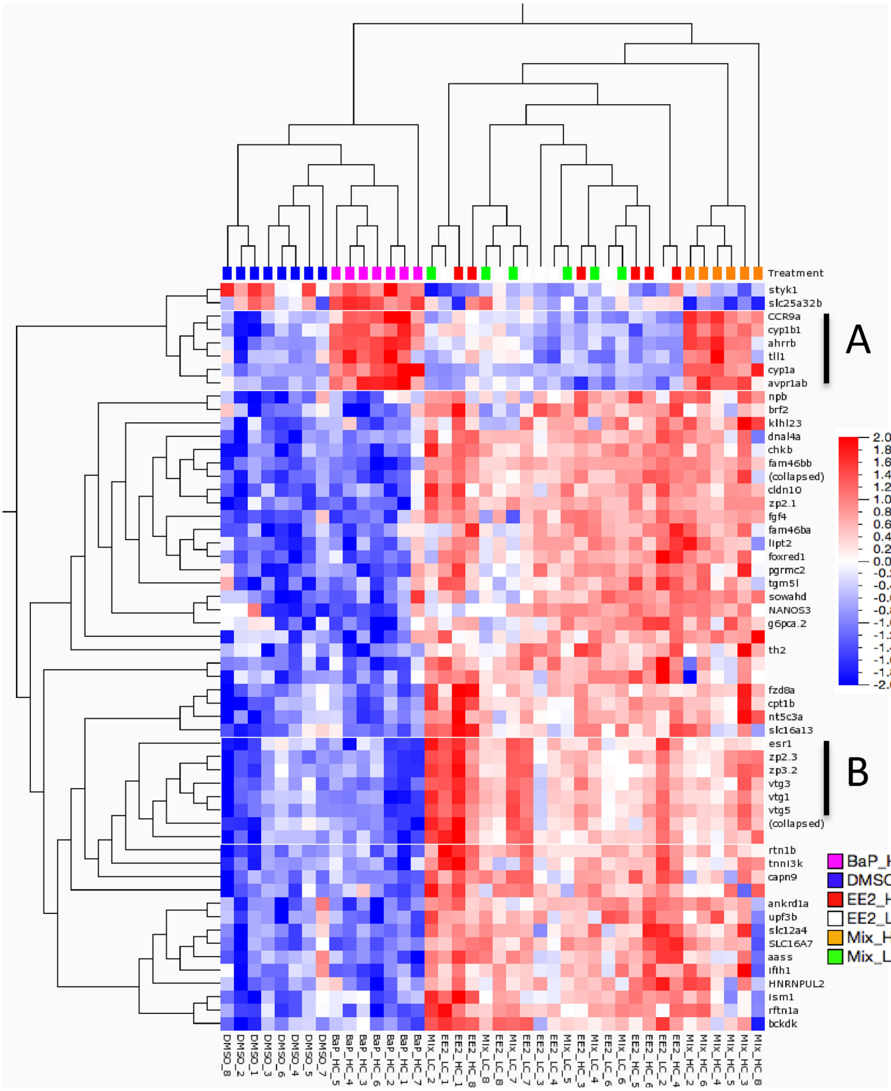


Fig. 3. Hierarchical clustering analysis of differentially expressed genes. Genes differentially expressed in DMSO control, low (10 nM) concentrations (LC) of BaP, EE2 and mixture (BaP and EE2), high (1000 nM) concentrations (HC) of BaP, EE2 and mixture treated groups. Analysis was performed based on log2-transformed ratio (treated/control) values of differentially expressed genes (q-value < 0.05) in Multi Group Comparison (Qlucore Omics Explorer). Rows represent genes and columns represent samples. The vertical lines indicate the top Ahr pathway (**A**) and core estrogen receptor pathway (**B**) genes. The group with low (10 nM) BaP concentrations with little effect on gene expression was removed for clarity. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. PCLS viability

MTT assay was performed with slices from each treatment and no statistically significant change in viability was observed in any of the treatment groups compared to the DMSO control (Fig. 6).

3.5. qPCR assay

Expression levels of eight genes from the list of genes differentially expressed in liver slices treated with BaP (*cyp1a* and *ahrrb*) (Table 1) and EE2 (*esr1*, *vtg1*, *zp2l1*, *fam46bb*, *fgf3* and *fgf4*) (Supplementary Tables S2 and S3) were also assessed using qPCR to confirm the RNA-Seq data (Fig. 7A–H). All the genes tested were significantly up-regulated in a similar manner to the RNA-Seq results, but in general lower fold-changes were obtained using qPCR compared to RNA-Seq (Fig. 7A–H, Table 1, Supplementary Tables S1-5). The amplicon sizes and sequences of primers used in the qPCR assays are given in Supplementary Table S9.

The anti-estrogenic effect of BaP was further investigated in additional PCLS exposure experiments using qPCR analysis of estrogen

Table 3
Functional annotation clusters for genes differentially expressed in EE2 treated PCLS.[a].

Cluster 1	Enrichment Score: 10.0	
Category	Term	FDR
INTERPRO	IPR015258:Vitellinogen, beta-sheet shell	5.26E-11
INTERPRO	IPR015255:Vitellinogen, open beta-sheet	8.62E-10
INTERPRO	IPR015817:Vitellinogen, open beta-sheet, subdomain 1	8.62E-10
INTERPRO	IPR015818:Vitellinogen, open beta-sheet, subdomain 2	8.62E-10
INTERPRO	IPR001747:Lipid transport protein, N-terminal	1.72E-09
INTERPRO	IPR011030:Vitellinogen, superhelical	1.72E-09
INTERPRO	IPR015816:Vitellinogen, beta-sheet N-terminal	1.72E-09
INTERPRO	IPR015819:Lipid transport protein, beta-sheet shell	1.72E-09
GOTERM_BP_FAT	GO:0006869 ~ lipid transport	0.00369
GOTERM_BP_FAT	GO:0010876 ~ lipid localization	0.00539
GOTERM_BP_FAT	GO:0033036 ~ macromolecule localization	42.92
Cluster 2	Enrichment Score: 7.1	
Category	Term	FDR
INTERPRO	IPR000519:P-type trefoil	5.59E-05
INTERPRO	IPR001507:Zona pellucida domain	6.62E-05
INTERPRO	IPR017977:Zona pellucida domain, conserved site	2.29E-04
Cluster 3	Enrichment Score: 3.2	
Category	Term	FDR
GOTERM_BP_FAT	GO:0071391 ~ cellular response to estrogen stimulus	0.264
GOTERM_BP_FAT	GO:0043627 ~ response to estrogen	0.449
GOTERM_BP_FAT	GO:0070887 ~ cellular response to chemical stimulus	6.378

[a] Annotation clusters were obtained using gene ontology biological process (GOTERM_BP_FAT) and protein domains (INTERPRO) enrichment analysis in DAVID.

receptor pathway (*vtg1* and *esr1*) (Supplementary Fig. S4A and B) and Ahr pathway (*cyp1a*) genes (Supplementary Fig. S4C). Exposure experiments with lower EE2 concentration (50 nM) and two different BaP concentrations (1 and 10 µM) showed stronger anti-estrogenic effects of BaP at the higher BaP: EE2 M ratios (Supplementary Fig. S4A and B). In the mixture treatments, higher induction of *cyp1a* is negatively associated with induction of *vtg1* and *esr1* genes (Supplementary Fig. S4A-C).

4. Discussion

4.1. Analysis of differential expression in PCLS exposed to BaP and EE2

In this study, we performed an in-depth mapping of transcriptome responses in Atlantic cod liver tissue treated with BaP and EE2 (singly and in combination) *in vitro* using PCLS. The number of genes reliably quantified here in Atlantic cod liver slices that could be mapped to human orthologs (about 13,000) is much higher compared to the corresponding number of genes quantified using microarrays in the cod liver (about 8300) (Yadetie et al., 2014). RNA-Seq appears to have enabled detection and quantification of low-expressed genes such as *fgf3* and *fgf4*. Our results are in agreement with the differential expression of several genes encoding proteins in the Ahr pathways by Ahr ligands (Jenny et al., 2009; Karchner et al., 2002), and oogenesis-related estrogen receptor pathway gene products by estrogens reported in previous studies (Arukwe and Goksøyr, 2003; Tyler and Sumpter, 1996). qPCR analysis of selected genes confirmed differential expression of the genes by RNA-Seq, but the RNA-Seq method resulted in higher fold-changes compared to qPCR, consistent with higher precision and dynamic range of the former (Wang et al., 2009). The results also demonstrate the usefulness of PCLS culture previously developed for Atlantic cod (Eide et al., 2014) in high throughput toxicogenomics studies. An important advantage of the slice culture method is the ability to perform large number of exposure experiments with only a limited number of animals in a paired sample experimental design, as demonstrated here. Although high variability was observed in gene expression responses in the slices from different fish, the paired sample test considerably improved the analysis of differential expression. Possible sex differences in gene expression responses to BaP or EE2 were also tested, and there were no significant differences in liver gene expression responses to BaP or EE2 between slices from the sexually immature male and female juvenile fish (data not shown). Juvenile male and female fish respond similarly in liver gene expression in responses to estrogenic compounds in many fish species such as salmonids (Shilling and Williams, 2000; Yadetie and Male, 2002).

4.2. Effects of BaP

The differentially expressed genes in the BaP exposed liver slices include *cyp1a*, *cyp1b* and *ahrrb*. Four genes encoding Cyp1 subfamily members (Cyp1a, Cyp1b, Cyp1c and Cyp1d) are present in Atlantic cod and other fish genomes (Goldstone et al., 2010; Goldstone and Stegeman, 2008; Karlsen et al., 2012). The *cyp1a*, *cyp1b* and *cyp1c* genes were previously shown to be inducible by Ahr ligands (Jonsson et al., 2007). In the present experiment, the *cyp1a* gene was the most abundantly and highly differentially expressed (fold-change 59.2), followed by *cyp1b* (fold-change 39.8) (Supplementary Table S1). Two other genes *cyp1c1* and *cyp1c2* were not significantly up-regulated (FDR > 0.05) (not shown). Although *cyp1d* is present in the cod genome (https://www.ensembl.org), its transcript was not detectable by the RNA-Seq experiment here. However, it has been shown that *cyp1d* is not inducible by Ahr ligands in zebrafish (Goldstone et al., 2009). Atlantic cod Cyp1a has long been widely used as a biomarker (Goksøyr, 1995; Goksøyr and Förlin, 1992). One of the up-regulated genes, the gene encoding aryl hydrocarbon receptor repressor (*ahrr*), is also known to be induced by Ahr ligands (Jenny et al., 2009; Karchner et al., 2002). Both *ahrra* and *ahrrb* genes are inducible by 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in an Ahr-dependent manner in zebrafish (Jenny et al., 2009). Of the differentially expressed genes by BaP treatment, most were up-regulated, and only two genes (*znf366* and *vtg1-1*) were down regulated (Table 1). Notably, ZNF366 protein is a repressor of ERα (Lopez-Garcia et al., 2006), and down regulation of the *znf366* gene might be related to the cross-talk between the aryl hydrocarbon receptor and estrogen receptor pathways (Ohtake et al.,

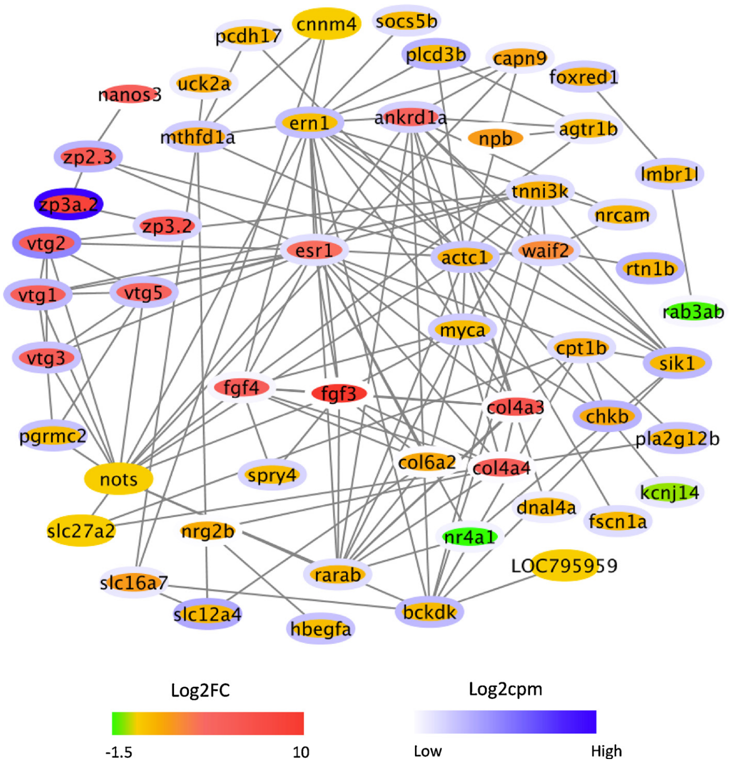


Fig. 4. Network of genes differentially expressed in EE2 treated PCLS. The network was generated using STRING (https://string-db.org/) and imported in Cytoscape for visualization. Color range green to deep red (node fill) and white to deep cyan (node border) indicates log2-transformed fold-changes (log2FC) and log2-transformed counts per million reads (log2cpm), respectively. Disconnected nodes were removed from the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2003). Genes encoding the inflammatory mediators IL1-β and CCR9a were also up-regulated, suggesting effects on immune modulation. IL1-β is has been recently shown to be up-regulated *via* AHR (Jacob et al., 2017). Thus, these results are consistent with reports that the aryl hydrocarbon receptor also has a role in immune modulation (Esser and Rannug, 2015). Among pathways significantly enriched in BaP-treated liver slices include chemical carcinogenesis and oxidative stress (Table 2A and B). This is likely related to possible metabolic activation of BaP by Cyp1 enzymes induced in the liver slices. Metabolic activation of BaP by Cyp1 enzymes can lead to formation of carcinogenic metabolites such as benzo[*a*]pyrene-7,8-diol-9,10-epoxide (BPDE) (Nebert and Dalton, 2006; Stansbury et al., 1994). Many of the Ahr pathway genes described here have also been shown to be affected by BaP exposure and RNA-Seq analysis of zebrafish embryos and mammalian a cell line (Fang et al., 2015; van Delft et al., 2012).

4.3. Effects of EE2

The major pathways affected by EE2 treatment in PCLS in this experiment are related to vitellogenesis and zonagenesis processes taking place in fish liver during oocyte maturation. The Atlantic cod has at least 3 vitellogenin genes that map to orthologs in zebrafish, which has at least 8 genes (https://www.ensembl.org). Most fish species have at least three vitellogenin genes, with lineage-specific gene duplications in some species including zebrafish (Finn and Kristoffersen, 2007; Finn et al., 2009). There are also more *zp* and *zp-like* genes in the zebrafish genome (12 paralogs) compared to cod (9 paralogs) (https://www.ensembl.org). *Vitellogenin* mRNA and proteins are known to be highly up-regulated in the liver of oviparous fish in response to estrogens (Arukwe and Goksøyr, 2003; Tyler and Sumpter, 1996). Genes encoding eggshell proteins and their protein products are also induced in the liver in response to estrogens in many fish species such as Atlantic cod and Atlantic salmon (Arukwe and Goksøyr, 2003; Arukwe et al., 1997; Hyllner et al., 1991; Oppen-Berntsen et al., 1999, 1992). The gene encoding estrogen receptor alpha (*esr1*), which was up-regulated here and appears as a hub in the core EE2-responsive module of the network (Fig. 4) is also known to be induced by estradiol and environmental estrogens in fish liver and primary hepatocytes (Marlatt et al., 2008; Pakdel et al., 1991; Yadetie et al., 1999). In addition to the core estrogen receptor pathway genes, many genes in these modules (*e.g. nots*, *fam20c* and *rtn1b*) are previously shown to be modulated in fish liver by estradiol (Levi et al., 2009; Uren Webster et al., 2015). Many genes related to vitellogenesis and zonagenesis that were differentially expressed in this study have also been shown to be differentially regulated in microarray analysis of primary hepatocyte cultures treated with EE2 (Hultman et al., 2015). The gene expression responses in EE2 treated cod liver slices correlated well with *in vivo* responses. For example, many of the top up-regulated genes constituting the core estrogen-responsive genes in fish liver (Feswick et al., 2017) were also differentially expressed in EE2-treated cod liver slices, which should facilitate the use of multiple biomarkers of exposure to estrogens in Atlantic cod.

MetaCore pathways and processes related to estrogenic effects in mammals were also enriched (Supplementary Tables S6 and S7). Many

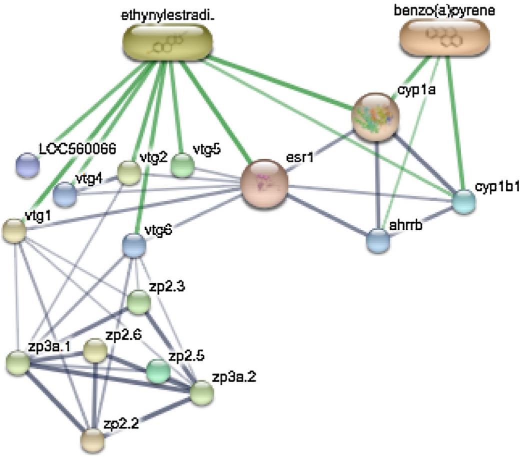


Fig. 5. Network of genes differentially expressed in mixture (BaP and EE2) treated PCLS. The network was generated using STITCH v5.0 (http://stitch.embl.de). The network represents "confidence view" with stronger associations represented by thicker lines. Protein-protein interactions are shown in grey, and chemical-protein interactions in green. Disconnected nodes were removed from the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

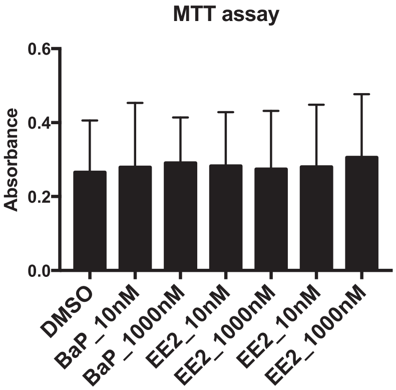


Fig. 6. MTT assay of PCLS. Cell viability test by MTT colorimetric assay after treatment with DMSO, low (10 nM) and high (1000 nM) concentrations of BaP, EE2 or mixture (BaP and EE2) for 48 h. Liver slices from each of the seven treatment groups (n = 8) were used in the MTT assay and statistical tests were performed using repeated measure ANOVA. No significant differences were found between controls and any of the treatments in viability. Data points represent mean ± SD of absorbance values.

of the gene products involved in MetaCore pathways and processes (Supplementary Tables S6 and S7) are shown to interact with *esr1* in the network of differentially expressed genes (Fig. 4). Among these, for example the *fgf3* and *fgf4* genes were highly up-regulated in EE2-exposed PCLS. This appears to be consistent with modulation of *fgf* genes by estrogen receptors in some mammalian cells (Smith et al., 2002). The mammalian FGF family growth factors are involved in a range of processes that include development, proliferation, survival and differentiation (Ornitz and Itoh, 2015). Estrogens are known to induce growth-promoting genes in normal and cancerous cells and tissues of reproductive organs in mammals (Hewitt et al., 2003; Platet et al., 2004). Indeed, increase in liver weight is one of the effects of estrogens related to vitellogenesis in fish (Emmerson et al., 1979; Skillman et al., 2006). Further, estrogens have been reported to be liver tumor promoters (Nunez et al., 1989; Cooke and Hinton, 1999). Thus, growth-promoting effects of estrogens appear to be conserved in fish. However, to our knowledge, up-regulation of the *fgf3* and *fgf4* genes by estrogens has not been shown before and suggests estrogen dependent activation of the FGF signaling pathway in fish liver. Further experiments are warranted to investigate the implications of the possible activation of the FGF signaling by estrogens.

4.4. Effects of BaP and EE2 mixtures

From our analysis BaP appears to have resulted in a lower number of differentially expressed putative estrogen-regulated genes in the liver slices exposed to mixture of and BaP and EE2, suggesting ant-estrogenic effects. In addition, two genes *vtg1-1* (an estrogen responsive gene) and *znf366*, coding for an estrogen receptor co-repressor (Lopez-Garcia et al., 2006) were down regulated by BaP treatment (Table 1). Down regulation of *znf366* and an estrogen responsive *vtg-1* gene seems to be consistent with known anti-estrogenic effects of Ahr ligands (see below). This was observed in samples that were not treated by EE2 and it suggests inhibition of background estrogen receptor activity by BaP treatment, which may have environmentally relevant implications since Ahr activating pollutants may interfere with the functioning of the endocrine system of fish.

The anti-estrogenic effect of BaP was further confirmed using qPCR analysis of estrogen receptor (*vtg1* and *esr1*) and Ahr pathway genes (*cyp1a*) after PCLS treatment with higher BaP: EE2 M ratios. The anti-estrogenic effect of BaP observed here is consistent with known cross-talk between the Ahr and estrogen receptor pathways (Goksøyr, 2006a; Safe and Wormke, 2003; Wormke et al., 2000). Mechanisms involved in anti-estrogenic effects of ligand-activated Ahr include inhibition of binding of the estrogen receptor to its response elements in the promoters of estrogen-responsive genes, through direct binding of Ahr to estrogen receptor and Ahr-mediated degradation of estrogen receptor (Matthews and Gustafsson, 2006; Ohtake et al., 2011, 2003). Many studies have documented anti-estrogenic effect of Ahr ligands in fish (Bemanian et al., 2004; Navas and Segner, 2000). In Atlantic salmon primary hepatocytes, TCDD showed anti-estrogenic effects by reducing expression of a gene encoding vitellogenin in primary hepatocytes treated with 17β-estradiol (E2) through inhibition of estrogen receptor binding to estrogen responsive elements (ERE) (Bemanian et al., 2004). Treatment of zebrafish by TCDD resulted in decreased serum E2 levels and vitellogenin synthesis (Heiden et al., 2006). More recently, TCDD has been shown to inhibit vitellogenin induction in EE2-treated zebrafish in an Ahr-dependent manner (Bugel et al., 2013).

Network analysis of BaP and EE2 mixture-affected genes in STITCH database suggested interactions of the estrogen receptor and Ahr pathways, that reflects known cross-talk between the two pathways (Goksøyr, 2006a; Safe and Wormke, 2003). For example, the interaction network shows interaction between the up-regulated *ahrrb* and *esr1* genes (Fig. 5). AHRR has been shown to have anti-estrogenic effects by binding to estrogen receptor 1 in human breast cancer MCF-7 cells (Kanno et al., 2008). Anti-estrogen effects may also be attributed to increased metabolism of estradiol by Cyp1 enzymes induced by Ahr ligands (James, 2011; Scornaienchi et al., 2010), as suggested in the network by the interactions of the Cyp1a and Cyp1b1 enzymes and EE2 (Fig. 5).

5. Conclusion

Combining high throughput RNA-Seq analysis and PCLS *ex vivo* tissue culture in Atlantic cod, we have mapped the transcriptome responses to BaP and EE2 treatments, offering further mechanistic insights into effects of these chemicals. BaP and EE2 treatments resulted in differential expression of several genes in the Ahr and estrogen

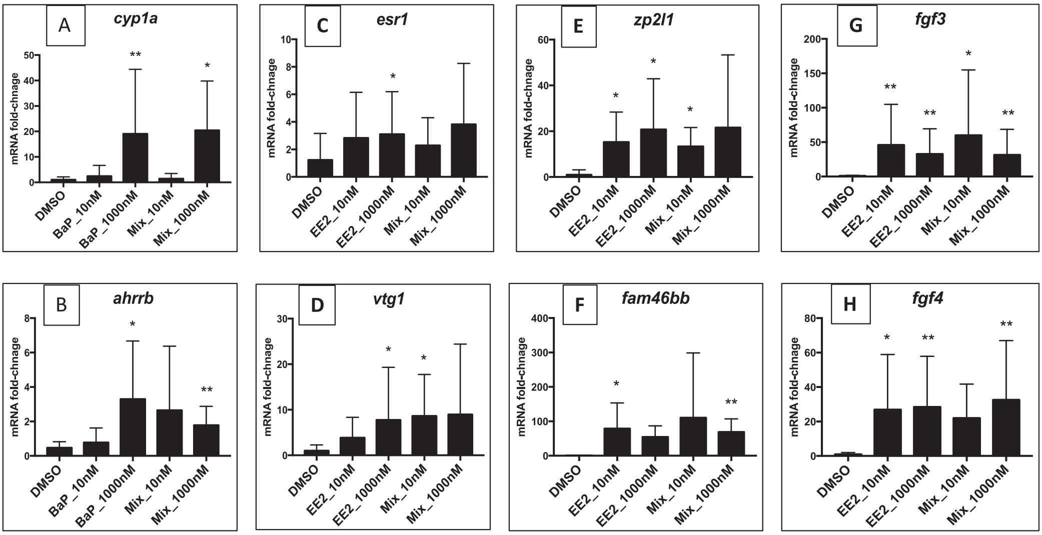


Fig. 7. qPCR analysis of genes differentially expressed in PCLS treated with BaP, EE2 and mixtures of BaP and EE2. The plots show fold-changes in mRNA levels for *cyp1a* (**A**) and *ahrrb* (**B**) genes in slices treated with BaP (10 nM and 1000 nM) and equimolar mixtures (Mix) of BaP and EE2 (10 nM and 1000 nM). The fold-changes in mRNA levels of the genes *esr1* (**C**), *vtg1* (**D**), *zp2l1* (**E**), *fam46bb* (**F**), *fgf3* (**G**), and *fgf4* (**H**) are from slices treated with EE2 (10 nM and 1000 nM) and equimolar Mix (10 nM and 1000 nM). RNA samples from liver slices of 7–8 fish used for RNA-Seq experiment were used (n = 7–8 per treatment group). **p* < 0.05; ***p* < 0.01; ****p* < 0.001 (one-way ANOVA and Dunnett's multiple comparison test for all except **C**, **F** and **H** for which Friedman test followed by Dunn's post hoc test was performed). Data points present mean ± SD.

receptor pathways, respectively. The up-regulation of *fgf3* and *fgf4* genes in EE2-treated liver slices suggests new mechanistic insights into effects of estrogens in fish liver. Finally, the functional significance of the changes detected at transcript levels needs to be further studied.

Conflict of interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Research Council of Norway to the Centre for Digital Life Norway (DLN) project dCod 1.0 [grant number 248840], and the iCod 2.0 project [grant number 244564].

Acknowledgements

We thank the Genomics Core Facility (GCF) at the Department of Clinical Science, University of Bergen, Norway for services rendered towards sequencing and analysis of genomic data, and Bergen Research Foundation (BFS) for partial funding of the sequencing project through grants to GCF. The results of this contribution were presented at PRIMO19 in Ehime University, Matsuyama, Japan.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aquatox.2018.06.003.

References

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29, 730–741.

Arukwe, A., Goksøyr, A., 2003. Eggshell and egg yolk proteins in fish: hepatic proteins for the next generation: oogenetic, population, and evolutionary implications of endocrine disruption. Comp. Hepatol. 2, 4.

Arukwe, A., Knudsen, F.R., Goksøyr, A., 1997. Fish zona radiata (eggshell) protein: a sensitive biomarker for environmental estrogens. Environ. Health Perspect. 105, 418–422.

Baker, M.E., Vidal-Dorsch, D.E., Ribecco, C., Sprague, L.J., Angert, M., Lekmine, N., Ludka, C., Martella, A., Ricciardelli, E., Bay, S.M., Gully, J.R., Kelley, K.M., Schlenk, D., Carnevali, O., Sasik, R., Hardiman, G., 2013. Molecular analysis of endocrine disruption in hornyhead turbot at wastewater outfalls in southern california using a second generation multi-species microarray. Plos One 8, e75553.

Balk, L., Hylland, K., Hansson, T., Berntssen, M.H.G., Beyer, J., Jonsson, G., Melbye, A., Grung, M., Torstensen, B.E., Borseth, J.F., Skarphedinsdottir, H., Klungsøyr, J., 2011. Biomarkers in natural fish populations indicate adverse biological effects of offshore oil production. PloS One 6.

Bemanian, V., Male, R., Goksøyr, A., 2004. The aryl hydrocarbon receptor-mediated disruption of vitellogenin synthesis in the fish liver: Cross-talk between AHR- and ERalpha-signalling pathways. Comp. Hepatol. 3, 2.

Boess, F., Kamber, M., Romer, S., Gasser, R., Muller, D., Albertini, S., Suter, L., 2003. Gene expression in two hepatic cell lines, cultured primary hepatocytes, and liver slices compared to the in vivo liver gene expression in rats: possible implications for toxicogenomics use of in vitro systems. Toxicol. Sci. 73, 386–402.

Bratberg, M., Olsvik, P.A., Edvardsen, R.B., Brekken, H.K., Vadla, R., Meier, S., 2013. Effects of oil pollution and persistent organic pollutants (POPs) on glycerophospholipids in liver and brain of male Atlantic cod (Gadus morhua). Chemosphere 90, 2157–2171.

Brockmeier, E.K., Hodges, G., Hutchinson, T.H., Butler, E., Hecker, M., Tollefsen, K.E., Garcia-Reyero, N., Kille, P., Becker, D., Chipman, K., Colbourne, J., Collette, T.W., Cossins, A., Cronin, M., Graystock, P., Gutsell, S., Knapen, D., Katsiadaki, I., Lange, A., Marshall, S., Owen, S.F., Perkins, E.J., Plaistow, S., Schroeder, A., Taylor, D., Viant, M., Ankley, G., Falciani, F., 2017. The role of Omics in the application of adverse outcome pathways for chemical risk assessment. Toxicol. Sci. 158, 252–262.

Bugel, S.M., White, L.A., Cooper, K.R., 2013. Inhibition of vitellogenin gene induction by 2,3,7,8-tetrachlorodibenzo-p-dioxin is mediated by aryl hydrocarbon receptor 2 (AHR2) in zebrafish (Danio rerio). Aquat. Toxicol. 126, 1–8.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A.R., Vailaya, A., Wang, P.L., Adler, A., Conklin, B.R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G.J., Ideker, T., Bader, G.D., 2007. Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. 2, 2366–2382.

Colli-Dula, R.C., Martyniuk, C.J., Kroll, K.J., Prucha, M.S., Kozuch, M., Barber, D.S., Denslow, N.D., 2014. Dietary exposure of 17-alpha ethinylestradiol modulates physiological endpoints and gene signaling pathways in female largemouth bass

(Micropterus salmoides). Aquat. Toxicol. 156, 148–160.
Cooke, J.B., Hinton, D.E., 1999. Promotion by 17β-estradiol and β-hexachlorocyclohexane of hepatocellular tumors in medaka, Oryzias latipes. Aquat. Toxicol. 45, 127–145.
Denslow, N.D., Garcia-Reyero, N., Barber, D.S., 2007. Fish' n' chips: the use of microarrays for aquatic toxicology. Mol. Biosyst. 3, 172–177.
Eide, M., Karlsen, O.A., Kryvi, H., Olsvik, P.A., Goksøyr, A., 2014. Precision-cut liver slices of Atlantic cod (Gadus morhua): an in vitro system for studying the effects of environmental contaminants. Aquat. Toxicol. 153, 110–115.
Emmerson, J., Korsgaard, B., Petersen, I., 1979. Dose response kinetics of serum vitellogenin, liver DNA, RNA, protein and lipid after induction by estradiol-17 beta in male flounders (Platichthys flesus L.). Comp. Biochem. Physiol. B 63, 1–6.
Esser, C., Rannug, A., 2015. The aryl hydrocarbon receptor in barrier organ physiology, immunology, and toxicology. Pharmacol. Rev. 67, 259–279.
Fang, X., Corrales, J., Thornton, C., Clerk, T., Scheffler, B.E., Willett, K.L., 2015. Transcriptomic changes in zebrafish embryos and larvae following benzo[a]pyrene exposure. Toxicol. Sci. 146, 395–411.
Feswick, A., Munkittrick, K.R., Martyniuk, C.J., 2017. Estrogen-responsive gene networks in the teleost liver: what are the key molecular indicators? Environ. Toxicol. Pharmacol. 56, 366–374.
Finn, R.N., Kristoffersen, B.A., 2007. Vertebrate vitellogenin gene duplication in relation to the "3R hypothesis": correlation to the pelagic egg and the oceanic radiation of teleosts. Plos One 2, e169.
Finn, R.N., Kolarevic, J., Kongshaug, H., Nilsen, F., 2009. Evolution and differential expression of a vertebrate vitellogenin gene cluster. BMC Evol. Biol. 9, 2.
Goksøyr, A., 1995. Use of cytochrome P450 1A (CYP1A) in fish as a biomarker of aquatic pollution. Arch. Toxicol. Suppl. 17, 80–95.
Goksøyr, A., 2006a. Endocrine disruptors in the marine environment: mechanisms of toxicity and their influence on reproductive processes in fish. J. Toxicol. Environ. Health-Part A-Curr. Issues 69, 175–184.
Goksøyr, A., 2006b. Endocrine disruptors in the marine environment: mechanisms of toxicity and their influence on reproductive processes in fish. J. Toxicol. Environ. Health A 69, 175–184.
Goksøyr, A., Förlin, L., 1992. The cytochrome P450 system in fish, aquatic toxicology, and environmental monitoring. Aquat. Toxicol. 22, 287–311.
Goldstone, J.V., Stegeman, J.J., 2008. Gene structure of the novel cytochrome P4501D1 genes in stickleback (Gasterosteus aculeatus) and medaka (Oryzias latipes). Mar. Environ. Res. 66, 19–20.
Goldstone, J.V., Jonsson, M.E., Behrendt, L., Woodin, B.R., Jenny, M.J., Nelson, D.R., Stegeman, J.J., 2009. Cytochrome P450 1D1: a novel CYP1A-related gene that is not transcriptionally activated by PCB126 or TCDD. Arch. Biochem. Biophys. 482, 7–16.
Goldstone, J.V., McArthur, A.G., Kubota, A., Zanette, J., Parente, T., Jonsson, M.E., Nelson, D.R., Stegeman, J.J., 2010. Identification and developmental expression of the full complement of Cytochrome P450 genes in Zebrafish. BMC Genomics 11, 643.
Hahn, M.E., 2002. Aryl hydrocarbon receptors: diversity and evolution. Chem. Biol. Interact. 141, 131–160.
Hahn, M.E., 2011. Mechanistic research in aquatic toxicology: perspectives and future directions. Aquat. Toxicol. 105, 67–71.
Hahn, M.E., McArthur, A.G., Karchner, S.I., Franks, D.G., Jenny, M.J., Timme-Laragy, A.R., Stegeman, J.J., Woodin, B.R., Cipriano, M.J., Linney, E., 2014. The transcriptional response to oxidative stress during vertebrate development: effects of tert-butylhydroquinone and 2,3,7,8-tetrachlorodibenzo-p-dioxin. Plos One 9, e113158.
Heiden, T.K., Carvan 3rd, M.J., Hutz, R.J., 2006. Inhibition of follicular development, vitellogenesis, and serum 17beta-estradiol concentrations in zebrafish following chronic, sublethal dietary exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin. Toxicol. Sci. 90, 490–499.
Hewitt, S.C., Deroo, B.J., Hansen, K., Collins, J., Grissom, S., Afshari, C.A., Korach, K.S., 2003. Estrogen receptor-dependent genomic responses in the uterus mirror the biphasic physiological response to estrogen. Mol. Endocrinol. 17, 2070–2083.
Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57.
Hultman, M.T., Song, Y., Tollefsen, K.E., 2015. 17alpha-ethinylestradiol (EE2) effect on global gene expression in primary rainbow trout (Oncorhynchus mykiss) hepatocytes. Aquat. Toxicol. 169, 90–104.
Hylland, K., Tollefsen, K.E., Ruus, A., Jonsson, G., Sundt, R.C., Sanni, S., Roe Utvik, T.I., Johnsen, S., Nilssen, I., Pinturier, L., Balk, L., Barsiene, J., Marigomez, I., Feist, S.W., Borseth, J.F., 2008. Water column monitoring near oil installations in the North Sea 2001-2004. Mar. Pollut. Bull. 56, 414–429.
Hyllner, S.J., Oppenberntsen, D.O., Helvik, J.V., Walther, B.T., Haux, C., 1991. Oestradiol-17 beta induces the major vitelline envelope proteins in both sexes in teleosts. J. Endocrinol. 131, 229–236.
Jacob, A., Tomkiewicz-Raulet, C., Jamet, C., Bendayan, R., Massicot, F., Coumoul, X., Decleves, X., 2017. Aryl hydrocarbon receptor upregulates IL-1beta expression in hCMEC/D3 human cerebral microvascular endothelial cells after TCDD exposure. Toxicol. In Vitro 41, 200–204.
James, M.O., 2011. Steroid catabolism in marine and freshwater fish. J. Steroid Biochem. Mol. Biol. 127, 167–175.
Jenny, M.J., Karchner, S.I., Franks, D.G., Woodin, B.R., Stegeman, J.J., Hahn, M.E., 2009. Distinct roles of two zebrafish AHR repressors (AHRRa and AHRRb) in embryonic development and regulating the response to 2,3,7,8-tetrachlorodibenzo-p-dioxin. Toxicol. Sci. 110, 426–441.
Jonsson, M.E., Orrego, R., Woodin, B.R., Goldstone, J.V., Stegeman, J.J., 2007. Basal and 3,3',4,4',5-pentachlorobiphenyl-induced expression of cytochrome P450 1A, 1B and 1C genes in zebrafish. Toxicol. Appl. Pharmacol. 221, 29–41.
Kanno, Y., Takane, Y., Takizawa, Y., Inouye, Y., 2008. Suppressive effect of aryl hydrocarbon receptor repressor on transcriptional activity of estrogen receptor alpha by protein-protein interaction in stably and transiently expressing cell lines. Mol. Cell. Endocrinol. 291, 87–94.
Karchner, S.I., Franks, D.G., Powell, W.H., Hahn, M.E., 2002. Regulatory interactions among three members of the vertebrate aryl hydrocarbon receptor family: AHR repressor, AHR1, and AHR2. J. Biol. Chem. 277, 6949–6959.
Karlsen, O.A., Bjørneklett, S., Berg, K., Brattas, M., Bohne-Kjersem, A., Grøsvik, B.E., Goksøyr, A., 2011. Integrative environmental genomics of Cod (Gadus morhua): the proteomics approach. J. Toxicol. Environ. Health-Part A-Curr. Issues 74, 494–507.
Karlsen, O.A., Puntervoll, P., Goksøyr, A., 2012. Mass spectrometric analyses of microsomal cytochrome P450 isozymes isolated from beta-naphthoflavone-treated Atlantic cod (Gadus morhua) liver reveal insights into the cod CYPome. Aquat. Toxicol. 108, 2–10.
Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360.
Koster, J., Rahmann, S., 2012. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522.
Krewski, D., Acosta Jr., D., Andersen, M., Anderson, H., Bailar 3rd, J.C., Boekelheide, K., Brent, R., Charnley, G., Cheung, V.G., Green Jr., S., Kelsey, K.T., Kerkvliet, N.I., Li, A.A., McCray, L., Meyer, O., Patterson, R.D., Pennie, W., Scala, R.A., Solomon, G.M., Stephens, M., Yager, J., Zeise, L., 2010. Toxicity testing in the 21st century: a vision and a strategy. J. Toxicol. Environ. Health B Crit. Rev. 13, 51–138.
Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Ma'ayan, A., 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44, W90–97.
Larsson, D.G.J., Adolfsson-Erici, M., Parkkonen, J., Pettersson, M., Berg, A.H., Olsson, P.E., Förlin, L., 1999. Ethinyloestradiol - an undesired fish contraceptive? Aquat. Toxicol. 45, 91–97.
Levi, L., Pekarski, I., Gutman, E., Fortina, P., Hyslop, T., Biran, J., Levavi-Sivan, B., Lubzens, E., 2009. Revealing genes associated with vitellogenesis in the liver of the zebrafish (Danio rerio) by transcriptome profiling. BMC Genomics 10, 141.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740.
Lopez-Garcia, J., Periyasamy, M., Thomas, R.S., Christian, M., Leao, M., Jat, P., Kindle, K.B., Heery, D.M., Parker, M.G., Buluwela, L., Kamalati, T., Ali, S., 2006. ZNF366 is an estrogen receptor corepressor that acts through CtBP and histone deacetylases. Nucleic Acids Res. 34, 6126–6136.
Marlatt, V.L., Martyniuk, C.J., Zhang, D., Xiong, H., Watt, J., Xia, X., Moon, T., Trudeau, V.L., 2008. Auto-regulation of estrogen receptor subtypes and gene expression profiling of 17beta-estradiol action in the neuroendocrine axis of male goldfish. Mol. Cell. Endocrinol. 283, 38–48.
Martyniuk, C.J., Griffitt, R.J., Denslow, N.D., 2011. Omics in aquatic toxicology: not just another microarray. Environ. Toxicol. Chem. 30, 263–264.
Matthews, J., Gustafsson, J.A., 2006. Estrogen receptor and aryl hydrocarbon receptor signaling pathways. Nucl. Recept Signal. 4, e016.
McCarthy, D.J., Chen, Y., Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40, 4288–4297.
Miller, M.G., Beyer, J., Hall, G.L., deGraffenried, L.A., Adams, P.E., 1993. Predictive value of liver slices for metabolism and toxicity in vivo: use of acetaminophen as a model hepatotoxicant. Toxicol. Appl. Pharmacol. 122, 108–116.
Navas, J.M., Segner, H., 2000. Antiestrogenicity of beta-naphthoflavone and PAHs in cultured rainbow trout hepatocytes: evidence for a role of the arylhydrocarbon receptor. Aquat. Toxicol. 51, 79–92.
Nebert, D.W., Dalton, T.P., 2006. The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. Nat. Rev.: Cancer 6, 947–960.
Nunez, O., Hendricks, J.D., Arbogast, D.N., Fong, A.T., Lee, B.C., Bailey, G.S., 1989. Promotion of aflatoxin B1 hepatocarcinogenesis in rainbow trout by 17β-estradiol. Aquat. Toxicol. 15, 289–302.
Ohtake, F., Takeyama, K., Matsumoto, T., Kitagawa, H., Yamamoto, Y., Nohara, K., Tohyama, C., Krust, A., Mimura, J., Chambon, P., Yanagisawa, J., Fujii-Kuriyama, Y., Kato, S., 2003. Modulation of oestrogen receptor signalling by association with the activated dioxin receptor. Nature 423, 545–550.
Ohtake, F., Fujii-Kuriyama, Y., Kawajiri, K., Kato, S., 2011. Cross-talk of dioxin and estrogen receptor signals through the ubiquitin system. J. Steroid Biochem. Mol. Biol. 127, 102–107.
Oppen-Berntsen, D.O., Hyllner, S.J., Haux, C., Helvik, J.V., Walther, B.T., 1992. Eggshell zona radiata-proteins from cod (Gadus morhua): extra-ovarian origin and induction by estradiol-17 beta. Int. J. Dev. Biol 36, 247–254.
Oppen-Berntsen, D.O., Arukwe, A., Yadetie, F., Lorens, J.B., Male, R., 1999. Salmon eggshell protein expression: a marker for environmental estrogens. Mar. Biotechnol. 1, 252–260.
Ornitz, D.M., Itoh, N., 2015. The Fibroblast Growth Factor signaling pathway. Wiley Interdisc. Rev.-Dev. Biol. 4, 215–266.
Pakdel, F., Feon, S., Le Gac, F., Le Menn, F., Valotaire, Y., 1991. In vivo estrogen induction of hepatic estrogen receptor mRNA and correlation with vitellogenin mRNA in rainbow trout. Mol. Cell. Endocrinol. 75, 205–212.
Platet, N., Cathiard, A.M., Gleizes, M., Garcia, M., 2004. Estrogens and their receptors in breast cancer progression: a dual role in cancer proliferation and invasion. Crit. Rev. Oncol. Hematol. 51, 55–67.
Safe, S., Wormke, M., 2003. Inhibitory aryl hydrocarbon receptor-estrogen receptor alpha cross-talk and mechanisms of action. Chem. Res. Toxicol. 16, 807–816.

Schmittgen, T.D., Livak, K.J., 2008. Analyzing real-time PCR data by the comparative C(T) method. Nat. Protoc. 3, 1101–1108.

Scornaienchi, M.L., Thornton, C., Willett, K.L., Wilson, J.Y., 2010. Cytochrome P450-mediated 17beta-estradiol metabolism in zebrafish (Danio rerio). J. Endocrinol. 206, 317–325.

Shilling, A.D., Williams, D.E., 2000. Determining relative estrogenicity by quantifying vitellogenin induction in rainbow trout liver slices. Toxicol. Appl. Pharmacol. 164, 330–335.

Singh, Y., Cooke, J.B., Hinton, D.E., Miller, M.G., 1996. Trout liver slices for metabolism and toxicity studies. Drug. Metab. Dispos. 24, 7–14.

Skillman, A.D., Nagler, J.J., Hook, S.E., Small, J.A., Schultz, I.R., 2006. Dynamics of 17alpha-ethynylestradiol exposure in rainbow trout (Oncorhynchus mykiss): absorption, tissue distribution, and hepatic gene expression pattern. Environ. Toxicol. Chem. 25, 2997–3005.

Smith, P., Rhodes, N.P., Ke, Y., Foster, C.S., 2002. Upregulation of estrogen and androgen receptors modulate expression of FGF-2 and FGF-7 in human, cultured, prostatic stromal cells exposed to high concentrations of estradiol. Prostate Cancer Prostatic Dis. 5, 105–110.

Stansbury, K.H., Flesher, J.W., Gupta, R.C., 1994. Mechanism of aralkyl-DNA adduct formation from benzo[a]pyrene in vivo. Chem. Res. Toxicol. 7, 254–259.

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., Wetten, O.F., Lanzen, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, O., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R.B., Tina, K.G., Espelund, M., Nepal, C., Previti, C., Karlsen, B.O., Moum, T., Skage, M., Berg, P.R., Gjoen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S.D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S.W., Stenseth, N.C., Jakobsen, K.S., 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477, 207–210.

Stegeman, J.J., Lech, J.J., 1991. Cytochrome P-450 monooxygenase systems in aquatic species: carcinogen metabolism and biomarkers for carcinogen and pollutant exposure. Environ. Health Perspect. 90, 101–109.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Sumpter, J.P., Jobling, S., 1995. Vitellogenesis as a biomarker for estrogenic contamination of the aquatic environment. Environ. Health Perspect. 103 (Suppl. 7), 173–178.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J., von Mering, C., 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–D452.

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., Kuhn, M., 2016. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 44, D380–384.

Torresen, O.K., Star, B., Jentoft, S., Reinar, W.B., Grove, H., Miller, J.R., Walenz, B.P., Knight, J., Ekholm, J.M., Peluso, P., Edvardsen, R.B., Tooming-Klunderud, A., Skage, M., Lien, S., Jakobsen, K.S., Nederbragt, A.J., 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics 18, 95.

Tyler, C.R., Sumpter, J.P., 1996. Oocyte growth and development in teleosts. Rev. Fish. Biol. Fish. 6, 287–318.

Uren Webster, T.M., Shears, J.A., Moore, K., Santos, E.M., 2015. Identification of conserved hepatic transcriptomic responses to 17beta-estradiol using high-throughput sequencing in brown trout. Physiol. Genomics 47, 420–431.

van Delft, J., Gaj, S., Lienhard, M., Albrecht, M.W., Kirpiy, A., Brauers, K., Claessen, S., Lizarraga, D., Lehrach, H., Herwig, R., Kleinjans, J., 2012. RNA-Seq provides new insights in the transcriptome responses induced by the carcinogen benzo[a]pyrene. Toxicol. Sci. 130, 427–439.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-e: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Williams, T.D., Diab, A.M., Gubbins, M., Collins, C., Matejusova, I., Kerr, R., Chipman, J.K., Kuiper, R., Vethaak, A.D., George, S.G., 2013. Transcriptomic responses of European flounder (Platichthys flesus) liver to a brominated flame retardant mixture. Aquat. Toxicol. 142-143, 45–52.

Wormke, M., Stoner, M., Saville, B., Safe, S., 2000. Crosstalk between estrogen receptor alpha and the aryl hydrocarbon receptor in breast cancer cells involves unidirectional activation of proteasomes. FEBS Lett. 478, 109–112.

Yadetie, F., Male, R., 2002. Effects of 4-nonylphenol on gene expression of pituitary hormones in juvenile Atlantic Salmon (Salmo salar). Aquat. Toxicol. 58, 113–129.

Yadetie, F., Arukwe, A., Goksøyr, A., Male, R., 1999. Induction of hepatic estrogen receptor in juvenile Atlantic salmon in vivo by the environmental estrogen, 4-nonylphenol. Sci. Total Environ. 233, 201–210.

Yadetie, F., Karlsen, O.A., Lanzen, A., Berg, K., Olsvik, P., Hogstrand, C., Goksøyr, A., 2013. Global transcriptome analysis of Atlantic cod (Gadus morhua) liver after in vivo methylmercury exposure suggests effects on energy metabolism pathways. Aquat. Toxicol. 126, 314–325.

Yadetie, F., Karlsen, O.A., Eide, M., Hogstrand, C., Goksøyr, A., 2014. Liver transcriptome analysis of Atlantic cod (Gadus morhua) exposed to PCB 153 indicates effects on cell cycle regulation and lipid metabolism. BMC Genomics 15.

Yadetie, F., Oveland, E., Doskeland, A., Berven, F., Goksøyr, A., Karlsen, O.A., 2017. Quantitative proteomics analysis reveals perturbation of lipid metabolic pathways in the liver of Atlantic cod (Gadus morhua) treated with PCB 153. Aquat. Toxicol. 185, 19–28.

Supplemental materials

MIQE checklist for authors, reviewers and editors.
Figure S1. Heatmap of expression profiles of top differentially expressed genes in BaP treated liver slices.
Figure S2. Gene set enrichment analysis (GSEA) of genes quantified by RNA-Seq in PCLS treated with BaP (1 μM) and DMSO.
Figure S3. Heatmap fromgene set enrichment analysis (GSEA) of genes quantified by RNA-Seq in PCLS treated with BaP (1 μM) and DMSO.
Figure S4. qPCR analysis of genes differentially expressed in PCLS treated with BaP and mixtures of BaP and EE2.
Table S1. Genes differentially expressed in BaP (10 and 1000 nM) treated PCLS.
Table S2. Genes differentially expressed in 10 nM EE2 treated PCLS.
Table S3. Genes differentially expressed in 1000 nM EE2 treated PCLS.
Table S4. Genes differentially expressed in 10 nM Mix (BaP + EE2, 10nM each) treated PCLS.
Table S5. Genes differentially expressed in 1000 nM Mix (BaP + EE2, 10nM each) treated PCLS.
Table S6. Enriched MetaCore map folders for genes differentially expressed in EE2 treated PCLS.
Table S7. Enriched MetaCore gene ontology (GO) biological process (BP) for genes differentially expressed in EE2 treated PCLS.
Table S8. The top gene set enrichment results for BaP treated liver slices.
Table S9. qPCR primer sequences, amplicon lengths and PCR efficiencies.

<u>MIQE checklist for authors, reviewers and editors.</u>

All essential information (E) must be submitted with the manuscript. Desirable information (D) should be submitted if possible.

If using primers obtained from RTPrimerDB, information on qPCR target, oligonucleotides, protocols and validation is available from that source.

ITEM TO CHECK	IMPORTANCE	Comments
EXPERIMENTAL DESIGN		
Definition of experimental and control groups	**E**	Given in the manuscript (experimental design)
Number within each group	**E**	Given in the manuscript (experimental design), also given in Figure captions.
Assay carried out by core lab or investigator's lab?	D	Investigator’s lab
Acknowledgement of authors' contributions	D	
SAMPLE		
Description	**E**	Precision-cut liver slices
Volume/mass of sample processed	D	Two 8 mm diameter (250 µm thick), about 20 mg wt slices per sample
Microdissection or macrodissection	**E**	Macrodissection (precision-cut –iver slices)
Processing procedure	**E**	
If frozen - how and how quickly?	**E**	Snap frozen in liquid nitrogen
If fixed - with what, how quickly?	**E**	Not fixed
Sample storage conditions and duration (especially for FFPE samples)	**E**	Tissue and RNA stored at -80 degree celcius. cDNAs and primers stored at -20 degree celcius

NUCLEIC ACID EXTRACTION		
Procedure and/or instrumentation	**E**	Manual
Name of kit and details of any modifications	**E**	mirVana™ miRNA Isolation Kit with phenol (Cat# AM1560, Ambion, Austin, TX, USA). No modifications
Source of additional reagents used	D	
Details of DNase or RNAse treatment	**E**	Used TURBO DNase (TURBO DNA-free kit, Ambion) according to manufaturer’s protocol.
Contamination assessment (DNA or RNA)	**E**	RNA run on gel. PCR products checked (melting curve analysis) as well as agarose gel for some of the products
Nucleic acid quantification	**E**	Nanodrop
Instrument and method	**E**	BioRad CFX96 real-time PCR detection system
Purity (A260/A280)	D	
Yield	D	
RNA integrity method/instrument	**E**	Agarose gel electrophoresis
RIN/RQI or Cq of 3' and 5' transcripts	**E**	Not calculated integriity checked on Agarose gel electrophoresis
Electrophoresis traces	D	
Inhibition testing (Cq dilutions, spike or other)	**E**	Used cDNAs dilutions to make standard curve and check efficiency
REVERSE TRANSCRIPTION		
Complete reaction conditions	**E**	Used iScript^{TM} cDNA synthesis kit and protocol (Bio-Rad Laboratories, Hercules, CA) as detailed in the manuscript text
Amount of RNA and reaction volume	**E**	1 µg RNA, in 20 µl reaction, details given in the manuscript text

Priming oligonucleotide (if using GSP) and concentration	**E**	Primer sequences given as supplemtary data (Table S9). Conc. (0.5 uM), given in text
Reverse transcriptase and concentration	**E**	1uL RT enzyme in 20 µL reaction according to iScript kit protocol (Biorad). Details given in the manuscript text
Temperature and time	**E**	Given in the manuscript text (according to kit protocol)
Manufacturer of reagents and catalogue numbers	D	
Cqs with and without RT	D*	
Storage conditions of cDNA	D	cDNAs stored at -20 degree Celsius
qPCR TARGET INFORMATION		
If multiplex, efficiency and LOD of each assay.	**E**	Not applicable (NA)
Sequence accession number	**E**	Given in primer table in supplemtary data (Table S9).
Location of amplicon	D	
Amplicon length	**E**	Given in primer table in supplemtary data (Table S9).
In silico specificity screen (BLAST, etc)	**E**	Performed, on cod geneome/transcriptome sequence databae
Pseudogenes, retropseudogenes or other homologs?	D	
Sequence alignment	D	
Secondary structure analysis of amplicon	D	
Location of each primer by exon or intron (if applicable)	**E**	Primer pairs designed from different exons and/or at exon-exon junctions.
What splice variants are targeted?	**E**	NA. No specific splice variants targeted, primers usually from 3´-regions.

qPCR OLIGONUCLEOTIDES		
Primer sequences	**E**	Primer sequences given as supplemtary data (Table S9).
RTPrimerDB Identification Number	D	NA
Probe sequences	D**	NA
Location and identity of any modifications	**E**	NA
Manufacturer of oligonucleotides	D	Sigma
Purification method	D	Desalt
qPCR PROTOCOL		
Complete reaction conditions	**E**	Given in the manuscript text
Reaction volume and amount of cDNA/DNA	**E**	20 µl reaction. 0.5 µl of cDNA template per reaction.
Primer, (probe), Mg++ and dNTP concentrations	**E**	Used standard kit (SYBR Green Master mix from Roche) as per manufacturer's protocol.
Polymerase identity and concentration	**E**	Given in the manuscript text
Buffer/kit identity and manufacturer	**E**	Given in the manuscript text
Exact chemical constitution of the buffer	D	Used standard kit, Given in the manuscript text
Additives (SYBR Green I, DMSO, etc.)	**E**	Used standard kit with (SYBR Green Master mix from Roche).
Manufacturer of plates/tubes and catalog number	D	
Complete thermocycling parameters	**E**	Given in the manuscript text
Reaction setup (manual/robotic)	D	manual
Manufacturer of qPCR instrument	**E**	Biorad

qPCR VALIDATION		
Evidence of optimisation (from gradients)	D	
Specificity (gel, sequence, melt, or digest)	**E**	Gel and melting point analysis
For SYBR Green I, Cq of the NTC	**E**	For NTC. No Cq (no amplification)
Standard curves with slope and y-intercept	**E**	Yes, stsndard curve analysis performed for every primer pair used
PCR efficiency calculated from slope	**E**	Yes, efficiency was 91-101%, as determined from from slope of std curves. Given in supplemtary data (Table S9).
Confidence interval for PCR efficiency or standard error	D	
r2 of standard curve	**E**	r2 (0.993-1.000). Given in supplemtary data (Table S9).
Linear dynamic range	**E**	Yes. Assessed using standard curves for each primer pair. Samples have Cq values within the linear range.
Cq variation at lower limit	**E**	Replicates should have Cq std dev less than 0.2
Confidence intervals throughout range	D	
Evidence for limit of detection	**E**	Assessed using standard curves.
If multiplex, efficiency and LOD of each assay.	**E**	NA
DATA ANALYSIS		
qPCR analysis program (source, version)	**E**	CFX Manager 3.0 software, Biorad
Cq method determination	**E**	As per CFX Manager 3.0 software
Outlier identification and disposition	**E**	Statistical test (Graphpad prism)
Results of NTCs	**E**	No amplication or Cq

Justification of number and choice of reference genes	**E**	Validated the refernce gene used (actb) by comparing with at least one other reference gene (no change in expression with conditions)
Description of normalisation method	**E**	described in the manuscript text
Number and concordance of biological replicates	D	Given in the manuscript text
Number and stage (RT or qPCR) of technical replicates	**E**	No RT replicate. Technical triplicates for qPCR
Repeatability (intra-assay variation)	**E**	NA
Reproducibility (inter-assay variation, %CV)	D	
Power analysis	D	
Statistical methods for result significance	**E**	Given in the manuscript text

* Assessing the absence of DNA using a no RT assay is essential when first extracting RNA. Once the sample has been validated as RDNA-free, inclusion of a no-RT control is desirable, but no longer essential.

** Disclosure of the probe sequence is highly desirable and strongly encouraged. However, since not all commercial pre-designed assay

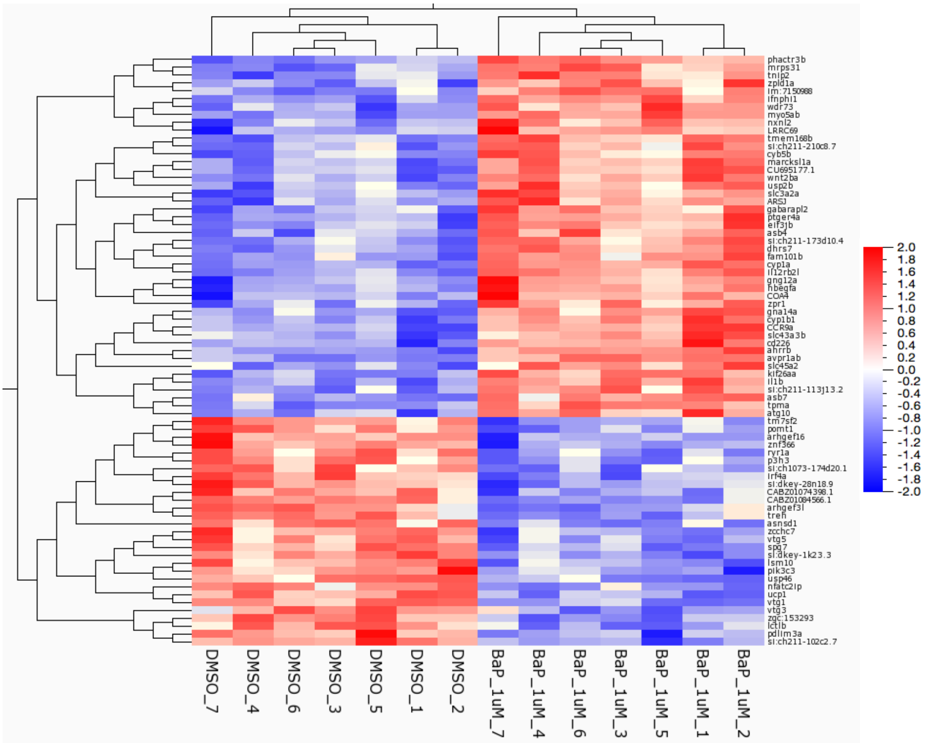


Figure S1. Heatmap of expression profiles of top differentially expressed genes in BaP treated liver slices. The top differentially expressed genes ($p < 0.05$) in DMSO control compared to BaP (1uM) were used in heirarchical cluster analysis performed based on log2-transformed ratio (treated/control) values with Two Group Comparison (Qlucore Omics Explorer). Rows represent genes and columns represent samples. Legend bar on the right indicates colour codes for log2-fold-changes ranging from deep red (2.0) to deep blue (-2.0).

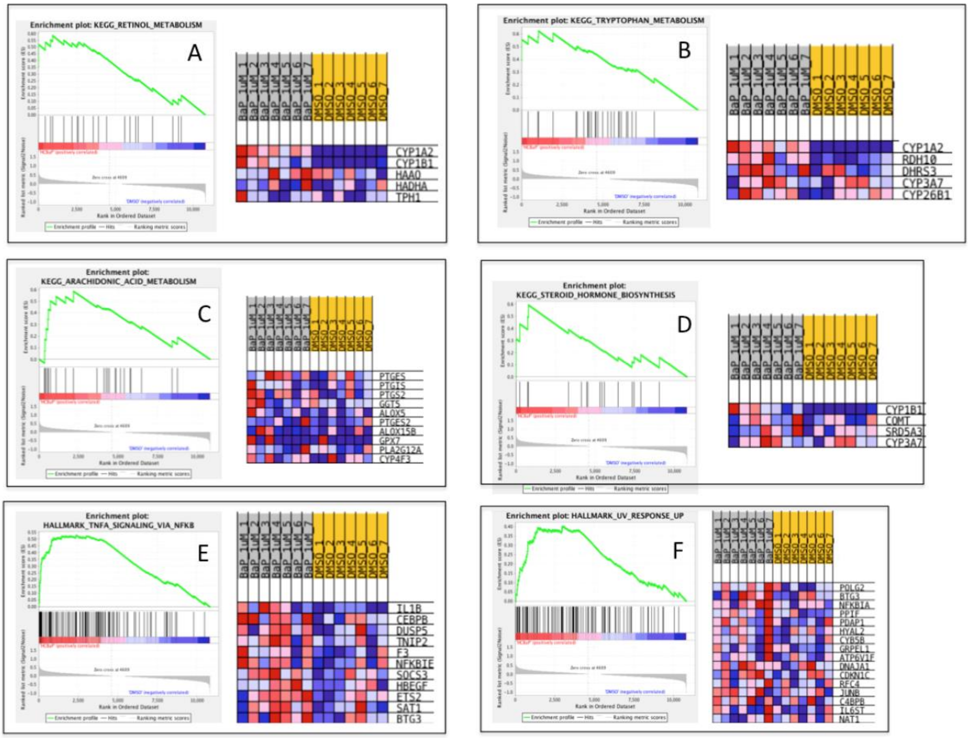


Figure S2. Gene set enrichment analysis (GSEA) of genes quantified by RNA-Seq in PCLS treated with BaP (1 µM) and DMSO. Significantly enriched KEGG gene sets (A-D) and Hallmarks gene sets (E and F) are shown here. In each case (A-F) the left and right panel represents enrichment plot and heatmap for the top leading edge genes. GSEA was performed with human orthologs of Atlantic cod genes. NES, Normalized Enrichment Score; FDR-q value, False Discovery Rate q-value. See Figure S3 for heatmap legends.

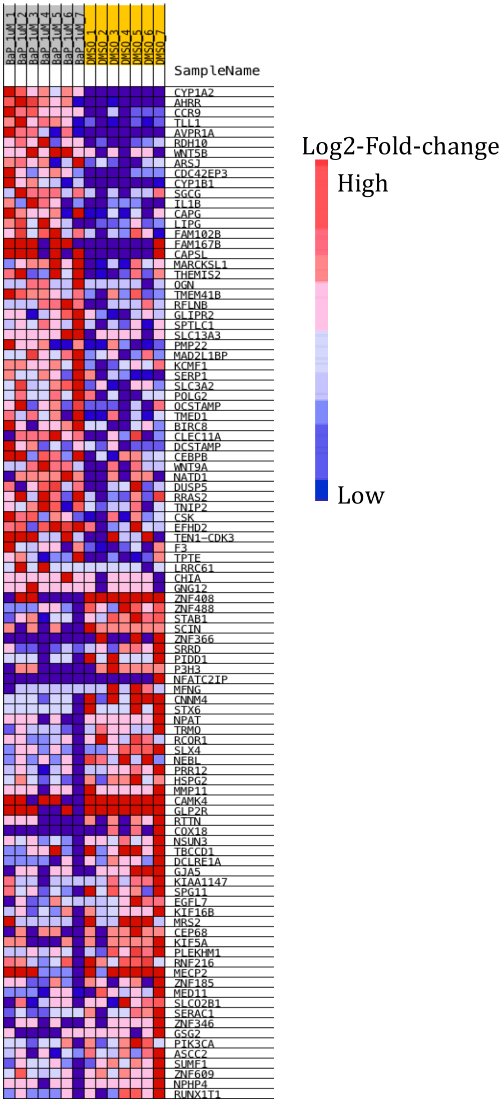


Figure S3. Heatmap from gene set enrichment analysis (GSEA) of genes quantified by RNA-Seq in PCLS treated with BaP (1 μM) and DMSO.

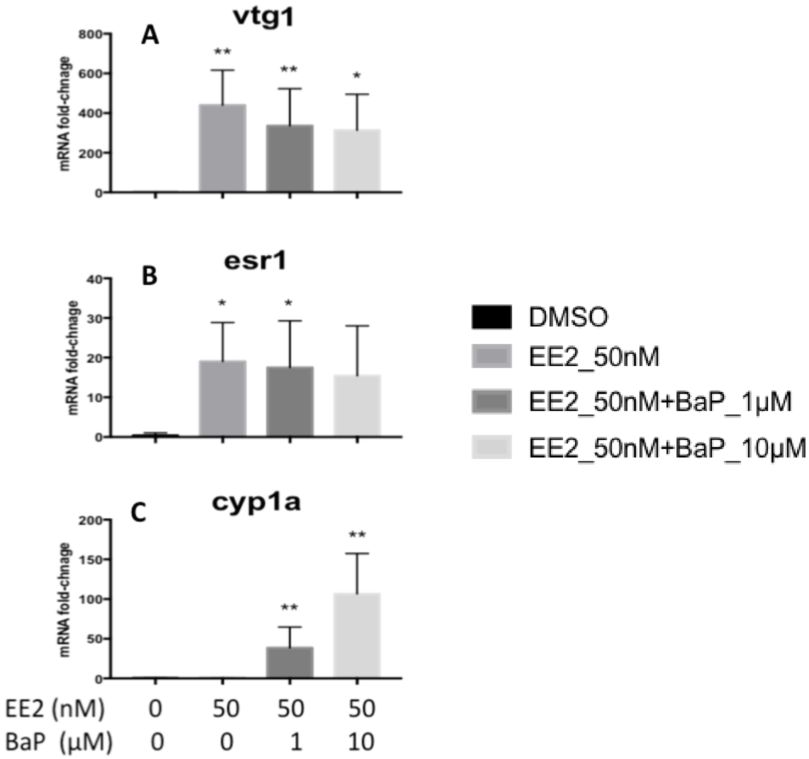
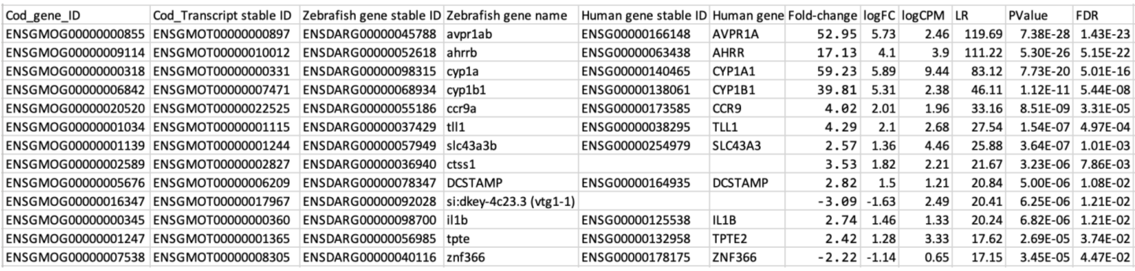


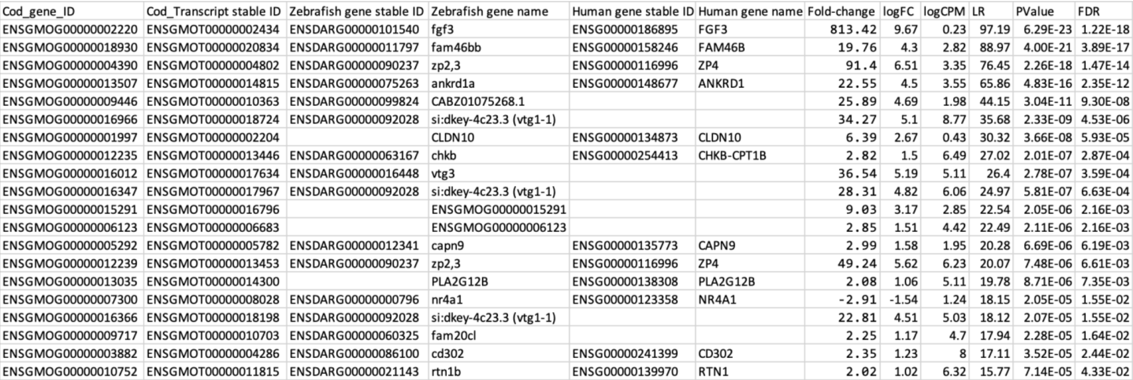
Figure S4. qPCR analysis of genes differentially expressed in PCLS treated with BaP and mixtures of BaP and EE2. The plots show fold-changes in mRNA levels for *vtg1* (**A**) and *esr1* (**B**) and *cyp1a* (**C**) genes in slices treated with the compounds as shown. RNA samples from liver slices of 4 fish liver samples (n = 4) were used for qPCR assays. The significant changes indicated are for comparison with DMSO control. **p* < 0.05; ***p* < 0.01 (one-way ANOVA followed by Dunnett's multiple comparison test). Data points present mean ± SD.

Table S1. Genes differentially expressed in BaP (10 and 1000 nM) treated PCLS.



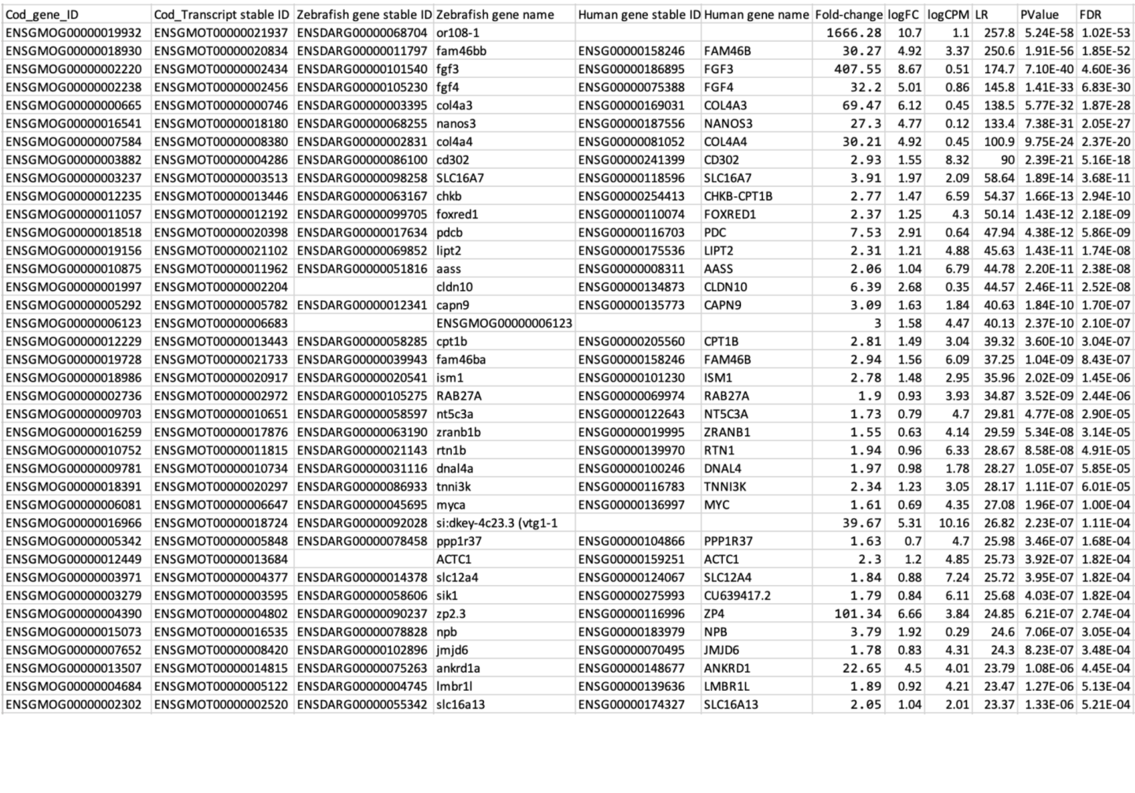
Cod_gene_ID	Cod_Transcript stable ID	Zebrafish gene stable ID	Zebrafish gene name	Human gene stable ID	Human gene	Fold-change	logFC	logCPM	LR	PValue	FDR
ENSGMOG00000000855	ENSGMOT00000000897	ENSDARG00000045788	avpr1ab	ENSG00000166148	AVPR1A	52.95	5.73	2.46	119.69	7.38E-28	1.43E-23
ENSGMOG00000009114	ENSGMOT00000010012	ENSDARG00000052618	ahrrb	ENSG00000063438	AHRR	17.13	4.1	3.9	111.22	5.30E-26	5.15E-22
ENSGMOG00000000318	ENSGMOT00000000331	ENSDARG00000098315	cyp1a	ENSG00000140465	CYP1A1	59.23	5.89	9.44	83.12	7.73E-20	5.01E-16
ENSGMOG00000006842	ENSGMOT00000007471	ENSDARG00000068934	cyp1b1	ENSG00000138061	CYP1B1	39.81	5.31	2.38	46.11	1.12E-11	5.44E-08
ENSGMOG00000020520	ENSGMOT00000022525	ENSDARG00000055186	ccr9a	ENSG00000173585	CCR9	4.02	2.01	1.96	33.16	8.51E-09	3.31E-05
ENSGMOG00000001034	ENSGMOT00000001115	ENSDARG00000037429	tll1	ENSG00000038295	TLL1	4.29	2.1	2.68	27.54	1.54E-07	4.97E-04
ENSGMOG00000001139	ENSGMOT00000001244	ENSDARG00000057949	slc43a3b	ENSG00000254979	SLC43A3	2.57	1.36	4.46	25.88	3.64E-07	1.01E-03
ENSGMOG00000002589	ENSGMOT00000002827	ENSDARG00000036940	ctss1			3.53	1.82	2.21	21.67	3.23E-06	7.86E-03
ENSGMOG00000005676	ENSGMOT00000006209	ENSDARG00000078347	DCSTAMP	ENSG00000164935	DCSTAMP	2.82	1.5	1.21	20.84	5.00E-06	1.08E-02
ENSGMOG00000016347	ENSGMOT00000017967	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			-3.09	-1.63	2.49	20.41	6.25E-06	1.21E-02
ENSGMOG00000000345	ENSGMOT00000000360	ENSDARG00000098700	il1b	ENSG00000125538	IL1B	2.74	1.46	1.33	20.24	6.82E-06	1.21E-02
ENSGMOG00000001247	ENSGMOT00000001365	ENSDARG00000056985	tpte	ENSG00000132958	TPTE2	2.42	1.28	3.33	17.62	2.69E-05	3.74E-02
ENSGMOG00000007538	ENSGMOT00000008305	ENSDARG00000040116	znf366	ENSG00000178175	ZNF366	-2.22	-1.14	0.65	17.15	3.45E-05	4.47E-02

Table S2. Genes differentially expressed in 10 nM EE2 treated PCLS.



Cod_gene_ID	Cod_Transcript stable ID	Zebrafish gene stable ID	Zebrafish gene name	Human gene stable ID	Human gene name	Fold-change	logFC	logCPM	LR	PValue	FDR
ENSGMOG00000002220	ENSGMOT00000002434	ENSDARG00000101540	fgf3	ENSG00000186895	FGF3	813.42	9.67	0.23	97.19	6.29E-23	1.22E-18
ENSGMOG00000018930	ENSGMOT00000020834	ENSDARG00000011797	fam46bb	ENSG00000158246	FAM46B	19.76	4.3	2.82	88.97	4.00E-21	3.89E-17
ENSGMOG00000004390	ENSGMOT00000004802	ENSDARG00000090237	zp2,3	ENSG00000116996	ZP4	91.4	6.51	3.35	76.45	2.26E-18	1.47E-14
ENSGMOG00000013507	ENSGMOT00000014815	ENSDARG00000075263	ankrd1a	ENSG00000148677	ANKRD1	22.55	4.5	3.55	65.86	4.83E-16	2.35E-12
ENSGMOG00000009446	ENSGMOT00000010363	ENSDARG00000099824	CABZ01075268.1			25.89	4.69	1.98	44.15	3.04E-11	9.30E-08
ENSGMOG00000016966	ENSGMOT00000018724	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			34.27	5.1	8.77	35.68	2.33E-09	4.53E-06
ENSGMOG00000001997	ENSGMOT00000002204		CLDN10	ENSG00000134873	CLDN10	6.39	2.67	0.43	30.32	3.66E-08	5.93E-05
ENSGMOG00000012235	ENSGMOT00000013446	ENSDARG00000063167	chkb	ENSG00000254413	CHKB-CPT1B	2.82	1.5	6.49	27.02	2.01E-07	2.87E-04
ENSGMOG00000016012	ENSGMOT00000017634	ENSDARG00000016448	vtg3			36.54	5.19	5.11	26.4	2.78E-07	3.59E-04
ENSGMOG00000016347	ENSGMOT00000017967	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			28.31	4.82	6.06	24.97	5.81E-07	6.63E-04
ENSGMOG00000015291	ENSGMOT00000016796		ENSGMOG00000015291			9.03	3.17	2.85	22.54	2.05E-06	2.16E-03
ENSGMOG00000006123	ENSGMOT00000006683		ENSGMOG00000006123			2.85	1.51	4.42	22.49	2.11E-06	2.16E-03
ENSGMOG00000005292	ENSGMOT00000005782	ENSDARG00000012341	capn9	ENSG00000135773	CAPN9	2.99	1.58	1.95	20.28	6.69E-06	6.19E-03
ENSGMOG00000012239	ENSGMOT00000013453	ENSDARG00000090237	zp2,3	ENSG00000116996	ZP4	49.24	5.62	6.23	20.07	7.48E-06	6.61E-03
ENSGMOG00000013035	ENSGMOT00000014300		PLA2G12B	ENSG00000138308	PLA2G12B	2.08	1.06	5.11	19.78	8.71E-06	7.35E-03
ENSGMOG00000007300	ENSGMOT00000008028	ENSDARG00000000796	nr4a1	ENSG00000123358	NR4A1	-2.91	-1.54	1.24	18.15	2.05E-05	1.55E-02
ENSGMOG00000016366	ENSGMOT00000018198	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			22.81	4.51	5.03	18.12	2.07E-05	1.55E-02
ENSGMOG00000009717	ENSGMOT00000010703	ENSDARG00000060325	fam20cl			2.25	1.17	4.7	17.94	2.28E-05	1.64E-02
ENSGMOG00000003882	ENSGMOT00000004286	ENSDARG00000086100	cd302	ENSG00000241399	CD302	2.35	1.23	8	17.11	3.52E-05	2.44E-02
ENSGMOG00000010752	ENSGMOT00000011815	ENSDARG00000021143	rtn1b	ENSG00000139970	RTN1	2.02	1.02	6.32	15.77	7.14E-05	4.33E-02

Table S3. Genes differentially expressed in 1000 nM EE2 treated PCLS.



Cod_gene_ID	Cod_Transcript stable ID	Zebrafish gene stable ID	Zebrafish gene name	Human gene stable ID	Human gene name	Fold-change	logFC	logCPM	LR	PValue	FDR
ENSGMOG00000019932	ENSGMOT00000021937	ENSDARG00000068704	or108-1			1666.28	10.7	1.1	257.8	5.24E-58	1.02E-53
ENSGMOG00000018930	ENSGMOT00000020834	ENSDARG00000011797	fam46bb	ENSG00000158246	FAM46B	30.27	4.92	3.37	250.6	1.91E-56	1.85E-52
ENSGMOG00000002220	ENSGMOT00000002434	ENSDARG00000101540	fgf3	ENSG00000186895	FGF3	407.55	8.67	0.51	174.7	7.10E-40	4.60E-36
ENSGMOG00000002238	ENSGMOT00000002456	ENSDARG00000105230	fgf4	ENSG00000075388	FGF4	32.2	5.01	0.86	145.8	1.41E-33	6.83E-30
ENSGMOG00000000665	ENSGMOT00000000746	ENSDARG00000003395	col4a3	ENSG00000169031	COL4A3	69.47	6.12	0.45	138.5	5.77E-32	1.87E-28
ENSGMOG00000016541	ENSGMOT00000018180	ENSDARG00000068255	nanos3	ENSG00000187556	NANOS3	27.3	4.77	0.12	133.4	7.38E-31	2.05E-27
ENSGMOG00000007584	ENSGMOT00000008380	ENSDARG00000002831	col4a4	ENSG00000081052	COL4A4	30.21	4.92	0.45	100.9	9.75E-24	2.37E-20
ENSGMOG00000003882	ENSGMOT00000004286	ENSDARG00000086100	cd302	ENSG00000241399	CD302	2.93	1.55	8.32	90	2.39E-21	5.16E-18
ENSGMOG00000003237	ENSGMOT00000003513	ENSDARG00000098258	SLC16A7	ENSG00000118596	SLC16A7	3.91	1.97	2.09	58.64	1.89E-14	3.68E-11
ENSGMOG00000012235	ENSGMOT00000013446	ENSDARG00000063167	chkb	ENSG00000254413	CHKB-CPT1B	2.77	1.47	6.59	54.37	1.66E-13	2.94E-10
ENSGMOG00000011057	ENSGMOT00000012192	ENSDARG00000099705	foxred1	ENSG00000110074	FOXRED1	2.37	1.25	4.3	50.14	1.43E-12	2.18E-09
ENSGMOG00000018518	ENSGMOT00000020398	ENSDARG00000017634	pdcb	ENSG00000116703	PDC	7.53	2.91	0.64	47.94	4.38E-12	5.86E-09
ENSGMOG00000019156	ENSGMOT00000021102	ENSDARG00000069852	lipt2	ENSG00000175536	LIPT2	2.31	1.21	4.88	45.63	1.43E-11	1.74E-08
ENSGMOG00000010875	ENSGMOT00000011962	ENSDARG00000051816	aass	ENSG00000008311	AASS	2.06	1.04	6.79	44.78	2.20E-11	2.38E-08
ENSGMOG00000001997	ENSGMOT00000002204		cldn10	ENSG00000134873	CLDN10	6.39	2.68	0.35	44.57	2.46E-11	2.52E-08
ENSGMOG00000005292	ENSGMOT00000005782	ENSDARG00000012341	capn9	ENSG00000135773	CAPN9	3.09	1.63	1.84	40.63	1.84E-10	1.70E-07
ENSGMOG00000006123	ENSGMOT00000006683		ENSGMOG00000006123			3	1.58	4.47	40.13	2.37E-10	2.10E-07
ENSGMOG00000012229	ENSGMOT00000013443	ENSDARG00000058285	cpt1b	ENSG00000205560	CPT1B	2.81	1.49	3.04	39.32	3.60E-10	3.04E-07
ENSGMOG00000019728	ENSGMOT00000021733	ENSDARG00000039943	fam46ba	ENSG00000158246	FAM46B	2.94	1.56	6.09	37.25	1.04E-09	8.43E-07
ENSGMOG00000018986	ENSGMOT00000020917	ENSDARG00000020541	ism1	ENSG00000101230	ISM1	2.78	1.48	2.95	35.96	2.02E-09	1.45E-06
ENSGMOG00000002736	ENSGMOT00000002972	ENSDARG00000105275	RAB27A	ENSG00000069974	RAB27A	1.9	0.93	3.93	34.87	3.52E-09	2.44E-06
ENSGMOG00000009703	ENSGMOT00000010651	ENSDARG00000058597	nt5c3a	ENSG00000122643	NT5C3A	1.73	0.79	4.7	29.81	4.77E-08	2.90E-05
ENSGMOG00000016259	ENSGMOT00000017876	ENSDARG00000063190	zranb1b	ENSG00000019995	ZRANB1	1.55	0.63	4.14	29.59	5.34E-08	3.14E-05
ENSGMOG00000010752	ENSGMOT00000011815	ENSDARG00000021143	rtn1b	ENSG00000139970	RTN1	1.94	0.96	6.33	28.67	8.58E-08	4.91E-05
ENSGMOG00000009781	ENSGMOT00000010734	ENSDARG00000031116	dnal4a	ENSG00000100246	DNAL4	1.97	0.98	1.78	28.27	1.05E-07	5.85E-05
ENSGMOG00000018391	ENSGMOT00000020297	ENSDARG00000086933	tnni3k	ENSG00000116783	TNNI3K	2.34	1.23	3.05	28.17	1.11E-07	6.01E-05
ENSGMOG00000006081	ENSGMOT00000006647	ENSDARG00000045695	myca	ENSG00000136997	MYC	1.61	0.69	4.35	27.08	1.96E-07	1.00E-04
ENSGMOG00000016966	ENSGMOT00000018724	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1			39.67	5.31	10.16	26.82	2.23E-07	1.11E-04
ENSGMOG00000005342	ENSGMOT00000005848	ENSDARG00000078458	ppp1r37	ENSG00000104866	PPP1R37	1.63	0.7	4.7	25.98	3.46E-07	1.68E-04
ENSGMOG00000012449	ENSGMOT00000013684		ACTC1	ENSG00000159251	ACTC1	2.3	1.2	4.85	25.73	3.92E-07	1.82E-04
ENSGMOG00000003971	ENSGMOT00000004377	ENSDARG00000014378	slc12a4	ENSG00000124067	SLC12A4	1.84	0.88	7.24	25.72	3.95E-07	1.82E-04
ENSGMOG00000003279	ENSGMOT00000003595	ENSDARG00000058606	sik1	ENSG00000275993	CU639417.2	1.79	0.84	6.11	25.68	4.03E-07	1.82E-04
ENSGMOG00000004390	ENSGMOT00000004802	ENSDARG00000090237	zp2.3	ENSG00000116996	ZP4	101.34	6.66	3.84	24.85	6.21E-07	2.74E-04
ENSGMOG00000015073	ENSGMOT00000016535	ENSDARG00000078828	npb	ENSG00000183979	NPB	3.79	1.92	0.29	24.6	7.06E-07	3.05E-04
ENSGMOG00000007652	ENSGMOT00000008420	ENSDARG00000102896	jmjd6	ENSG00000070495	JMJD6	1.78	0.83	4.31	24.3	8.23E-07	3.48E-04
ENSGMOG00000013507	ENSGMOT00000014815	ENSDARG00000075263	ankrd1a	ENSG00000148677	ANKRD1	22.65	4.5	4.01	23.79	1.08E-06	4.45E-04
ENSGMOG00000004684	ENSGMOT00000005122	ENSDARG00000004745	lmbr1l	ENSG00000139636	LMBR1L	1.89	0.92	4.21	23.47	1.27E-06	5.13E-04
ENSGMOG00000002302	ENSGMOT00000002520	ENSDARG00000055342	slc16a13	ENSG00000174327	SLC16A13	2.05	1.04	2.01	23.37	1.33E-06	5.21E-04

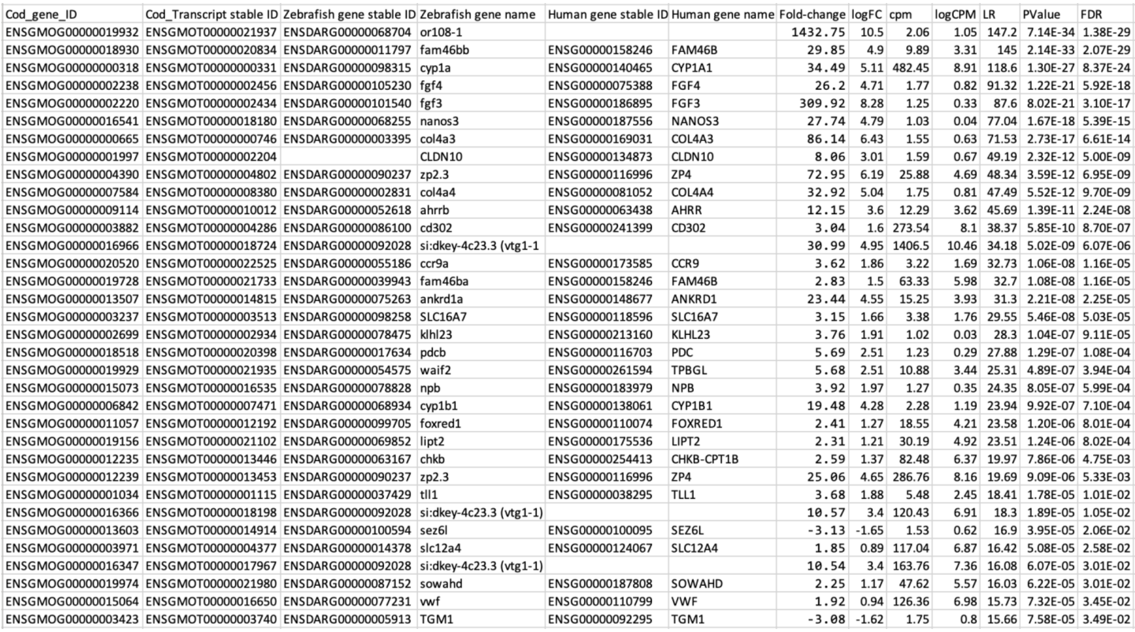
Table S3 (cont.).

ENSGMOG00000013035	ENSGMOT00000014300		PLA2G12B	ENSG00000138308	PLA2G12B	1.56	0.64	4.65	23.36	1.34E-06	5.21E-04
ENSGMOG00000018437	ENSGMOT00000020331	ENSDARG00000042122	acot11b	ENSG00000162390	ACOT11	1.7	0.77	4.17	22.82	1.78E-06	6.66E-04
ENSGMOG00000011560	ENSGMOT00000012702	ENSDARG00000070831	rftn1a	ENSG00000131378	RFTN1	2.3	1.2	1.46	22.02	2.70E-06	9.90E-04
ENSGMOG00000005198	ENSGMOT00000005662	ENSDARG00000030237	pgrmc2	ENSG00000164040	PGRMC2	1.64	0.72	5.31	21.85	2.95E-06	1.06E-03
ENSGMOG00000009714	ENSGMOT00000010671	ENSDARG00000016904	bckdk	ENSG00000103507	BCKDK	1.7	0.76	5.89	21.66	3.26E-06	1.15E-03
ENSGMOG00000003762	ENSGMOT00000004115		CLEC18C	ENSG00000157335	CLEC18C	2.32	1.21	0.78	21.59	3.38E-06	1.17E-03
ENSGMOG00000016102	ENSGMOT00000018127	ENSDARG00000036237	slc27a2a	ENSG00000140284	SLC27A2	-1.67	-0.74	5.84	20.98	4.65E-06	1.59E-03
ENSGMOG00000015089	ENSGMOT00000016574	ENSDARG00000092350	si:dkey-222h21.12	ENSG00000274252	GGTLC3	3.04	1.6	0.95	20.82	5.05E-06	1.69E-03
ENSGMOG00000006837	ENSGMOT00000007474	ENSDARG00000062745	socs5b	ENSG00000171150	SOCS5	1.81	0.86	2.92	20.36	6.43E-06	2.08E-03
ENSGMOG00000010514	ENSGMOT00000011569	ENSDARG00000030263	mfsd2b	ENSG00000205639	MFSD2B	2.57	1.36	0.32	20.12	7.26E-06	2.31E-03
ENSGMOG00000011179	ENSGMOT00000012272	ENSDARG00000043835	rab3ab	ENSG00000105649	RAB3A	-2.33	-1.22	0.9	20.07	7.47E-06	2.34E-03
ENSGMOG00000012239	ENSGMOT00000013453	ENSDARG00000090237	zp2.3	ENSG00000116996	ZP4	41.66	5.38	7.36	19.19	1.19E-05	3.57E-03
ENSGMOG00000006817	ENSGMOT00000007453	ENSDARG00000075121	hbegfa	ENSG00000113070	HBEGF	1.59	0.67	5.23	19.17	1.19E-05	3.57E-03
ENSGMOG00000004610	ENSGMOT00000005036	ENSDARG00000013880	spata20	ENSG00000006282	SPATA20	1.65	0.72	6.17	19.04	1.28E-05	3.77E-03
ENSGMOG00000008920	ENSGMOT00000009787	ENSDARG00000102380	FO704641.1			1.74	0.79	2.07	18.86	1.41E-05	4.08E-03
ENSGMOG00000002650	ENSGMOT00000002888	ENSDARG00000034893	rarab	ENSG00000131759	RARA	1.93	0.95	3.11	18.75	1.49E-05	4.25E-03
ENSGMOG00000001777	ENSGMOT00000001944	ENSDARG00000045443	agtr1b	ENSG00000144891	AGTR1	1.89	0.92	2.01	18.72	1.51E-05	4.26E-03
ENSGMOG00000013259	ENSGMOT00000014562	ENSDARG00000098837	tgm5l	ENSG00000125780	TGM3	2.32	1.21	3.36	18.23	1.96E-05	5.37E-03
ENSGMOG00000007683	ENSGMOT00000008457		SYNPO2	ENSG00000172403	SYNPO2	4.13	2.05	0.29	18.16	2.04E-05	5.50E-03
ENSGMOG00000000575	ENSGMOT00000000592	ENSDARG00000079572	plcd3b	ENSG00000161714	PLCD3	1.73	0.79	6.38	17.89	2.34E-05	6.24E-03
ENSGMOG00000014528	ENSGMOT00000015979		COL6A2	ENSG00000142173	COL6A2	2.78	1.47	0.45	17.62	2.70E-05	7.11E-03
ENSGMOG00000019929	ENSGMOT00000021935	ENSDARG00000054575	waif2	ENSG00000261594	TPBGL	5.86	2.55	3.39	17.1	3.55E-05	9.09E-03
ENSGMOG00000011232	ENSGMOT00000012342	ENSDARG00000013997	ern1	ENSG00000178607	ERN1	1.52	0.6	5.44	16.66	4.48E-05	1.12E-02
ENSGMOG00000005554	ENSGMOT00000006084	ENSDARG00000086585	nrg2b	ENSG00000158458	NRG2	2.46	1.3	0.34	16.35	5.26E-05	1.26E-02
ENSGMOG00000019974	ENSGMOT00000021980	ENSDARG00000087152	sowahd	ENSG00000187808	SOWAHD	1.83	0.87	5.39	16.29	5.42E-05	1.29E-02
ENSGMOG00000007600	ENSGMOT00000008367	ENSDARG00000006074	uck2a	ENSG00000143179	UCK2	2.6	1.38	1.73	16.26	5.51E-05	1.29E-02
ENSGMOG00000000785	ENSGMOT00000000817	ENSDARG00000059680	fscn1a	ENSG00000075618	FSCN1	1.74	0.8	3.02	15.54	8.06E-05	1.76E-02
ENSGMOG00000000846	ENSGMOT00000001017		NRCAM	ENSG00000091129	NRCAM	1.95	0.96	3.19	14.57	1.35E-04	2.85E-02
ENSGMOG00000000777	ENSGMOT00000000816	ENSDARG00000074337	cbfa2t2	ENSG00000078699	CBFA2T2	1.86	0.89	0.66	14.19	1.65E-04	3.45E-02
ENSGMOG00000001894	ENSGMOT00000002072		PCDH17	ENSG00000118946	PCDH17	1.89	0.92	2.23	14.07	1.76E-04	3.59E-02
ENSGMOG00000005645	ENSGMOT00000006170	ENSDARG00000075914	kcnj14	ENSG00000182324	KCNJ14	-1.55	-0.63	2.26	13.86	1.97E-04	3.92E-02
ENSGMOG00000005971	ENSGMOT00000006528	ENSDARG00000068732	spry4	ENSG00000187678	SPRY4	1.62	0.7	3.79	13.7	2.14E-04	4.13E-02
ENSGMOG00000001992	ENSGMOT00000002186	ENSDARG00000077396	tlcd2	ENSG00000185561	TLCD2	-1.81	-0.86	0.36	13.65	2.20E-04	4.20E-02
ENSGMOG00000011783	ENSGMOT00000012946	ENSDARG00000042008	BX936298.1	ENSG00000197540	GZMM	-1.58	-0.66	2.27	13.62	2.23E-04	4.22E-02
ENSGMOG00000005873	ENSGMOT00000006446	ENSDARG00000098777	CU693379.1	ENSG00000176092	CRYBG2	-1.57	-0.65	3.03	13.34	2.60E-04	4.73E-02

Table S4. Genes differentially expressed in 10 nM Mix (BaP + EE2, 10nM each) treated PCLS.

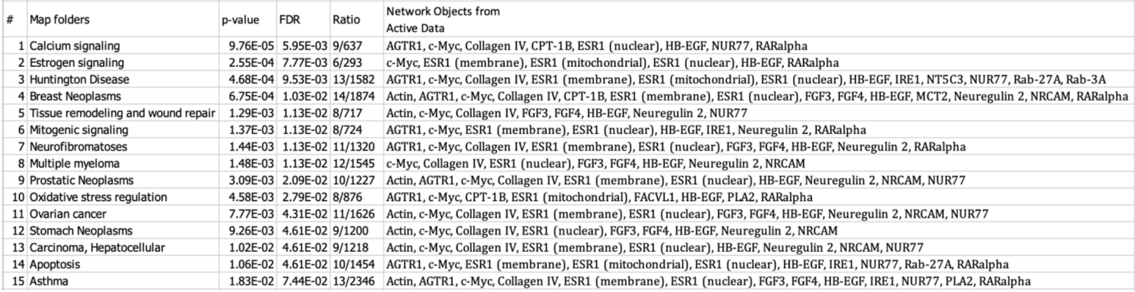
Cod_gene_ID	Cod_Transcript stable ID	Zebrafish gene stable ID	Zebrafish gene name	Human gene stable ID	Human gene	Fold-change	logFC	logCPM	LR	PValue	FDR
ENSGMOG00000018930	ENSGMOT00000020834	ENSDARG00000011797	fam46bb	ENSG00000158246	FAM46B	18.98	4.25	2.72	188.9	5.45E-43	1.05E-38
ENSGMOG00000002238	ENSGMOT00000002456	ENSDARG00000105230	fgf4	ENSG00000075388	FGF4	26.02	4.7	0.58	114.2	1.15E-26	1.11E-22
ENSGMOG00000004390	ENSGMOT00000004802	ENSDARG00000090237	zp2.3	ENSG00000116996	ZP4	79.22	6.31	3.15	111.6	4.28E-26	2.75E-22
ENSGMOG00000013507	ENSGMOT00000014815	ENSDARG00000075263	ankrd1a	ENSG00000148677	ANKRD1	32.56	5.03	3.26	55.74	8.27E-14	2.28E-10
ENSGMOG00000001997	ENSGMOT00000002204		CLDN10	ENSG00000134873	CLDN10	8.37	3.07	1.11	52.18	5.07E-13	1.09E-09
ENSGMOG00000012449	ENSGMOT00000013684		ACTC1	ENSG00000159251	ACTC1	2.76	1.46	5.14	42.9	5.76E-11	1.05E-07
ENSGMOG00000015291	ENSGMOT00000016796		ENSGMOG00000015291			7.02	2.81	2.25	42.82	6.00E-11	1.05E-07
ENSGMOG00000003882	ENSGMOT00000004286	ENSDARG00000086100	cd302	ENSG00000241399	CD302	2.28	1.19	8.02	41	1.52E-10	2.44E-07
ENSGMOG00000003237	ENSGMOT00000003513	ENSDARG00000098258	SLC16A7	ENSG00000118596	SLC16A7	3.08	1.62	1.77	38.78	4.73E-10	7.03E-07
ENSGMOG00000012235	ENSGMOT00000013446	ENSDARG00000063167	chkb	ENSG00000254413	CHKB-CPT1B	2.59	1.37	6.58	38.2	6.38E-10	8.21E-07
ENSGMOG00000017460	ENSGMOT00000019243		ENSGMOG00000017460			4.91	2.3	0.15	35.76	2.23E-09	2.70E-06
ENSGMOG00000018518	ENSGMOT00000020398	ENSDARG00000017634	pdcb	ENSG00000116703	PDC	5.04	2.33	0.29	35.37	2.72E-09	3.09E-06
ENSGMOG00000016966	ENSGMOT00000018724	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			45.77	5.52	9.5	33.95	5.65E-09	5.94E-06
ENSGMOG00000012239	ENSGMOT00000013453	ENSDARG00000090237	zp2.3	ENSG00000116996	ZP4	37.41	5.23	6.31	29.25	6.35E-08	6.13E-05
ENSGMOG00000009446	ENSGMOT00000010363	ENSDARG00000099824	CABZ01075268.1			23.72	4.57	0.88	25.19	5.19E-07	4.56E-04
ENSGMOG00000009781	ENSGMOT00000010734	ENSDARG00000031116	dnal4a	ENSG00000100246	DNAL4	1.93	0.95	1.65	25.08	5.50E-07	4.61E-04
ENSGMOG00000019835	ENSGMOT00000021840	ENSDARG00000045444	fzd8a	ENSG00000177283	FZD8	2.3	1.2	2.84	24.74	6.56E-07	5.27E-04
ENSGMOG00000018986	ENSGMOT00000020917	ENSDARG00000020541	ism1	ENSG00000101230	ISM1	1.96	0.97	2.97	24.1	9.17E-07	7.08E-04
ENSGMOG00000005264	ENSGMOT00000005772	ENSDARG00000059361	slc39a4	ENSG00000147804	SLC39A4	2.85	1.51	0.53	23.22	1.44E-06	1.04E-03
ENSGMOG00000005292	ENSGMOT00000005782	ENSDARG00000012341	capn9	ENSG00000135773	CAPN9	2.84	1.51	2.02	22.19	2.46E-06	1.70E-03
ENSGMOG00000003971	ENSGMOT00000004377	ENSDARG00000014378	slc12a4	ENSG00000124067	SLC12A4	1.79	0.84	7.14	21.91	2.86E-06	1.90E-03
ENSGMOG00000012244	ENSGMOT00000013458	ENSDARG00000042130	zp3a.2	ENSG00000188372	ZP3	13.13	3.72	8.33	21.82	3.00E-06	1.93E-03
ENSGMOG00000006123	ENSGMOT00000006683		ENSGMOG00000006123			2.36	1.24	3.36	21.03	4.53E-06	2.82E-03
ENSGMOG00000011560	ENSGMOT00000012702	ENSDARG00000070831	rftn1a	ENSG00000131378	RFTN1	2.2	1.13	1.61	20.09	7.39E-06	4.45E-03
ENSGMOG00000009714	ENSGMOT00000010671	ENSDARG00000016904	bckdk	ENSG00000103507	BCKDK	1.66	0.73	5.66	19.68	9.15E-06	5.20E-03
ENSGMOG00000009114	ENSGMOT00000010012	ENSDARG00000052618	ahrrb	ENSG00000063438	AHRR	2.12	1.09	1.11	19.38	1.07E-05	5.90E-03
ENSGMOG00000004684	ENSGMOT00000005122	ENSDARG00000004745	lmbr1l	ENSG00000139636	LMBR1L	1.96	0.97	3.95	18.07	2.13E-05	1.08E-02
ENSGMOG00000014936	ENSGMOT00000016396	ENSDARG00000021432	odf3l2	ENSG00000181781	ODF3L2	2.58	1.37	0.83	17.8	2.45E-05	1.21E-02
ENSGMOG00000016347	ENSGMOT00000017967	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			31.54	4.98	6.7	17.26	3.26E-05	1.53E-02
ENSGMOG00000010875	ENSGMOT00000011962	ENSDARG00000051816	aass	ENSG00000008311	AASS	1.7	0.77	6.57	17.15	3.45E-05	1.58E-02
ENSGMOG00000010366	ENSGMOT00000011416	0	ENSGMOG00000010366	ENSG00000154134	ROBO3	2.06	1.04	1.45	17	3.73E-05	1.67E-02
ENSGMOG00000019929	ENSGMOT00000021935	ENSDARG00000054575	waif2	ENSG00000261594	TPBGL	3.43	1.78	3.03	16.84	4.07E-05	1.78E-02
ENSGMOG00000010752	ENSGMOT00000011815	ENSDARG00000021143	rtn1b	ENSG00000139970	RTN1	1.62	0.7	6.19	15.21	9.60E-05	3.86E-02
ENSGMOG00000018564	ENSGMOT00000020448	0	DLX3	ENSG00000064195	DLX3	2.38	1.25	1.21	15.09	1.02E-04	3.97E-02
ENSGMOG00000014898	ENSGMOT00000016376	ENSDARG00000004111	esr1	ENSG00000091831	ESR1	12.64	3.66	2.81	15.08	1.03E-04	3.97E-02
ENSGMOG00000009785	ENSGMOT00000010753	ENSDARG00000053375	nptxrb	ENSG00000221890	NPTXR	1.94	0.96	1.77	14.97	1.09E-04	4.05E-02
ENSGMOG00000020045	ENSGMOT00000022050	ENSDARG00000027285	kcnf1b	ENSG00000162975	KCNF1	-2.44	-1.27	0.07	14.49	1.41E-04	4.77E-02
ENSGMOG00000006837	ENSGMOT00000007474	ENSDARG00000062745	socs5b	ENSG00000171150	SOCS5	1.7	0.76	2.74	14.34	1.52E-04	4.98E-02

Table S5. Genes differentially expressed in 1000 nM Mix (BaP + EE2, 10nM each) treated PCLS.



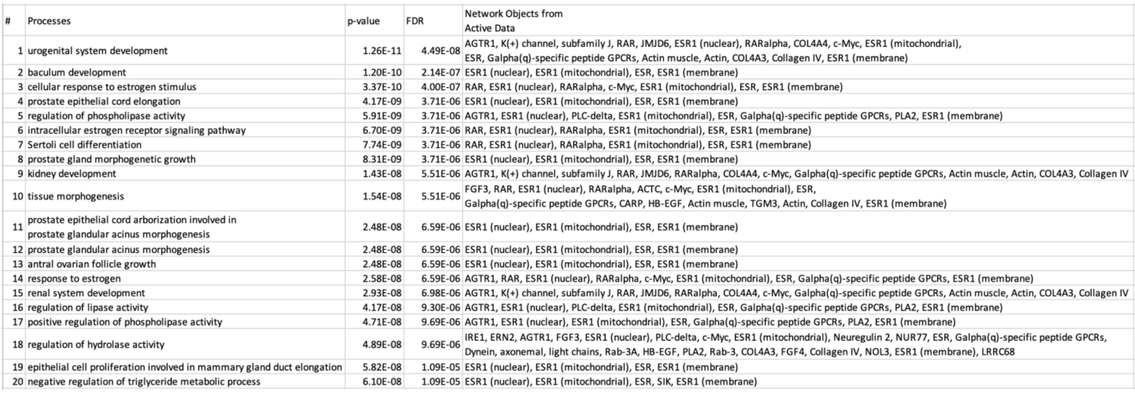
Cod_gene_ID	Cod_Transcript stable ID	Zebrafish gene stable ID	Zebrafish gene name	Human gene stable ID	Human gene name	Fold-change	logFC	cpm	logCPM	LR	PValue	FDR
ENSGMOG00000019932	ENSGMOT00000021937	ENSDARG00000068704	or108-1			1432.75	10.5	2.06	1.05	147.2	7.14E-34	1.38E-29
ENSGMOG00000018930	ENSGMOT00000020834	ENSDARG00000011797	fam46bb	ENSG00000158246	FAM46B	29.85	4.9	9.89	3.31	145	2.14E-33	2.07E-29
ENSGMOG00000000318	ENSGMOT00000000331	ENSDARG00000098315	cyp1a	ENSG00000140465	CYP1A1	34.49	5.11	482.45	8.91	118.6	1.30E-27	8.37E-24
ENSGMOG00000002238	ENSGMOT00000002456	ENSDARG00000105230	fgf4	ENSG00000075388	FGF4	26.2	4.71	1.77	0.82	91.32	1.22E-21	5.92E-18
ENSGMOG00000002220	ENSGMOT00000002434	ENSDARG00000101540	fgf3	ENSG00000186895	FGF3	309.92	8.28	1.25	0.33	87.6	8.02E-21	3.10E-17
ENSGMOG00000016541	ENSGMOT00000018180	ENSDARG00000068255	nanos3	ENSG00000187556	NANOS3	27.74	4.79	1.03	0.04	77.04	1.67E-18	5.39E-15
ENSGMOG00000000665	ENSGMOT00000000746	ENSDARG00000003395	col4a3	ENSG00000169031	COL4A3	86.14	6.43	1.55	0.63	71.53	2.73E-17	6.61E-14
ENSGMOG00000001997	ENSGMOT00000002204		CLDN10	ENSG00000134873	CLDN10	8.06	3.01	1.59	0.67	49.19	2.32E-12	5.00E-09
ENSGMOG00000004390	ENSGMOT00000004802	ENSDARG00000090237	zp2.3	ENSG00000116996	ZP4	72.95	6.19	25.88	4.69	48.34	3.59E-12	6.95E-09
ENSGMOG00000007584	ENSGMOT00000008380	ENSDARG00000002831	col4a4	ENSG00000081052	COL4A4	32.92	5.04	1.75	0.81	47.49	5.52E-12	9.70E-09
ENSGMOG00000009114	ENSGMOT00000010012	ENSDARG00000052618	ahrrb	ENSG00000063438	AHRR	12.15	3.6	12.29	3.62	45.69	1.39E-11	2.24E-08
ENSGMOG00000003882	ENSGMOT00000004286	ENSDARG00000086100	cd302	ENSG00000241399	CD302	3.04	1.6	273.54	8.1	38.37	5.85E-10	8.70E-07
ENSGMOG00000016966	ENSGMOT00000018724	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1			30.99	4.95	1406.5	10.46	34.18	5.02E-09	6.07E-06
ENSGMOG00000020520	ENSGMOT00000022525	ENSDARG00000055186	ccr9a	ENSG00000173585	CCR9	3.62	1.86	3.22	1.69	32.73	1.06E-08	1.16E-05
ENSGMOG00000019728	ENSGMOT00000021733	ENSDARG00000039943	fam46ba	ENSG00000158246	FAM46B	2.83	1.5	63.33	5.98	32.7	1.08E-08	1.16E-05
ENSGMOG00000013507	ENSGMOT00000014815	ENSDARG00000075263	ankrd1a	ENSG00000148677	ANKRD1	23.44	4.55	15.25	3.93	31.3	2.21E-08	2.25E-05
ENSGMOG00000003237	ENSGMOT00000003513	ENSDARG00000098258	SLC16A7	ENSG00000118596	SLC16A7	3.15	1.66	3.38	1.76	29.55	5.46E-08	5.03E-05
ENSGMOG00000002699	ENSGMOT00000002934	ENSDARG00000078475	klhl23	ENSG00000213160	KLHL23	3.76	1.91	1.02	0.03	28.3	1.04E-07	9.11E-05
ENSGMOG00000018518	ENSGMOT00000020398	ENSDARG00000017634	pdcb	ENSG00000116703	PDC	5.69	2.51	1.23	0.29	27.88	1.29E-07	1.08E-04
ENSGMOG00000019929	ENSGMOT00000021935	ENSDARG00000054575	waif2	ENSG00000261594	TPBGL	5.68	2.51	10.88	3.44	25.31	4.89E-07	3.94E-04
ENSGMOG00000015073	ENSGMOT00000016535	ENSDARG00000078828	npb	ENSG00000183979	NPB	3.92	1.97	1.27	0.35	24.35	8.05E-07	5.99E-04
ENSGMOG00000006842	ENSGMOT00000007471	ENSDARG00000068934	cyp1b1	ENSG00000138061	CYP1B1	19.48	4.28	2.28	1.19	23.94	9.92E-07	7.10E-04
ENSGMOG00000011057	ENSGMOT00000012192	ENSDARG00000099705	foxred1	ENSG00000110074	FOXRED1	2.41	1.27	18.55	4.21	23.58	1.20E-06	8.01E-04
ENSGMOG00000019156	ENSGMOT00000021102	ENSDARG00000069852	lipt2	ENSG00000175536	LIPT2	2.31	1.21	30.19	4.92	23.51	1.24E-06	8.02E-04
ENSGMOG00000012235	ENSGMOT00000013446	ENSDARG00000063167	chkb	ENSG00000254413	CHKB-CPT1B	2.59	1.37	82.48	6.37	19.97	7.86E-06	4.75E-03
ENSGMOG00000012239	ENSGMOT00000013453	ENSDARG00000090237	zp2.3	ENSG00000116996	ZP4	25.06	4.65	286.76	8.16	19.69	9.09E-06	5.33E-03
ENSGMOG00000001034	ENSGMOT00000001115	ENSDARG00000037429	tll1	ENSG00000038295	TLL1	3.68	1.88	5.48	2.45	18.41	1.78E-05	1.01E-02
ENSGMOG00000016366	ENSGMOT00000018198	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			10.57	3.4	120.43	6.91	18.3	1.89E-05	1.05E-02
ENSGMOG00000013603	ENSGMOT00000014914	ENSDARG00000100594	sez6l	ENSG00000100095	SEZ6L	-3.13	-1.65	1.53	0.62	16.9	3.95E-05	2.06E-02
ENSGMOG00000003971	ENSGMOT00000004377	ENSDARG00000014378	slc12a4	ENSG00000124067	SLC12A4	1.85	0.89	117.04	6.87	16.42	5.08E-05	2.58E-02
ENSGMOG00000016347	ENSGMOT00000017967	ENSDARG00000092028	si:dkey-4c23.3 (vtg1-1)			10.54	3.4	163.76	7.36	16.08	6.07E-05	3.01E-02
ENSGMOG00000019974	ENSGMOT00000021980	ENSDARG00000087152	sowahd	ENSG00000187808	SOWAHD	2.25	1.17	47.62	5.57	16.03	6.22E-05	3.01E-02
ENSGMOG00000015064	ENSGMOT00000016650	ENSDARG00000077231	vwf	ENSG00000110799	VWF	1.92	0.94	126.36	6.98	15.73	7.32E-05	3.45E-02
ENSGMOG00000003423	ENSGMOT00000003740	ENSDARG00000005913	TGM1	ENSG00000092295	TGM1	-3.08	-1.62	1.75	0.8	15.66	7.58E-05	3.49E-02

Table S6. Enriched MetaCore map folders for genes differentially expressed in EE2 treated PCLS.



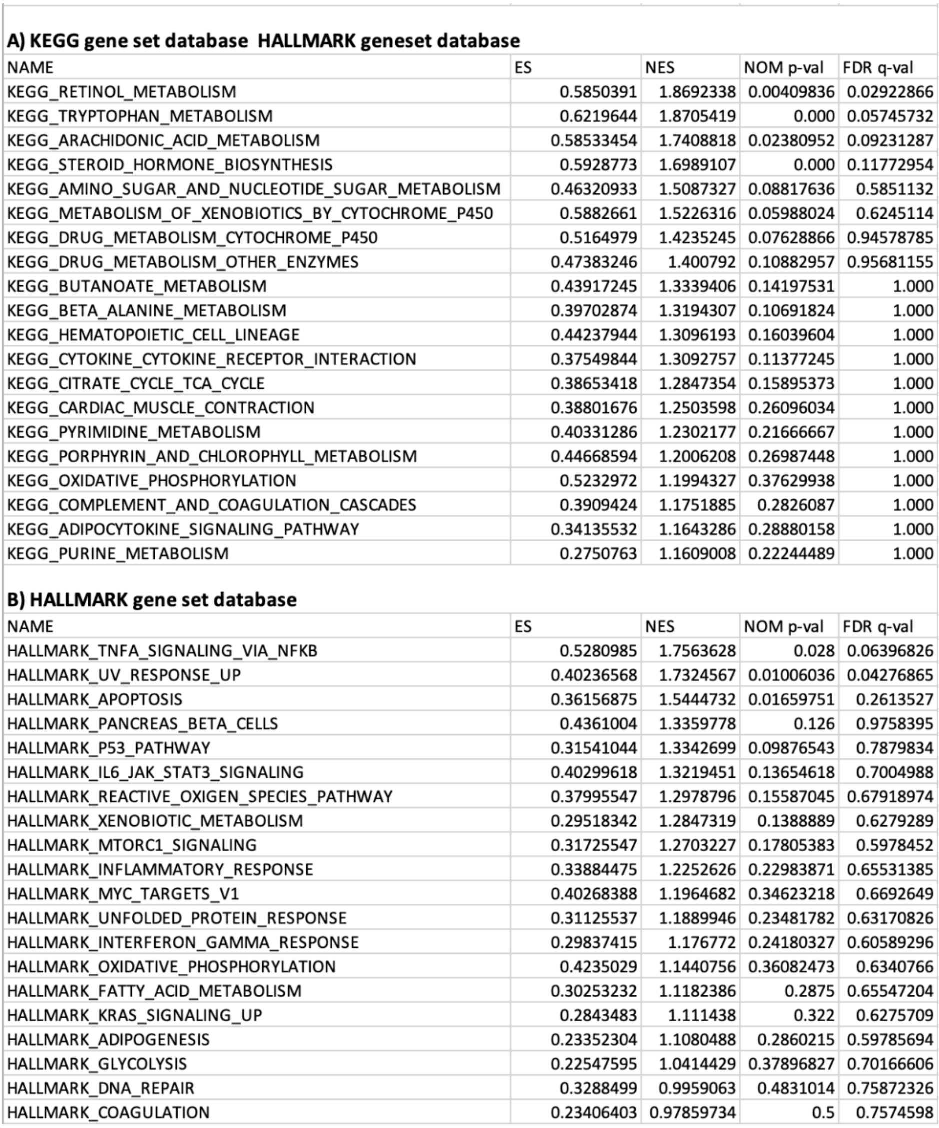
#	Map folders	p-value	FDR	Ratio	Network Objects from Active Data
1	Calcium signaling	9.76E-05	5.95E-03	9/637	AGTR1, c-Myc, Collagen IV, CPT-1B, ESR1 (nuclear), HB-EGF, NUR77, RARalpha
2	Estrogen signaling	2.55E-04	7.77E-03	6/293	c-Myc, ESR1 (membrane), ESR1 (mitochondrial), ESR1 (nuclear), HB-EGF, RARalpha
3	Huntington Disease	4.68E-04	9.53E-03	13/1582	AGTR1, c-Myc, Collagen IV, ESR1 (membrane), ESR1 (mitochondrial), ESR1 (nuclear), HB-EGF, IRE1, NT5C3, NUR77, Rab-27A, Rab-3A
4	Breast Neoplasms	6.75E-04	1.03E-02	14/1874	Actin, AGTR1, c-Myc, Collagen IV, CPT-1B, ESR1 (membrane), ESR1 (nuclear), FGF3, FGF4, HB-EGF, MCT2, Neuregulin 2, NRCAM, RARalpha
5	Tissue remodeling and wound repair	1.29E-03	1.13E-02	8/717	Actin, c-Myc, Collagen IV, FGF3, FGF4, HB-EGF, Neuregulin 2, NUR77
6	Mitogenic signaling	1.37E-03	1.13E-02	8/724	AGTR1, c-Myc, ESR1 (membrane), ESR1 (nuclear), HB-EGF, IRE1, Neuregulin 2, RARalpha
7	Neurofibromatoses	1.44E-03	1.13E-02	11/1320	AGTR1, c-Myc, Collagen IV, ESR1 (membrane), ESR1 (nuclear), FGF3, FGF4, HB-EGF, Neuregulin 2, RARalpha
8	Multiple myeloma	1.48E-03	1.13E-02	12/1545	c-Myc, Collagen IV, ESR1 (nuclear), FGF3, FGF4, HB-EGF, Neuregulin 2, NRCAM
9	Prostatic Neoplasms	3.09E-03	2.09E-02	10/1227	Actin, AGTR1, c-Myc, Collagen IV, ESR1 (membrane), ESR1 (nuclear), HB-EGF, Neuregulin 2, NRCAM, NUR77
10	Oxidative stress regulation	4.58E-03	2.79E-02	8/876	AGTR1, c-Myc, CPT-1B, ESR1 (mitochondrial), FACVL1, HB-EGF, PLA2, RARalpha
11	Ovarian cancer	7.77E-03	4.31E-02	11/1626	Actin, c-Myc, Collagen IV, ESR1 (membrane), ESR1 (nuclear), FGF3, FGF4, HB-EGF, Neuregulin 2, NRCAM, NUR77
12	Stomach Neoplasms	9.26E-03	4.61E-02	9/1200	Actin, c-Myc, Collagen IV, ESR1 (nuclear), FGF3, FGF4, HB-EGF, Neuregulin 2, NRCAM
13	Carcinoma, Hepatocellular	1.02E-02	4.61E-02	9/1218	Actin, c-Myc, Collagen IV, ESR1 (membrane), ESR1 (nuclear), HB-EGF, Neuregulin 2, NRCAM, NUR77
14	Apoptosis	1.06E-02	4.61E-02	10/1454	AGTR1, c-Myc, ESR1 (membrane), ESR1 (mitochondrial), ESR1 (nuclear), HB-EGF, IRE1, NUR77, Rab-27A, RARalpha
15	Asthma	1.83E-02	7.44E-02	13/2346	Actin, AGTR1, c-Myc, Collagen IV, ESR1 (membrane), ESR1 (nuclear), FGF3, FGF4, HB-EGF, IRE1, NUR77, PLA2, RARalpha

Table S7. Enriched MetaCore gene ontology (GO) biological process (BP) for genes differentially expressed in EE2 treated PCLS.



#	Processes	p-value	FDR	Network Objects from Active Data
1	urogenital system development	1.26E-11	4.49E-08	AGTR1, K(+) channel, subfamily J, RAR, JMJD6, ESR1 (nuclear), RARalpha, COL4A4, c-Myc, ESR1 (mitochondrial), ESR, Galpha(q)-specific peptide GPCRs, Actin muscle, Actin, COL4A3, Collagen IV, ESR1 (membrane)
2	baculum development	1.20E-10	2.14E-07	ESR1 (nuclear), ESR1 (mitochondrial), ESR, ESR1 (membrane)
3	cellular response to estrogen stimulus	3.37E-10	4.00E-07	RAR, ESR1 (nuclear), RARalpha, c-Myc, ESR1 (mitochondrial), ESR, ESR1 (membrane)
4	prostate epithelial cord elongation	4.17E-09	3.71E-06	ESR1 (nuclear), ESR1 (mitochondrial), ESR, ESR1 (membrane)
5	regulation of phospholipase activity	5.91E-09	3.71E-06	AGTR1, ESR1 (nuclear), PLC-delta, ESR1 (mitochondrial), ESR, Galpha(q)-specific peptide GPCRs, PLA2, ESR1 (membrane)
6	intracellular estrogen receptor signaling pathway	6.70E-09	3.71E-06	RAR, ESR1 (nuclear), RARalpha, ESR1 (mitochondrial), ESR, ESR1 (membrane)
7	Sertoli cell differentiation	7.74E-09	3.71E-06	RAR, ESR1 (nuclear), RARalpha, ESR1 (mitochondrial), ESR, ESR1 (membrane)
8	prostate gland morphogenetic growth	8.31E-09	3.71E-06	ESR1 (nuclear), ESR1 (mitochondrial), ESR, ESR1 (membrane)
9	kidney development	1.43E-08	5.51E-06	AGTR1, K(+) channel, subfamily J, RAR, JMJD6, RARalpha, COL4A4, c-Myc, Galpha(q)-specific peptide GPCRs, Actin muscle, Actin, COL4A3, Collagen IV
10	tissue morphogenesis	1.54E-08	5.51E-06	FGF3, RAR, ESR1 (nuclear), RARalpha, ACTC, c-Myc, ESR1 (mitochondrial), ESR, Galpha(q)-specific peptide GPCRs, CARP, HB-EGF, Actin muscle, TGM3, Actin, Collagen IV, ESR1 (membrane)
11	prostate epithelial cord arborization involved in prostate glandular acinus morphogenesis	2.48E-08	6.59E-06	ESR1 (nuclear), ESR1 (mitochondrial), ESR, ESR1 (membrane)
12	prostate glandular acinus morphogenesis	2.48E-08	6.59E-06	ESR1 (nuclear), ESR1 (mitochondrial), ESR, ESR1 (membrane)
13	antral ovarian follicle growth	2.48E-08	6.59E-06	ESR1 (nuclear), ESR1 (mitochondrial), ESR, ESR1 (membrane)
14	response to estrogen	2.58E-08	6.59E-06	AGTR1, RAR, ESR1 (nuclear), RARalpha, c-Myc, ESR1 (mitochondrial), ESR, Galpha(q)-specific peptide GPCRs, ESR1 (membrane)
15	renal system development	2.93E-08	6.98E-06	AGTR1, K(+) channel, subfamily J, RAR, JMJD6, RARalpha, COL4A4, c-Myc, Galpha(q)-specific peptide GPCRs, Actin muscle, Actin, COL4A3, Collagen IV
16	regulation of lipase activity	4.17E-08	9.30E-06	AGTR1, ESR1 (nuclear), PLC-delta, ESR1 (mitochondrial), ESR, Galpha(q)-specific peptide GPCRs, PLA2, ESR1 (membrane)
17	positive regulation of phospholipase activity	4.71E-08	9.69E-06	AGTR1, ESR1 (nuclear), ESR1 (mitochondrial), ESR, Galpha(q)-specific peptide GPCRs, PLA2, ESR1 (membrane)
18	regulation of hydrolase activity	4.89E-08	9.69E-06	IRE1, ERN2, AGTR1, FGF3, ESR1 (nuclear), PLC-delta, c-Myc, ESR1 (mitochondrial), Neuregulin 2, NUR77, ESR, Galpha(q)-specific peptide GPCRs, Dynein, axonemal, light chains, Rab-3A, HB-EGF, PLA2, Rab-3, COL4A3, FGF4, Collagen IV, NOL3, ESR1 (membrane), LRRC68
19	epithelial cell proliferation involved in mammary gland duct elongation	5.82E-08	1.09E-05	ESR1 (nuclear), ESR1 (mitochondrial), ESR, ESR1 (membrane)
20	negative regulation of triglyceride metabolic process	6.10E-08	1.09E-05	ESR1 (nuclear), ESR1 (mitochondrial), ESR, SIK, ESR1 (membrane)

Table S8. The top gene set enrichment results for BaP treated liver slices.



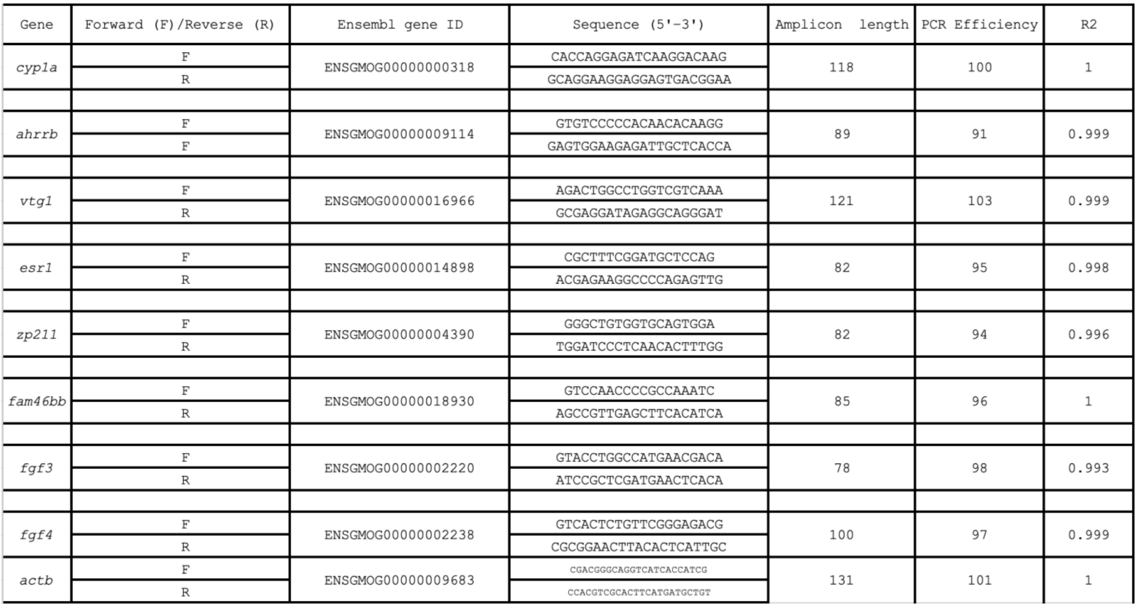
A) KEGG gene set database HALLMARK geneset database

NAME	ES	NES	NOM p-val	FDR q-val
KEGG_RETINOL_METABOLISM	0.5850391	1.8692338	0.00409836	0.02922866
KEGG_TRYPTOPHAN_METABOLISM	0.6219644	1.8705419	0.000	0.05745732
KEGG_ARACHIDONIC_ACID_METABOLISM	0.58533454	1.7408818	0.02380952	0.09231287
KEGG_STEROID_HORMONE_BIOSYNTHESIS	0.5928773	1.6989107	0.000	0.11772954
KEGG_AMINO_SUGAR_AND_NUCLEOTIDE_SUGAR_METABOLISM	0.46320933	1.5087327	0.08817636	0.5851132
KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450	0.5882661	1.5226316	0.05988024	0.6245114
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	0.5164979	1.4235245	0.07628866	0.94578785
KEGG_DRUG_METABOLISM_OTHER_ENZYMES	0.47383246	1.400792	0.10882957	0.95681155
KEGG_BUTANOATE_METABOLISM	0.43917245	1.3339406	0.14197531	1.000
KEGG_BETA_ALANINE_METABOLISM	0.39702874	1.3194307	0.10691824	1.000
KEGG_HEMATOPOIETIC_CELL_LINEAGE	0.44237944	1.3096193	0.16039604	1.000
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	0.37549844	1.3092757	0.11377245	1.000
KEGG_CITRATE_CYCLE_TCA_CYCLE	0.38653418	1.2847354	0.15895373	1.000
KEGG_CARDIAC_MUSCLE_CONTRACTION	0.38801676	1.2503598	0.26096034	1.000
KEGG_PYRIMIDINE_METABOLISM	0.40331286	1.2302177	0.21666667	1.000
KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM	0.44668594	1.2006208	0.26987448	1.000
KEGG_OXIDATIVE_PHOSPHORYLATION	0.5232972	1.1994327	0.37629938	1.000
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	0.3909424	1.1751885	0.2826087	1.000
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	0.34135532	1.1643286	0.28880158	1.000
KEGG_PURINE_METABOLISM	0.2750763	1.1609008	0.22244489	1.000

B) HALLMARK gene set database

NAME	ES	NES	NOM p-val	FDR q-val
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.5280985	1.7563628	0.028	0.06396826
HALLMARK_UV_RESPONSE_UP	0.40236568	1.7324567	0.01006036	0.04276865
HALLMARK_APOPTOSIS	0.36156875	1.5444732	0.01659751	0.2613527
HALLMARK_PANCREAS_BETA_CELLS	0.4361004	1.3359778	0.126	0.9758395
HALLMARK_P53_PATHWAY	0.31541044	1.3342699	0.09876543	0.7879834
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.40299618	1.3219451	0.13654618	0.7004988
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	0.37995547	1.2978796	0.15587045	0.67918974
HALLMARK_XENOBIOTIC_METABOLISM	0.29518342	1.2847319	0.1388889	0.6279289
HALLMARK_MTORC1_SIGNALING	0.31725547	1.2703227	0.17805383	0.5978452
HALLMARK_INFLAMMATORY_RESPONSE	0.33884475	1.2252626	0.22983871	0.65531385
HALLMARK_MYC_TARGETS_V1	0.40268388	1.1964682	0.34623218	0.6692649
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.31125537	1.1889946	0.23481782	0.63170826
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.29837415	1.176772	0.24180327	0.60589296
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.4235029	1.1440756	0.36082473	0.6340766
HALLMARK_FATTY_ACID_METABOLISM	0.30253232	1.1182386	0.2875	0.65547204
HALLMARK_KRAS_SIGNALING_UP	0.2843483	1.111438	0.322	0.6275709
HALLMARK_ADIPOGENESIS	0.23352304	1.1080488	0.2860215	0.59785694
HALLMARK_GLYCOLYSIS	0.22547595	1.0414429	0.37896827	0.70166606
HALLMARK_DNA_REPAIR	0.3288499	0.9959063	0.4831014	0.75872326
HALLMARK_COAGULATION	0.23406403	0.97859734	0.5	0.7574598

Table S9. qPCR primer sequences, amplicon lengths and PCR efficiencies.



Gene	Forward (F)/Reverse (R)	Ensembl gene ID	Sequence (5'-3')	Amplicon length	PCR Efficiency	R2
cyp1a	F	ENSGMOG00000000318	CACCAGGAGATCAAGGACAAG	118	100	1
	R		GCAGGAAGGAGGAGTGACGGAA			
ahrrb	F	ENSGMOG00000009114	GTGTCCCCCACAACACAAGG	89	91	0.999
	F		GAGTGGAAGAGATTGCTCACCA			
vtg1	F	ENSGMOG00000016966	AGACTGGCCTGGTCGTCAAA	121	103	0.999
	R		GCGAGGATAGAGGCAGGGAT			
esr1	F	ENSGMOG00000014898	CGCTTTCGGATGCTCCAG	82	95	0.998
	R		ACGAGAAGGCCCCAGAGTTG			
zp2l1	F	ENSGMOG00000004390	GGGCTGTGGTGCAGTGGA	82	94	0.996
	R		TGGATCCCTCAACACTTTGG			
fam46bb	F	ENSGMOG00000018930	GTCCAACCCCGCCAAATC	85	96	1
	R		AGCCGTTGAGCTTCACATCA			
fgf3	F	ENSGMOG00000002220	GTACCTGGCCATGAACGACA	78	98	0.993
	R		ATCCGCTCGATGAACTCACA			
fgf4	F	ENSGMOG00000002238	GTCACTCTGTTCGGGAGACG	100	97	0.999
	R		CGCGGAACTTACACTCATTGC			
actb	F	ENSGMOG00000009683	CGACGGGCAGGTCATCACCATCG	131	101	1
	R		CCACGTCGCACTTCATGATGCTGT			

Paper II

RASflow: an RNA-Seq analysis workflow with Snakemake

Xiaokang Zhang, Inge Jonassen
BMC Bioinformatics, **21/110** (2020)

Zhang and Jonassen *BMC Bioinformatics* (2020) 21:110
https://doi.org/10.1186/s12859-020-3433-x

BMC Bioinformatics

SOFTWARE Open Access

RASflow: an RNA-Seq analysis workflow with Snakemake

Xiaokang Zhang and Inge Jonassen*

Abstract

Background: With the cost of DNA sequencing decreasing, increasing amounts of RNA-Seq data are being generated giving novel insight into gene expression and regulation. Prior to analysis of gene expression, the RNA-Seq data has to be processed through a number of steps resulting in a quantification of expression of each gene/transcript in each of the analyzed samples. A number of workflows are available to help researchers perform these steps on their own data, or on public data to take advantage of novel software or reference data in data re-analysis. However, many of the existing workflows are limited to specific types of studies. We therefore aimed to develop a maximally general workflow, applicable to a wide range of data and analysis approaches and at the same time support research on both model and non-model organisms. Furthermore, we aimed to make the workflow usable also for users with limited programming skills.

Results: Utilizing the workflow management system Snakemake and the package management system Conda, we have developed a modular, flexible and user-friendly RNA-Seq analysis workflow: RNA-Seq Analysis Snakemake Workflow (RASflow). Utilizing Snakemake and Conda alleviates challenges with library dependencies and version conflicts and also supports reproducibility. To be applicable for a wide variety of applications, RASflow supports the mapping of reads to both genomic and transcriptomic assemblies. RASflow has a broad range of potential users: it can be applied by researchers interested in any organism and since it requires no programming skills, it can be used by researchers with different backgrounds. The source code of RASflow is available on GitHub: https://github.com/zhxiaokang/RASflow.

Conclusions: RASflow is a simple and reliable RNA-Seq analysis workflow covering many use cases.

Keywords: RNA-Seq, Workflow, Snakemake

Background

RNA sequencing (RNA-Seq) was introduced more than ten years ago and has become one of the most important tools to map and identify genes and understand their regulation and roles across species [1, 2]. A large number of studies have been performed using RNA-Seq and resulted in gene expression datasets available in databases such as GEO [3] and ArrayExpress [4]. Underlying reads are typically deposited to the Sequence Read Archive (SRA) [5], currently containing reads for more than 1,7 million samples (https://www.ncbi.nlm.nih.gov/sra/?term=RNA-Seq). One of the most popular applications of RNA-Seq is for Differential Expression Analysis (DEA) where one identifies genes that are expressed at different levels between two classes of samples (e.g., healthy, disease) [6].

When RNA-Seq is used in a DEA project, the sequencing reads need to be taken through several steps of processing and analysis. Often, the steps are organized into a workflow that can be executed in a fully or partially automated fashion. The steps include: quality control (QC) and trimming, mapping of reads to a reference genome (or transcriptome), quantification on gene (or transcript) level, statistical analysis of expression statistics to report genes (or transcripts) being differentially

*Correspondence: inge.jonassen@uib.no
Computational Biology Unit, Department of Informatics, University of Bergen, Thormohlens Gate 55, 5009 Bergen, Norway

expressed between two predefined sets of samples, along with associated *P*-values or False Discovery Rate (FDR) values. Aligning reads to the genome is the most computationally intensive and time-consuming step. An alternative approach is to perform a pseudo alignment to a transcriptome. This has gained more popularity recently, due to its high speed and high accuracy [7–9]. It has been shown that lightweight pseudo alignment improves gene expression estimation and at the same time is computationally more efficient, compared with the standard alignment/counting methods [10]. But if the purpose of analysis is to call genomic variants, then it is still better to map the reads to the genome [11]. Considering this, a workflow should provide both quantification strategies to satisfy users with different research interests.

There is a large number of RNA-Seq analysis workflows and many have been published and made available to the user community. We reviewed seven workflows published in the past three years [12–18] (see "Discussion" section for more details). We found that none of these workflows cover all the needs outlined above while also being usable for less computer fluent users. So more complete and easy-to-use workflows are still needed.

In this article, we present RNA-Seq Analysis Snakemake Workflow (RASflow) that is usable for a wide range of applications. RASflow can be applied to data from any organism and can map reads to either a genome or a transcriptome, allowing the user to refer to public databases such as ENSEMBL [19] or to supply their own genomes or transcriptomes [20, 21]. The latter can for example be useful for projects on non-model species for which there is no public high-quality reference genome/transcriptome. RASflow is scalable: it can be run on either supercomputers with many cores (which enable parallel computing) or on a personal computer with limited computing resources; it can process data from hundreds of samples and still consumes very little storage space because it temporarily copies or downloads the FASTQ file(s) of one sample (one file for single end and two files for pair end) to the working directory at the time, and it stores only the necessary intermediate and final outputs. Using Conda [22], the whole workflow with all dependencies (version already specified) can be installed simply with one single command in a virtual environment. This ensures quick and smooth installation. Using Snakemake [23], the whole analysis is completely reproducible and highly user-friendly also for users with limited programming skills. In the DEA step, RASflow supports use of paired tests that can help to strengthen the statistical power and bring out expression differences related to the phenomenon under study [24].

Implementation

Figure 1 shows a schematic representation of the RASflow workflow. It starts with performing QC of the raw FASTQ files using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The QC report is presented to the user along with a question of whether the reads should be trimmed. When opted for, trimming is performed using the tool Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and subsequently, an additional QC report is generated.

When the user is satisfied with the quality of the reads, the workflow proceeds to the next step: quantification of read abundance or expression level for transcripts or genes. The user decides whether to map the reads to a transcriptome or a genome depending on the goal of the analysis and availability of data. If the purpose of the analysis is to identify differentially expressed genes, it is suggested to map the reads to a transcriptome using pseudo alignment with Salmon [9]. A quantification table of the transcripts is generated from this step. Alternatively, the user can choose for the reads to be mapped to a genome. The aligner used in RASflow is HISAT2 [25] which has relatively modest memory requirements (∼4.3GB for the human genome) compared with for example the STAR aligner (requiring ∼27GB for the human genome) [26]. The alignment step is followed by a quality evaluation performed by Qualimap2 [27] and feature counting done by featureCounts [28] or htseq-count [29]. To be noted, after most of the steps, a summary report is generated using MultiQC [30].

When a quantification matrix for the genes/transcripts has been produced, RASflow can proceed to perform a DEA analysis using edgeR [31, 32] or DESeq2 [33]. RASflow supports both single and paired statistical tests. The user specifies which statistical test mode to be applied in the configuration file based on their experimental design. If the reads were mapped to a transcriptome, DEA will be done on both transcript- and gene-level. In any case, the outputs of DEA include three types of tables: normalized quantification tables, some important statistics for the whole gene or transcript list, and the list of significantly differentially expressed genes or transcripts (with default threshold of $FDR < 0.05$). The raw count is normalized based on Trimmed Mean of M values (TMM) [34] (if edgeR is used) or the median-of-ratios method [35] (if DESeq2 is used) when the reads are mapped to a genome. But if the reads are mapped to a transcriptome, the normalized values are estimated Transcripts Per Million (TPM) from Salmon scaled using the average transcript length over samples and then the library size by "tximport" [36]. The results of DEA is also visualized with a volcano plot enabling visual identification of genes with high fold change whose differential expression is also statistically significant, and a heatmap that not only

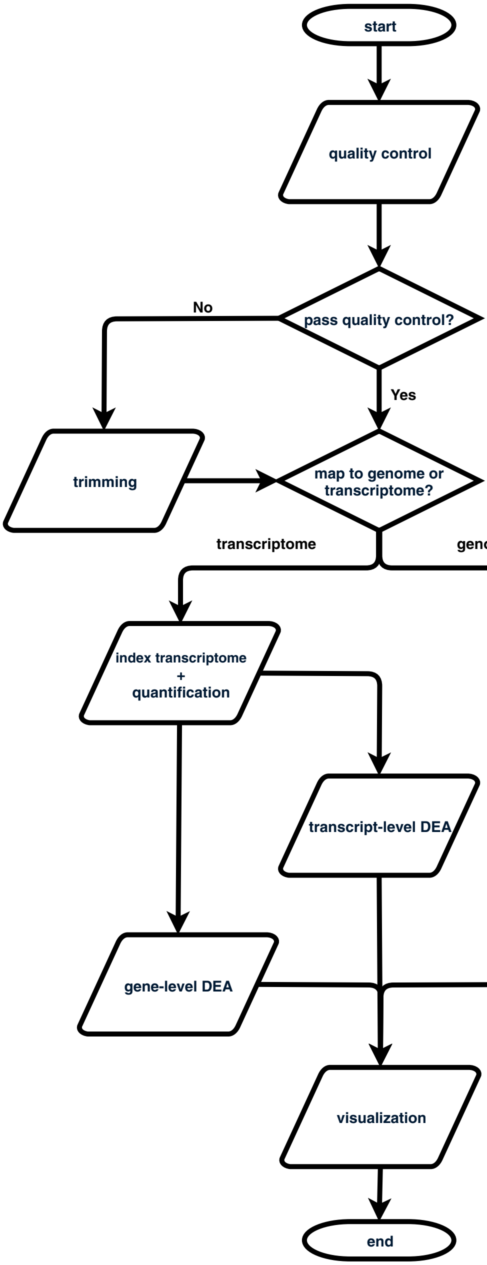


Fig. 1 Overview of the steps performed by RNA-Seq Analysis Snakemake Workflow (RASflow)

visualizes the expression pattern of the identified differentially expressed genes, but also a clustering of the samples based on those genes, so that the user can get an idea of how well separated the groups are.

To ensure smooth installation and reproducibility of the workflow, all the tools included are fixed to a specific version which can be found in the environment configuration file (env.yaml).

Results

To show users how RASflow works and to familiarize them with RASflow, we provide some small example datasets. They are generated as subsets of the original real data [37]. The figures in this section were generated by RASflow using the example data as input. RASflow was also tested on four real datasets: pair-end RNA-seq of prostate cancer and adjacent normal tissues from 14 patients (ArrayExpress accession: E-MTAB-567) [38], single-end RNA-Seq of mesenchymal stem cells (MSCs) and cancer-associated fibroblasts (CAFs) from EG7 tumor-bearing mice (GEO accession: GSE141199), pair-end RNA-Seq of Atlantic cod liver slices exposed to benzo[a]pyrene (BaP) and 17α-ethynylestradiol (EE2) (GEO accession: GSE106968) [39], and a benchmarking dataset, single-end RNA-Seq of highly purified human classical and nonclassical monocyte subsets from a clinical cohort (SRA accession: SRP082682) [40].

The output of the example dataset can be found on the GitHub page of RASflow and an overview of the output folder is shown in Additional file 1: Fig. S1. The output of the four real datasets can be found here: https://git.app.uib.no/Xiaokang.Zhang/rasflow_realdata.

Quality control of raw reads and alignments

FastQC checks the quality of the sequencing reads and produces one report for each FASTQ file. MultiQC is used to summarize all the reports and merge them into one document, as shown in Fig. 2a and b. Users are asked to check the report and decide whether trimming is needed. If the quality of the reads is good enough, it is recommended that trimming should not be performed since it would lead to loss of information; but if the quality is low, trimming is suggested to improve the quality. The raw reads quality of the human prostate dataset is not good enough and trimming was therefore performed. The QC reports of raw reads and trimmed reads can be found in Additional file 2: Fig. S2.

After the alignment to the genome, the intermediate output, the BAM files, will be provided to Qualimap2 to evaluate the alignment quality. Figure 2c shows an example report from Qualimap2.

MultiQC is used to generate a report on the mapping ratios using the output of feature counting (Fig. 2d).

Quantification of transcripts or genes

If a transcriptome was used as mapping reference, a file containing the estimated relative abundance and length of the target transcript is generated for each sample. If the reads were aligned to a genome, the direct outputs from alignment are genes' raw count tables for each sample.

Differential expression analysis

In the first step, the user-specified information on sample groups is used to produce one count or abundance file for each group. The raw count or abundance in those files is then normalized by either edgeR or DESeq2 generating a corresponding file for each of them. When a transcriptome is used as mapping reference, depending on user parameters, gene-level raw and normalized abundance can also be generated, and the downstream DEA will also be done on both transcript- and gene-level.

During DEA, a statistical test is performed on the raw abundance (both edgeR and DESeq2 prefer raw other than normalized abundance) tables of transcripts/genes. The result includes important statistics such as Log Fold Change, false discovery rates (FDRs) or adjusted *P*-value for each transcript/gene. With a predefined threshold of FDR (default value is 0.05), the transcripts/genes with a lower FDR are reported as significantly differentially expressed, and they are included in a second table. Besides the tables mentioned above, DEA also generates visualizations including a volcano plot (Fig. 3a) and a heatmap (Fig. 3b).

Williams et al. evaluated hundreds of combinatorial implementations of the most commonly used tools for their impact on DEA results, and they concluded that the method of differential expression analysis exhibited the strongest impact compared with the choice of tools in the other steps [40]. We have evaluated RASflow on the benchmarking dataset they generated using both the transcriptome and the genome as mapping reference, and in both cases, DESeq2 has a higher recall and edgeR has a higher precision, meaning that edgeR is more conservative in reporting a gene as differentially expressed in this study case. The differentially expressed gene list of each workflow and their performance, including values and ranks for recall and precision against the evaluated workflows in [40], can be found in Additional file 3.

Runtime

The most time-consuming part of the whole workflow is the alignment step. As already mentioned, pseudo alignment to a transcriptome is much faster than alignment to a genome. RASflow was run on four real datasets using a 1TB RAM 60 cores Dell PowerEdge R910 machine and the runtime is shown in Table 1. RASflow was also tested on the mouse dataset using Windows Subsystem for Linux on an 8GB RAM 4 cores Intel Core 2 machine,

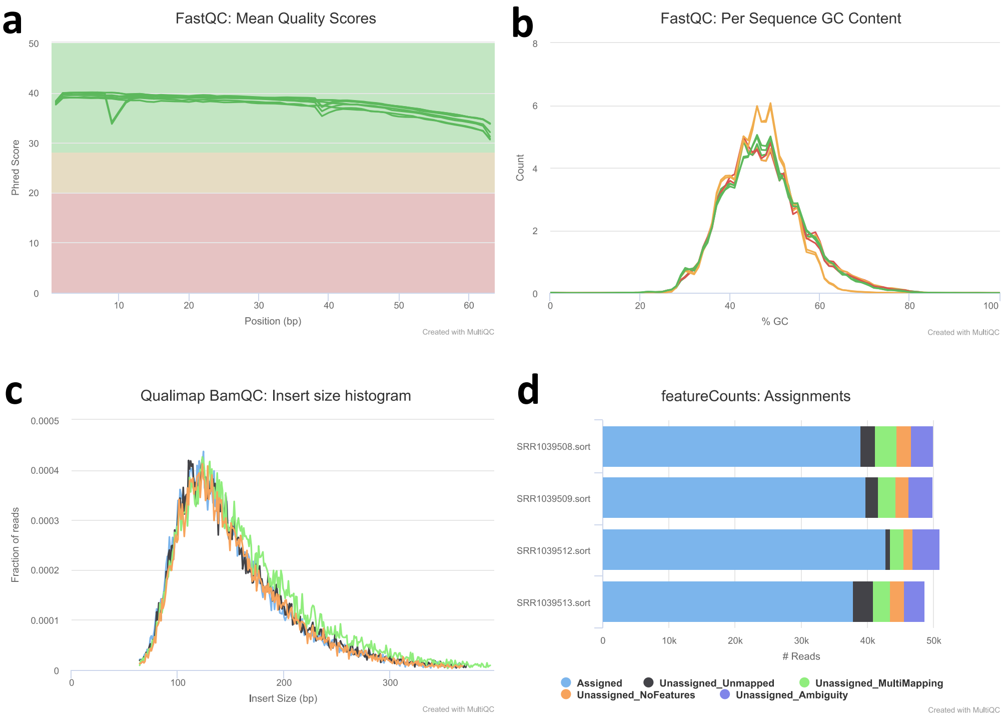


Fig. 2 Quality control of raw reads and alignment. **a** The mean quality value across each base position in the read. **b** The average GC content of reads. A normal random library typically has a roughly normal distribution of GC content. **c** Distribution of estimated insert sizes of mapped reads. **d** A brief mapping summary

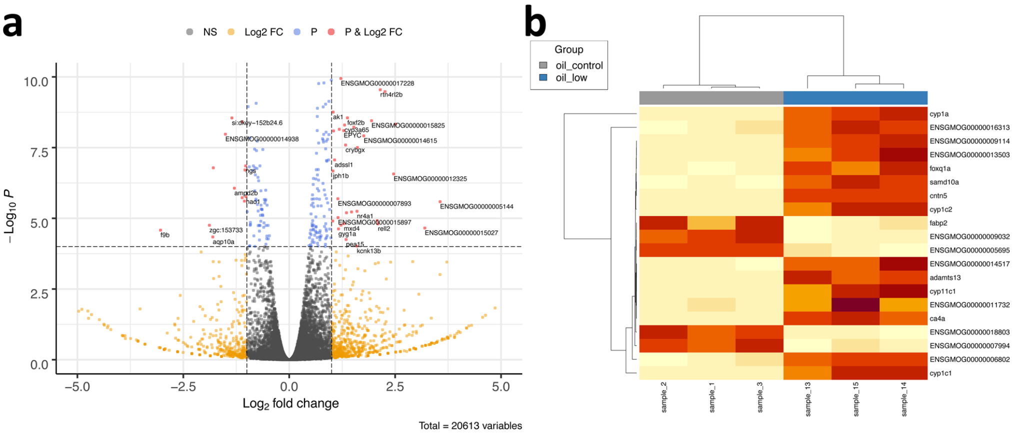


Fig. 3 Visualization of DEA results. **a** Volcano plot with labeled genes who pass the thresholds of both Fold Change and *P*-value. **b** Hierarchical clustering heatmap with samples along the x-axis and differentially expressed genes along the y-axis

Table 1 Alignment runtime of three datasets

Dataset	Number of samples	Size of raw data (GB)	Runtime of alignment (HH:MM)	
			Transcriptome as reference	Genome as reference
Cod	47	244	05:32	69:18
Human	28	137	03:14	20:03
Benchmark	32	36	02:37	11:22
Mouse	8	9.3	00:28	03:46
Mouse_pc*	8	9.3	01:11	19:31

*This was run on a personal computer

and the runtime is shown in Table 2. As Table 1 shows, alignment using a genome as reference takes much longer than using a transcriptome, especially when the dataset is large (datasets “Cod" and “Human") or the job is run on a personal computer (dataset “Mouse_pc").

Discussion

Virtual environment by Conda

The whole workflow is installed and run in a virtual environment created by Conda. While creating the virtual environment, all dependencies using the specified versions are installed at once. This ensures not only the smooth installation and running of RASflow, but also a reproducible analysis independent of the operating system and machine.

Snakemake as framework

Snakemake is a scalable workflow engine that helps to manage workflows in an easy way. It divides the whole workflow into rules with each rule accomplishing one step of the workflow. The input of one rule is the output from the rule corresponding to the previous step, making the dataflow easy to track. Thanks to this logic, the whole workflow becomes highly modular, so users can easily expand the workflow or replace part of it, also for complicated workflows.

RASflow organizes the rules carrying out one big step of the workflow in one file (with extension .rules). All the files are then integrated into one main file (main.py). For the users who are satisfied with RASflow’s default setting, they can manage the workflow simply through the configuration file to tell RASflow which pipeline and which tools they want to use. Advanced users may change the settings and parameters in the .rules files and may also substitute tools for example to try out new methods as they are published.

Transciptome and genome as reference

RASflow allows users to supply their own genomic or transcriptomic reference. This enables users to study expression in species where no public reference is available or the users have alternative references that they wish to utilize. It should be noted that if one aims for transcript-level analysis, a transcriptome should be used as reference.

But some analyses other than DEA require the reads to be mapped to a genome and gene-level DEA is more robust and experimentally actionable, so RASflow still

Table 2 Comparison of RASflow with the other workflows published between 2017 and 2019

workflow	quality control	organism	mapping reference	workflow for DEA*	hardware requirement	installation	programming requirement	year	ref
RASflow	yes	all	genome transcriptome	GB & TB	low	easy	low	2020	NA
UTAP	yes	5	genome	GB	high	easy	low	2019	[12]
ARMOR	yes	all	genome transcriptome	TB	high	easy	low	2019	[13]
VIPER	yes	2	genome	GB	high	easy	low	2018	[14]
BioJupies	no	2	genome	GB	low	web application	low	2018	[15]
hppRNA	yes	2	genome transcriptome	GB & TB	low	medium	medium	2018	[16]
aRNApipe	yes	all	genome	GB	high	hard	high	2017	[17]
RNACocktail	no	all	genome transcriptome	GB & TB	low	hard	high	2017	[18]

*GB: genome based — gene/transcript quantification and DEA based on reads mapped to a genome; TB: transcriptome based provides the traditional workflow of genome alignment and DEA based on gene counts.

Comparison with other tools

We compared RASflow to other existing workflows as shown in Table 2. As we can see from the table, some workflows do not include QC steps [15, 18]. Some of the workflows are limited to specific organisms typically human or mouse and in some cases other model organisms [12, 14–16]. Some of them have functionality only for mapping reads to a reference genome and do not support the use of a transcriptome reference [12, 14, 15, 17]. ARMOR includes both genome and transcriptome as mapping reference but does not support genome-based quantification of expression and subsequent DEA.

Considering hardware requirement, BioJupies is marked as "low" because it is a web application and the compute capacity is offered on the server side. The workflows marked with "high" use STAR for genome alignment which requires about 27GB of RAM to align reads to the human genome. hppRNA and RNACocktail support both STAR and other aligners which require comparably low RAM, such as HISAT2 which is used in RASflow. Tests performed show that RASflow can be used to run human genome alignment smoothly on a personal computer with only 8GB of RAM.

As for workflow installation, RASflow, UTAP, ARMOR, and VIPER all use Conda to create a virtual environment and to install the required software, making workflow installation easy and robust. hppRNA provides scripts to automatically install all the required software but as it is not done through the use of a virtual environment, some software may conflict with software already installed on the machine. The aRNApipe and RNACocktail workflows require the user to install all the software manually which is time-consuming and can also easily lead to version conflicts.

After installation, executing the workflow can also present challenges. In order to use the aRNApipe and RNACocktail workflows on their own data, the user needs to know programming very well. The hppRNA workflow comes with a very detailed and useful manual for the user to follow which helps a lot. The UTAP and BioJupies workflows both provide graphical user interfaces and can be used without any programming skills. While the remaining workflows do not provide graphical interfaces, they use Snakemake to manage all the steps in the workflow, making them easy to use also for those with limited programming skills.

Extension of RASflow

Thanks to the high modularity of RASflow, it is very easy to exchange the tools applied in RASflow with other tools if they are more appropriate for specific research interest or they are newly developed. Thanks to the feedback from users, we have already added the htseq-count tool for feature counting and the DESeq2 tool for DEA as extra options since the first version of RASflow. Advanced users can also do this by themselves without much effort. We welcome any feedback and contribution through GitHub page to improve RASflow.

RASflow can also be extended to realize other functions, such as Single Nucleotide Variant (SNV) detection, pathway analysis, and so on.

Conclusions

RASflow is a light-weight and easy-to-manage RNA-Seq analysis workflow. It includes the complete workflow for RNA-Seq analysis, starting with QC of the raw FASTQ files, going through optional trimming, alignment and feature counting (if the reads are mapped to a genome), pseudo alignment (if transcriptome is used as mapping reference), gene- or transcript- level DEA, and visualization of the output from DEA.

RASflow is designed in such a way that it can be applied by a wide range of users. It requires little programming skills and a well-written tutorial helps users go through the whole workflow making it very easy to set up and run RASflow from scratch. RASflow has low hardware requirements so that it can be run on almost any personal computer. It can also be scaled up to make full use of the computing power of a supercomputer or cluster. RASflow can be applied to data of any organism and the user can choose to map the reads to a transcriptome or a genome. It also supports the use of user-supplied transcriptome or genome references.

RASflow is built on the basis of Conda and Snakemake, making installation and management very easy. All the required tools are available on the Anaconda cloud (https://anaconda.org/) and are wrapped in a virtual environment managed by Conda, making RASflow independent of the underlying system thus avoiding package/library version conflicts. The whole workflow is defined by rules managed by Snakemake, which makes it highly modular. This means that the advanced users can easily extract parts of the workflow or expand it based on their own research needs, and replace the tools used in RASflow with other tools to explore new pipelines for analyzing RNA-Seq data.

Availability and requirements

Project name: RASflow.
Project home page: https://github.com/zhxiaokang/RASflow
Operating system(s): Linux, macOS and Windows.

Programming language: Python, R, Shell
Other requirements: Conda
License: MIT License
Any restrictions to use by non-academics: N/A.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3433-x.

Additional file 1: Figure S1. An overview of output folder of example data.

Additional file 2: Figure S2. (a) The mean quality scores of raw reads from human prostate cancer data. (b) The mean quality scores of trimmed reads from human prostate cancer data.

Additional file 3: Tables of differentially expressed gene lists of RASflow using both the transcriptome and the genome as mapping reference and using DESeq2 and edgeR as differential expression analysis methods and their performance.

Abbreviations

DEA: Differential expression analysis; FDR: False discovery rate; QC: Quality control; RASflow: RNA-Seq analysis Snakemake workflow; RNA-Seq: RNA sequencing; SNV: Single nucleotide variant; SRA: Sequence read archive; TMM: Trimmed mean of M values; TPM: Transcript per million

Acknowledgements

We would like to thank the colleagues in dCod 1.0 project for which the workflow was initially designed. The feedback from the biologists in the project greatly helped the improvement of the workflow. We would also like to thank the research school NORBIS where XZ took the training on analysis of next generation sequencing.

Authors' contributions

XZ designed and developed RASflow and wrote the tutorial. XZ wrote the initial draft of the manuscript. IJ supervised the work and finalized the manuscript. All authors have read and approved the final manuscript.

Funding

The dCod 1.0 project is funded under the Digital Life Norway initiative of the BIOTEK 2021 program of the Research Council of Norway (project no. 248840). The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All the datasets and source codes are available on GitHub.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 25 November 2019 Accepted: 26 February 2020
Published online: 18 March 2020

References

1. Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing,. Genome Res. 2007;17(1):69–73. https://doi.org/10.1101/gr.5145806.
2. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell. 2008;133(3):523–36. https://doi.org/10.1016/J.CELL.2008.03.029.
3. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets–update,. Nucleic Acids Res. 2013;41(Database issue):991–5. https://doi.org/10.1093/nar/gks1193.
4. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, Sarkans U, Brazma A. ArrayExpress update - From bulk to single-cell expression data. Nucleic Acids Res. 2019;47(D1):711–5. https://doi.org/10.1093/nar/gky964.
5. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011;39(Database):19–21. https://doi.org/10.1093/nar/gkq1019.
6. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 20191–26. https://doi.org/10.1038/s41576-019-0150-2.
7. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32(5):462–4. https://doi.org/10.1038/nbt.2862.
8. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7. https://doi.org/10.1038/nbt.3519.
9. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9. https://doi.org/10.1038/nmeth.4197.
10. Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol. 2015;16(1):177. https://doi.org/10.1186/s13059-015-0734-x.
11. Adetunji MO, Lamont SJ, Abasht B, Schmidt CJ. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. PLoS ONE. 2019;14(9):0216838. https://doi.org/10.1371/journal.pone.0216838.
12. Kohen R, Barlev J, Hornung G, Stelzer G, Feldmesser E, Kogan K, Safran M, Leshkowitz D. UTAP: User-friendly Transcriptome Analysis Pipeline. BMC Bioinformatics. 2019;20(1):154. https://doi.org/10.1186/s12859-019-2728-2.
13. Orjuela S, Huang R, Hembach KM, Robinson MD, Soneson C. ARMOR: an Automated Reproducible MOdular workflow for preprocessing and differential analysis of RNA-seq data. G3: Genes, Genomes, Genetics. 2019. https://doi.org/10.1534/g3.119.400185.
14. Cornwell M, Vangala M, Taing L, Herbert Z, Köster J, Li B, Sun H, Li T, Zhang J, Qiu X, Pun M, Jeselsohn R, Brown M, Liu XS, Long HW. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinformatics. 2018;19(1):135. https://doi.org/10.1186/s12859-018-2139-9.
15. Torre D, Lachmann A, Ma'ayan A. BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. Cell Syst. 2018;7(5):556–5613. https://doi.org/10.1016/j.cels.2018.10.007.
16. Wang D. hppRNA—a Snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples. Brief Bioinforma. 2017;19(4):143. https://doi.org/10.1093/bib/bbw143.
17. Alonso A, Lasseigne BN, Williams K, Nielsen J, Ramaker RC, Hardigan AA, Johnston B, Roberts BS, Cooper SJ, Marsal S, Myers RM. aRNApipe: A balanced, efficient and distributed pipeline for processing RNA-seq data in high performance computing environments. Bioinformatics. 2017;33(11):023. https://doi.org/10.1093/bioinformatics/btx023.
18. Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, Bani Asadi N, Gerstein MB, Wong WH, Snyder MP, Schadt E, Lam HYK. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun. 2017;8(1):59. https://doi.org/10.1038/s41467-017-00050-4.
19. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Girón CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Laird MR, Lavidas I, Liu Z, Loveland JE, Marugán JC, Maurel T, McMahon AC, Moore B, Morales J, Mudge JM, Nuhn M, Ogeh D, Parker A, Parton A, Patricio M, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sparrow H, Stapleton E, Szuba M, Taylor K, Threadgold G, Thormann A, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Yates AD, Zerbino DR, Flicek P. Ensembl 2019. Nucleic Acids Res. 2019;47(D1):745–51. https://doi.org/10.1093/nar/gky1113.

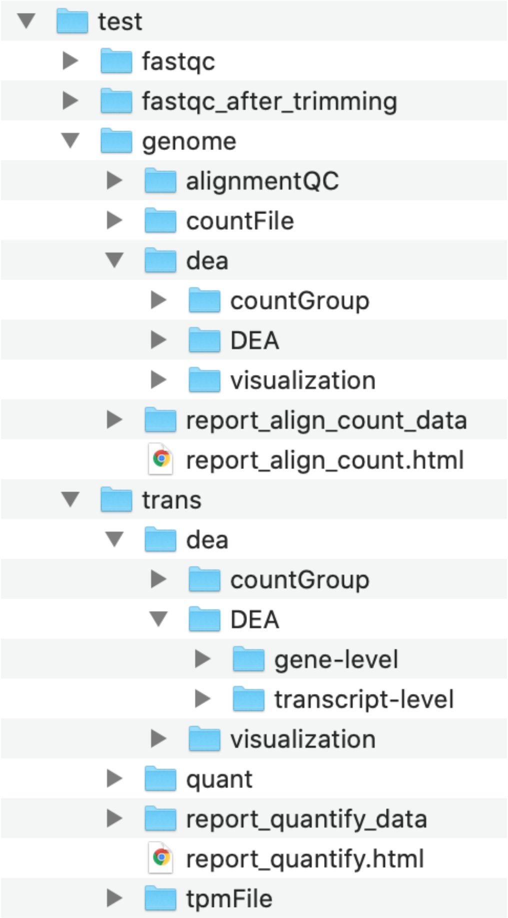
20. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95(6):315–27. https://doi.org/10.1016/J.YGENO.2010.03.001.
21. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14(3):157–67. https://doi.org/10.1038/nrg3367.
22. Analytics C. Anaconda software distribution. Comput Softw Vers. 20162.
23. Koster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2. https://doi.org/10.1093/bioinformatics/bts480.
24. Mcdonald JH. Handbook of Biological Statistics. Baltimore: Sparky House Publishing; 2009, pp. 6–59. http://www.biostathandbook.com.
25. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60. https://doi.org/10.1038/nmeth.3317.
26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
27. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2015;32(2):566. https://doi.org/10.1093/bioinformatics/btv566.
28. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.
29. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9. https://doi.org/10.1093/bioinformatics/btu638.
30. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8. https://doi.org/10.1093/bioinformatics/btw354.
31. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
32. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97. https://doi.org/10.1093/nar/gks042.
33. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
34. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):25. https://doi.org/10.1186/gb-2010-11-3-r25.
35. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106. https://doi.org/10.1186/gb-2010-11-10-r106.
36. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4:1521. https://doi.org/10.12688/f1000research.7563.2.
37. Himes BE, Jiang X, Wagner P, Hu R, Wang Q, Klanderman B, Whitaker RM, Duan Q, Lasky-Su J, Nikolos C, Jester W, Johnson M, Panettieri RA, Tantisira KG, Weiss ST, Lu Q. RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells. PLoS ONE. 2014;9(6):99625. https://doi.org/10.1371/journal.pone.0099625.
38. Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W, Chen C, Wang F, Guo X, Lu J, Yang J, Wei M, Tian Z, Guan Y, Tang L, Xu C, Wang L, Gao X, Tian W, Wang J, Yang H, Wang J, Sun Y. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res. 2012;22(5):806–21. https://doi.org/10.1038/cr.2012.30.
39. Yadetie F, Zhang X, Hanna EM, Aranguren-Abadía L, Eide M, Blaser N, Brun M, Jonassen I, Goksøyr A, Karlsen OA. Rna-seq analysis of transcriptome responses in atlantic cod (gadus morhua) precision-cut liver slices exposed to benzo [a] pyrene and 17α-ethynylestradiol. Aquat Toxicol. 2018;201:174–86. https://doi.org/10.1016/j.aquatox.2018.06.003.
40. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinformatics. 2017;18(1):. https://doi.org/10.1186/s12859-016-1457-z.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplemental materials

Additional file 1: Figure S1. An overview of output folder of example data.
Additional file 2: Figure S2. **(a)** The mean quality scores of raw reads from human prostate cancer data. **(b)** The mean quality scores of trimmed reads from human prostate cancer data.
Additional file 3: Tables of differentially expressed gene lists of RASflow using both the transcriptome and the genome as mapping reference and using DESeq2 and edgeR as differential expression analysis methods and their performance.



```
▼ test
    ▶ fastqc
    ▶ fastqc_after_trimming
    ▼ genome
        ▶ alignmentQC
        ▶ countFile
        ▼ dea
            ▶ countGroup
            ▶ DEA
            ▶ visualization
        ▶ report_align_count_data
          report_align_count.html
    ▼ trans
        ▼ dea
            ▶ countGroup
            ▼ DEA
                ▶ gene-level
                ▶ transcript-level
            ▶ visualization
        ▶ quant
        ▶ report_quantify_data
          report_quantify.html
        ▶ tpmFile
```

Figure S1. An overview of output folder of example data.

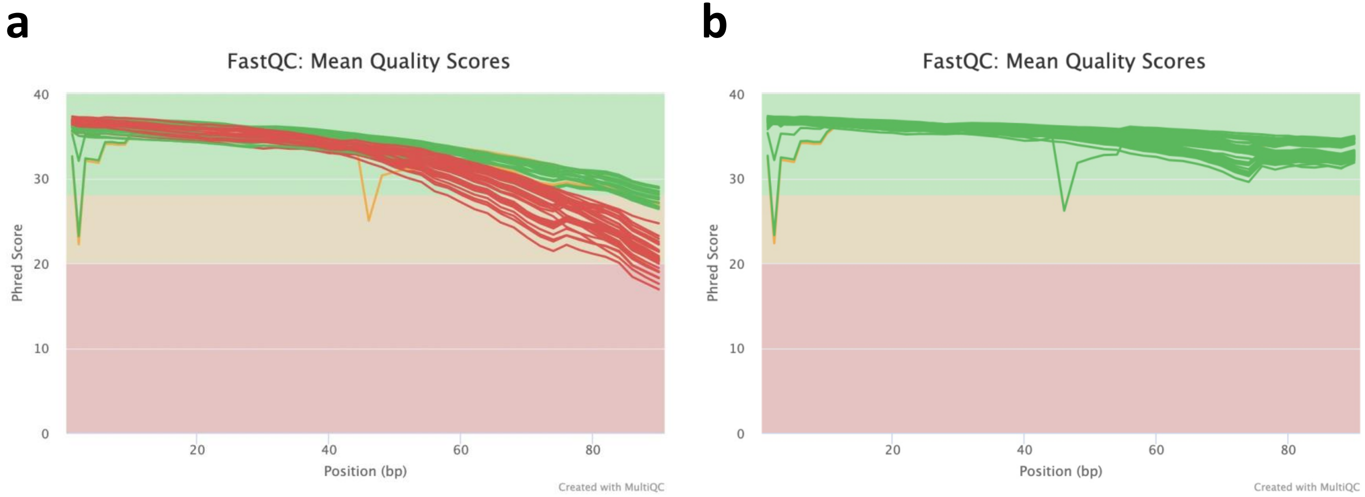


Figure S2. (a) The mean quality scores of raw reads from human prostate cancer data. **(b)** The mean quality scores of trimmed reads from human prostate cancer data.

Table S1.* Differentially expressed gene lists of RASflow using the transcriptome as mapping reference and using DESeq2 as differential expression analysis method.

Table S2.* Differentially expressed gene lists of RASflow using the transcriptome as mapping reference and using edgeR as differential expression analysis method.

Table S3.* Differentially expressed gene lists of RASflow using the genome as mapping reference and using DESeq2 as differential expression analysis method.

Table S4.* Differentially expressed gene lists of RASflow using the genome as mapping reference and using edgeR as differential expression analysis method.

* Full tables can be found at https://doi.org/10.1186/s12859-020-3433-x.

Table S5. Performance of DESeq2 using transcriptome as mapping reference.

		Frankenberger		Haniffa		Ingersoll		Wong		Average			
Workflow	Unit	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform		
RASflow	counts	1103	3910	583	3731	1101	3910	876	3368	915.75	3729.75		
SigGenesRef		2803		1178		2776		1998		2188.75			Rank
Recall		0.393506957		0.49490662		0.39661383		0.43843844		0.43086646			82/239
Precision		0.282097187		0.15625838		0.28158568		0.26009501		0.24500906			188/239

Table S6. Performance of edgeR using transcriptome as mapping reference.

		Frankenberger		Haniffa		Ingersoll		Wong		Average			
Workflow	Unit	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform		
RASflow	counts	957	3164	534	3038	942	3164	771	2749	801	3028.75		
SigGenesRef		2803		1178		2776		1998		2188.75			Rank
Recall		0.34141991		0.4533107		0.33933718		0.38588589		0.37998842			107/239
Precision		0.30246523		0.17577354		0.2977244		0.28046562		0.2641072			127/239

Table S7. Performance of DESeq2 using genome as mapping reference.

		Frankenberger		Haniffa		Ingersoll		Wong		Average			
Workflow	Unit	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform		
RASflow	counts	1068	3744	589	3506	1063	3744	867	3155	896.75	3537.25		
SigGenesRef		2803		1178		2776		1998		2188.75			Rank
Recall		0.38102034		0.5		0.38292507		0.43393393		0.42446984			29/256
Precision		0.28525641		0.16799772		0.28392094		0.2748019		0.25299424			222/256

Table S8. Performance of edgeR using genome as mapping reference.

		Frankenberger		Haniffa		Ingersoll		Wong		Average			
Workflow	Unit	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform	Sig Genes Also Sig By Ref	Sig Genes Present On Ref Platform		
RASflow	counts	705	2168	428	2053	695	2168	583	1860	602.75	2062.25		
SigGenesRef		2803		1178		2776		1998		2188.75			Rank
Recall		0.25151623		0.36332767		0.25036023		0.29179179		0.28924898			198/256
Precision		0.3251845		0.2084754		0.32057196		0.31344086		0.29191818			77/256

Paper III

A Comparative Analysis of Feature Selection Methods for Biomarker Discovery in Study of Toxicant-Treated Atlantic Cod (*Gadus Morhua*) Liver

Xiaokang Zhang, Inge Jonassen

Symposium of the Norwegian AI Society, Communications in Computer and Information Science, **1056/114–123** (2019)

A Comparative Analysis of Feature Selection Methods for Biomarker Discovery in Study of Toxicant-Treated Atlantic Cod (*Gadus Morhua*) Liver

Xiaokang Zhang and Inge Jonassen(✉)

Computational Biology Unit, Department of Informatics,
University of Bergen, Bergen, Norway
{xiaokang.zhang,inge.jonassen}@uib.no
https://www.cbu.uib.no/jonassen/

Abstract. Univariate and multivariate feature selection methods can be used for biomarker discovery in analysis of toxicant exposure. Among the univariate methods, differential expression analysis (DEA) is often applied for its simplicity and interpretability. A characteristic of methods for DEA is that they treat genes individually, disregarding the correlation that exists between them. On the other hand, some multivariate feature selection methods are proposed for biomarker discovery. Provided with various biomarker discovery methods, how to choose the most suitable method for a specific dataset becomes a problem. In this paper, we present a framework for comparison of potential biomarker discovery methods: three methods that stem from different theories are compared by how stable they are and how well they can improve the classification accuracy. The three methods we have considered are: Significance Analysis of Microarrays (SAM) which identifies the differentially expressed genes; minimum Redundancy Maximum Relevance (mRMR) based on information theory; and Characteristic Direction (GeoDE) inspired by a graphical perspective. Tested on the gene expression data from two experiments exposing the cod fish to two different toxicants (MeHg and PCB 153), different methods stand out in different cases, so a decision upon the most suitable method should be made based on the dataset under study and the research interest.

Keywords: Feature selection · Stability · Classification · Biomarker discovery

1 Introduction

Atlantic cod (*Gadus morhua*) is one of the most important commercial fish species in Norway [1], forming the basis for fisheries, trade, and, historically, civilization. Unfortunately, cod is increasingly susceptible to marine pollution from petroleum activities [2,3]. Atlantic cod is commonly used as an indicator species in marine environmental monitoring programs, and a useful model organism to investigate

K. Bach and M. Ruocco (Eds.): NAIS 2019, CCIS 1056, pp. 114–123, 2019.
https://doi.org/10.1007/978-3-030-35664-4_11

the effect of toxicants [4–6]. Finding the best set of biomarkers for Atlantic cod exposed to toxicants is of high research and commercial value. Biomarkers can for example be defined based on the expression level of a set of genes or proteins. Biomarker discovery is an essential part in study of toxicant exposure, and many methods have been proposed to find biomarkers [7]. However, a remaining question is, provided with numbers of biomarker discovery methods, which method is the most suitable one for a particular dataset. This paper provides a framework to compare potential biomarker discovery methods and to give researchers a better basis for choosing which one to use for the task at hand.

In the context of statistics and machine learning, biomarker discovery corresponds to a feature selection problem, where the purpose is to identify the most distinguishing features, for example, distinguishing normal and toxicant-treated cod livers. The task of feature selection is to identify, from a wide range of features, those that are best suited for classification.

The strategies of feature selection methods can be divided into two categories [7]:

1. Classical univariate statistical methods, where the features are considered as independent from each other. Genes that are differentially expressed are regarded as biomarkers.
2. Multivariate methods, which take the interaction between features into consideration when selecting the important features allowing to distinguish samples coming from different groups.

The classical univariate methods try to find the features having significantly different values between the different groups, e.g. control group and treated group. One of the most popular and basic methods is Student's t-test [8]. Some similar research also adopted Analysis of Variance (ANOVA) and Significance Analysis of Microarrays (SAM) to find the differential expressed genes [9–13]. A main drawback of such approaches is that they rest on the assumption that all the genes or proteins are independent from each other, which is clearly not true, since both genes and proteins are part of a biological system where they interact with each other [14,15].

On the other hand, multivariate methods will take the interaction among features into consideration, reflecting that the features are acting in groups. Many feature selection and machine learning methods try to find the features most correlated with the class labels and take the interaction among features into consideration at the same time.

Feature selection methods are often divided into three categories: filter methods which focus on the relation between feature values and class labels; wrapper methods which use an objective function (can be the classification accuracy of the classifier) to evaluate features; and embedded methods where the classifier selects the features automatically [16]. The latter two are both classifier-dependent, and filter methods are more like a one-way decision without feedback from prediction accuracy. In order to find a more general feature selection method, which does not only work well with one specific classifier, we will only focus on the filter methods.

In toxicant exposure study, or more generally, in the context of biology, very often, researchers are faced with the high-dimension-small-sample-size issue, since it is hard and expensive to get a high number of samples (it is often around 10 or even lower), but the number of features (genes or proteins) is usually very high (over one thousand). In such cases, two problems are difficult to avoid: finding a reliable feature subset, as in this case the possibility of chance correlation is quite high; assuring that the selected features are true biomarkers. The true biomarkers should be data-independent, meaning that a small change in the samples should not lead to a large change in the selected features, which requires the feature selection method to be stable. Besides of that, they should also be qualified to be treated as the representatives of the whole feature list and should therefore be able to improve a classifier's prediction accuracy while classifying samples from different biological conditions. Therefore, we will compare the feature selection methods based on two aspects of their performance: stability to find a reliable feature subset and ability to improve a classifier's prediction accuracy.

To make the work reproducible, all the data sets and source codes are publicly available at https://github.com/zhxiaokang/FScompare.

2 Methods

2.1 Data Sets

Two datasets from study of toxicant-treated Atlantic cod liver are used here. One is from the study of the hepatic proteome of MeHg-exposed Atlantic cod, where there are 10 samples in control group, 9 samples in low-dose treated group (0.5 mg/kg Body Weight MeHg), and 9 samples in high-dose treated group (2 mg/kg BW MeHg). The abundances of 1143 proteins were measured after the samples were exposed in vivo to MeHg for two weeks [12]. The other study is from the quantitative proteomics analysis of Atlantic cod livers treated with PCB 153 of various doses of PCB 153 (0, 0.5, 2 and 8 mg/kg BW PCB 153) for two weeks. There are 10 samples in each control group, low-dose treated group, medium-dose treated group, and high-dose treated group. Then 1272 liver proteins are quantified [13].

2.2 Principle of Method and Notations

Consider a set of m samples $\{x_i, y_i\}$ $(i = 1, 2, \ldots m)$. Each sample has n input variables $x_{i,j}$ $(j = 1, 2, \ldots n)$ and one output variable y_i. From the original feature set F, a feature selection method will select a subset S of k variables.

Suppose that there are P feature selection methods to be compared. Using Leave-One-Out Cross-Validation (LOOCV), m feature subsets will be generated for each pre-defined value of k. The stability of each feature selection method $Stab_{p,k}$ $(p = 1, 2, \ldots P)$ can be calculated based on those m subsets.

To test their ability to improve a classifier's prediction accuracy, the generated feature subsets will then be applied to train a classifier and the prediction accuracy of the corresponding classifier will also be measured. Area Under the Curve (AUC) is used to measure the classifier's prediction accuracy [17]. If tested on Q classifiers, the prediction accuracy of each classifier can be calculated $AUC_{p,q,k}$ $(q = 1, 2, ...Q)$. Considering both matrices $Stab$ and AUC, a general evaluation of each feature selection method can finally be achieved so that researchers can choose a proper method for their data.

But the stability does not necessarily agree with the prediction accuracy: the most stable feature selection method may not achieve the highest prediction accuracy. Then the researchers need to balance between these two measures according to their preference and the needs of the project.

2.3 Feature Selection Methods

Some representatives of those two strategies (univariate and multivariate) are compared. For the univariate methods, SAM is applied here, since it was used in the literature from where our data comes. SAM was designed to identify genes with significantly differential expression in microarray experiments. For the multivariate methods, we utilize minimum Redundancy Maximum Relevance (mRMR) [18] and Characteristic Direction from a geometrical aspect (GeoDE) [19]. mRMR is based on information theory. It tries to find out the feature subset in which the redundancy among the features are minimized and the relevance of features and the targeted classes are maximized. GeoDE uses linear discriminant analysis to define a separating hyperplane and the orientation of the hyperplane is used to identify the differentially expressed genes.

Those methods are selected for our comparison because they are based on different theories so that our results are more likely to be valid in general, and they are all widely used biomarker discovery methods. So P equals 3 in this case, but researchers can always compare as many feature selection methods as they want.

2.4 Performance Measurement

Performance of feature selection methods is measured by two factors: stability and accuracy.

Many measures of stability have been proposed. Nogueira et al. studied 15 different measures proposed between 2002 and 2018 and also proposed their novel measure [20]. In our case where the purpose is to compare the stability of different feature selection methods, the absolute values of stability are not that important as long as they are comparable for different methods under the same settings. In each round of comparison, the number of selected features k is a constant, so the stability measure does not need to be able to cope with various numbers of features. LOOCV will generate more than two feature sets based on which the stability is calculated, so the measures which are defined for a pair of feature sets are not proper choices. Considering the measures that satisfy all the requirements, we chose StabPerf [21] for its simplicity and interpretability. The stability is defined as:

$$Stab_{p,k} = \frac{\sum_{f \in F}(freq(f)/m)}{|F|} \tag{1}$$

Where $Stab_{p,k}$ is the stability of a given feature selection method p with a predefined k; m is the number of feature subsets analyzed; F is the set of features that appear in at least one of the m subsets and $|F|$ indicates the cardinality of F; $freq(f)$ is the frequency of feature $f \in F$ that appears in those m subsets.

To test the ability to improve a classifier's prediction accuracy, four popular classification methods are utilized here: Random Forest (RF) [22], Support Vector Machine (SVM) [23], and extended two-class logistic regression (RIDGE and LASSO are applied) [24].

2.5 Cross-Validation Approach

We characterize our problem as a two-class classification problem: the control group versus the treated group. In the process of classification, we need to divide the samples into training set and testing set. But since the number of samples is quite limited, we apply the strategy of LOOCV, which means that in every training-prediction process, we leave one sample out as testing set, and use the other samples as training set to search for the most important features and to train a classifier. With m samples, we will use the i^{th} sample to test the prediction accuracy of the classifier trained from the other $m - 1$ samples. The average of performance observed over all m predictions will be regarded as the estimate of the performance of the model trained over the whole sample set. To avoid overfitting or an overly optimistic estimate, it should be noted that the feature selection and training of classifiers are only limited to the training set, to avoid the information from the testing set leaking into the model training procedure [25]. That makes the size of testing set decided by the number of samples in one classification problem, e.g. 19 in MeHg's high-dose case. Moreover, 19 samples indicate 19 rounds of feature selection and prediction, resulting in 19 selected feature subsets and $19 * 4$ classifiers. Therefore, if a feature selection method is stable enough, there should be a big overlap among these 19 selected feature subsets; at best the feature subsets would be identical. And if the selected features are true biomarkers, the resulting 76 classifiers should yield high prediction accuracies.

To make our comparison more stable, avoiding the accidental findings, and to analyze the characteristic of the feature selection methods, we repeat the above process with different numbers of selected features (ranging from 40 to 400 with a step of 40, but also including 12 and 24 to look into more details with small numbers of selected features where the output varies a lot).

Tukey's Honestly Significant Difference Test (Tukey HSD Test) [26] is also applied to test the significance of the differences between different methods' performance on stability and prediction accuracy.

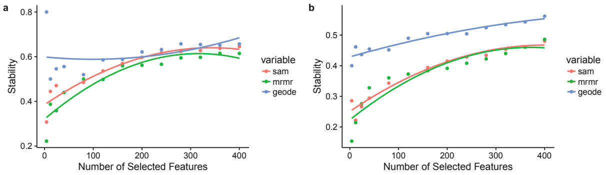


Fig. 1. Stability of feature selection methods on MeHg data. (a) Experiment on high-dose group versus control group. (b) Experiment on low-dose group versus control group.

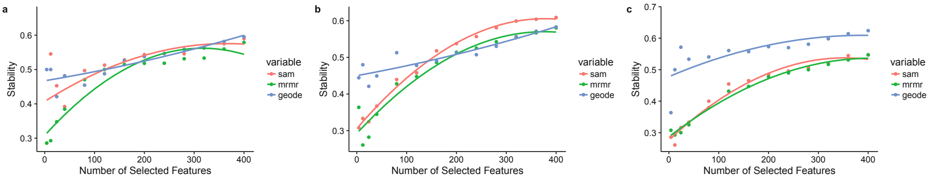


Fig. 2. Stability of feature selection methods on PCB 153 data. (a) Experiment on high-dose group versus control group. (b) Experiment on medium-dose group versus control group. (c) Experiment on low-dose group versus control group.

3 Results

3.1 Stability

We can see from Figs. 1 and 2 that the performance of GeoDE is more stable than SAM and mRMR across different numbers of selected features (with the smallest variance). Another big difference between GeoDE and the other two methods can be seen in low-dose condition of both MeHg and PCB 153: with all numbers of selected features, GeoDE consistently outperforms SAM and mRMR (Figs. 1b and 2c).

The results from Tukey HSD Test on stability are shown in Table 1. We limit the family error rate to 0.05, so the cases with an adjusted p-value (p-adj) smaller than 0.05 are regarded as significantly different. In accordance with the previous analysis, in low-dose condition both for MeHg and PCB 153, GeoDE is much more stable than the other two feature selection methods.

Table 1. Tukey HSD test on stability

Toxicant	Dose condition	Comparison	p-adj
MeHg	low	GeoDE is better than SAM	0.0006
MeHg	low	GeoDE is better than mRMR	0.0005
PCB 153	low	GeoDE is better than SAM	0.0014
PCB 153	low	GeoDE is better than mRMR	0.0007

Table 2. Tukey HSD test on prediction accuracy

Toxicant	Dose condition	Classifier	Comparison	p-adj
MeHg	high	RIDGE	mRMR is better than GeoDE	0.0107
MeHg	high	RIDGE	mRMR is better than SAM	0.0344
MeHg	high	LASSO	mRMR is better than GeoDE	0.0002
MeHg	high	RIDGE	SAM is better than GeoDE	0.0003
MeHg	low	LASSO	GeoDE is better than SAM	0.0004
PCB 153	high	LASSO	mRMR is better than GeoDE	0.0003
PCB 153	high	LASSO	SAM is better than GeoDE	0.0006
PCB 153	medium	SVM	mRMR is better than GeoDE	0.0077
PCB 153	medium	LASSO	SAM is better than GeoDE	0.0009
PCB 153	medium	LASSO	mRMR is better than GeoDE	0.0009
PCB 153	low	RF	GeoDE is better than mRMR	0.0002
PCB 153	low	RF	GeoDE is better than SAM	0.0082
PCB 153	low	SVM	GeoDE is better than SAM	0.0183

3.2 Accuracy

We find that the results of accuracy are not straightforward, since we will get different answers when asking which feature selection method performs the best. In each dose condition, all four classification methods are applied to assess the feature selection methods' ability to improve the prediction accuracy. Across different numbers of selected features, the AUCs of prediction are calculated. Figure 3 is an example in the condition of low-dose MeHg. It shows that SAM performs the best when the classifier is SVM, but GeoDE turns out to be the best with the other three classifiers. To make it simple, for every experiment (each dose of each toxicant), we select the best classification method for it: a classifier that can give a high prediction accuracy for all three feature selection methods. For example, in low-dose condition of MeHg (Fig. 3), RIDGE gets the highest prediction accuracy compared with the other three classifiers regardless of the used feature selection method. Then Fig. 4 gives us all results for all conditions. As we can see, different feature selection methods stand out as the best. In low-dose condition of MeHg and PCB 153 (Figs. 4b and e), GeoDE performs the best, because it has a higher AUC than the other two in most cases of different numbers of selected features. For the other conditions, in high-dose condition of both MeHg and PCB 153 (Figs. 4a and c), and medium-dose condition of PCB 153 (Fig. 4d), mRMR stands out, especially with a low number of selected features.

Another phenomenon we can see from Fig. 4 is that based on gene expression data and our analysis, MeHg appears to influence cods more than PCB 153 does, since it is easier for classifiers to distinguish between control group and treated group with a small number of features (higher prediction accuracy), and the performance is also more stable.

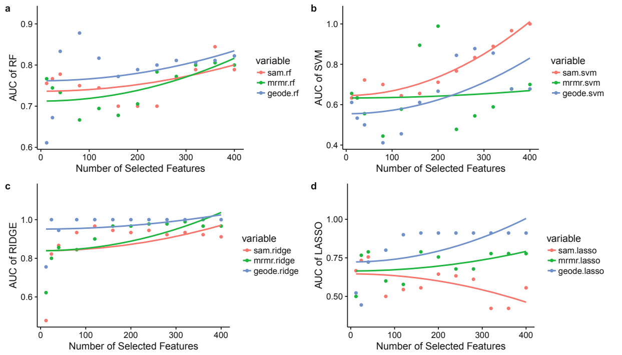


Fig. 3. Prediction accuracy on MeHg low dose data. (a) using RF (b) using SVM (c) using RIDGE (d) using LASSO.

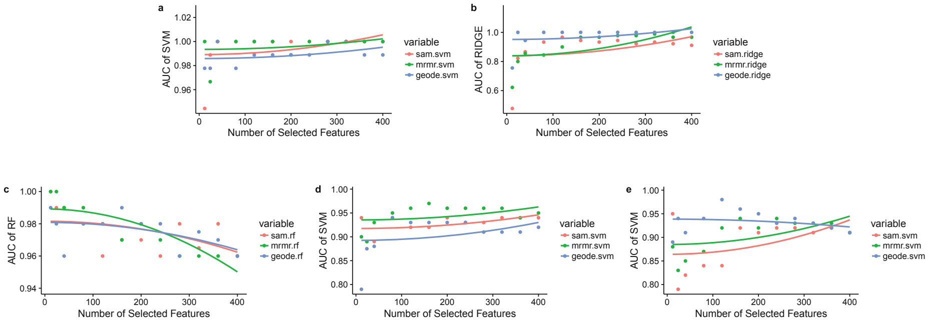


Fig. 4. Prediction accuracy. (a) in high-dose condition of MeHg (b) in low-dose condition of MeHg (c) in high-dose condition of PCB 153 (d) in medium-dose condition of PCB 153 (e) in low-dose condition of PCB 153.

According to the result of Tukey HSD Test on prediction accuracy (Table 2), in different dose conditions and with different classifiers, different feature selection methods will stand out. However, generally speaking, in high-dose condition, mRMR seems to outperform the other two feature selection methods, and in low-dose condition, GeoDE outperforms the other two.

4 Discussion and Conclusion

In this article, we have presented a framework to choose the most suitable biomarker discovery method for a specific dataset by comparing the potential candidates from two aspects: stability, reflecting whether the selected feature subset is robust to changes in the training data, and resulting prediction accuracy.

On the aspect of stability to find a reliable feature subset, our results show that GeoDE is more stable than SAM and mRMR in two ways: its stability varies little across different numbers of selected features for all conditions, and the absolute values of stability are always the highest for all numbers of selected features in low-dose condition.

On the aspect of feature selection methods' ability to improve a classifier's prediction accuracy, in different dose conditions, different feature selection methods show up as the best. mRMR performs well in high-dose condition, but in low-dose condition, GeoDE outperforms the other two.

To conclude this case study, the choice of the most suitable biomarker discovery method quite depends on the dataset under study. If the experiments are conducted in high dose, then mRMR is the best choice, since it gives the highest prediction accuracy and its stability is comparable with the other two. If it's in low dose, then GeoDE is definitely the best choice, considering its excellent performance both in stability and prediction accuracy.

The framework of the comparative analysis is not limited to only this case study, but can be applied to any other similar study.

Acknowledgements. We would like to thank the colleagues in Jonassen Group for helpful discussions and Computational Biology Unit at University of Bergen, where the work was carried out. We also would like to thank the Centre for Digital Life Norway (DLN) and the dCod 1.0 project to which the work is related.

Funding. The dCod 1.0 project is funded under the Digital Life Norway initiative of the BIOTEK 2021 program of the Research Council of Norway (project no. 248840).

References

1. Ageeva, T.N., et al.: Gender-specific responses of mature Atlantic cod (Gadus morhua L.) to feed deprivation. Fish. Res. **188**, 95–99 (2017)
2. Goksøyr, A., Solberg, T.S., Serigstad, B.: Immunochemical detection of cytochrome P450IA1 induction in cod larvae and juveniles exposed to a water soluble fraction of North Sea crude oil. Mar. Pollut. Bull. **22**(3), 122–127 (1991)
3. Balk, L., et al.: Biomarkers in natural fish populations indicate adverse biological effects of offshore oil production. PLoS ONE **6**(5), e19735 (2011)
4. Sundt, et al.: WCM 2010, 2012. NIVA, IMR, IRIS report (2012)
5. Chesman, B.S., et al.: Hepatic metallothionein and total oxyradical scavenging capacity in Atlantic cod Gadus morhua caged in open sea contamination gradients. Aquat. Toxicol. **84**(3), 310–20 (2007)
6. Olsvik, P.A., et al.: Are Atlantic cod in store Lungegrdsvann, a seawater recipient in Bergen, affected by environmental contaminants? A qRT-PCR survey. J. Toxicol. Environ. Health Part A Curr. Issues **72**(3–4), 140–154 (2009)
7. Robotti, E., Manfredi, M., Marengo, E.: Biomarkers discovery through multivariate statistical methods: a review of recently developed methods and applications in proteomics. J. Proteomics Bioinform. **3**, 20 (2014)
8. De Winter, J.C.: Using the student's t-test with extremely small sample sizes. Pract. Assess. Res. Eval. **18**(10), 1–12 (2013)

9. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Nat. Acad. Sci. **98**(9), 5116–5121 (2001)
10. Yadetie, F., et al.: Global transcriptome analysis of Atlantic cod (Gadus morhua) liver after in vivo methylmercury exposure suggests effects on energy metabolism pathways. Aquat. Toxicol. **126**, 314–325 (2013)
11. Yadetie, F., et al.: Liver transcriptome analysis of Atlantic cod (Gadus morhua) exposed to PCB 153 indicates effects on cell cycle regulation and lipid metabolism. BMC Genom. **15**(1), 481 (2014)
12. Yadetie, F., et al.: Quantitative analyses of the hepatic proteome of methylmercury-exposed Atlantic cod (Gadus morhua) suggest oxidative stress-mediated effects on cellular energy metabolism. BMC Genom. **17**(1), 554 (2016)
13. Yadetie, F., et al.: Quantitative proteomics analysis reveals perturbation of lipid metabolic pathways in the liver of Atlantic cod (Gadus morhua) treated with PCB 153. Aquat. Toxicol. **185**, 19–28 (2017)
14. Shannon, P., et al.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. **13**(11), 2498–2504 (2003)
15. Tong, A.H.Y., et al.: Global mapping of the yeast genetic interaction network. Science **303**(5659), 808–813 (2004)
16. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Comput. Biol. Chem. **34**(4), 215–225 (2010)
17. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
18. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
19. Clark, N.R., et al.: The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinform. **15**(1), 79 (2014)
20. Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. J. Mach. Learn. Res. **18**, 1–54 (2018)
21. Davis, C.A., et al.: Reliable gene signatures for microarray classification: assessment of stability and performance. Bioinformatics **22**(19), 2356–2363 (2006)
22. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
23. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
24. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1 (2010)
25. Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. **11**, 2079–2107 (2010)
26. Yandell, B.: Practical Data Analysis for Designed Experiments. Routledge, Abingdon (2017)

Paper IV

An Ensemble Feature Selection Framework Integrating Stability

Xiaokang Zhang, Inge Jonassen
IEEE International Conference on Bioinformatics and Biomedicine, **2792–2798** (2019)

An Ensemble Feature Selection Framework Integrating Stability

1st Xiaokang Zhang
Computational Biology Unit
Department of Informatics
University of Bergen
Bergen, Norway
xiaokang.zhang@uib.no

2nd Inge Jonassen
Computational Biology Unit
Department of Informatics
University of Bergen
Bergen, Norway
inge.jonassen@uib.no

***Abstract*—Ensemble feature selection has drawn more and more attention in recent years. There are mainly two strategies for ensemble feature selection, namely data perturbation and function perturbation. Data perturbation performs feature selection on data subsets sampled from the original dataset and then selects the features consistently ranked highly across those data subsets. Function perturbation frees the user from having to decide on the most appropriate selector for any given situation and works by aggregating multiple selectors. Our study showed that function perturbation resulted in a low stability. We therefore propose a framework, Ensemble Feature Selection Integrating Stability (EFSIS), combining these two strategies and integrating stability during the aggregation of selectors. Empirical results indicate that EFSIS highly improves stability and meanwhile, maintains the prediction accuracy.**

***Index Terms*—feature selection, ensemble learning, stability**

I. Introduction

Feature selection is a crucial technique in machine learning especially for high-dimensional data [1]. It is widely used in many fields to help to find the most important features. In classification tasks, feature selection can help to improve the prediction accuracy by removing the noisy features and avoiding overfitting [2], [3]. But feature selection can also be very challenging, especially when there are a large number of features (high-dimension) and very few training samples (small-size), which is quite often the case in biomedicine and genomics [3]. In such cases a small change in the samples used as training set, can sometimes lead to a large change in the set of selected features. The ability of a feature selection method to give a consistent set of features when the training data changes, is called stability [4]. So a good feature selection method should enable the chosen classifier to obtain high prediction accuracy and also be stable to provide similar selected feature subsets.

In the field of prediction, ensemble learning has been shown to improve the stability and prediction accuracy of the individual learners [5]. The ensemble logic has been more and more applied to feature selection problem in recent years.

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

Ensemble feature selection methods can mainly be divided into two categories: data perturbation and function perturbation [6], [7].

In data perturbation (sometimes referred to as the homogeneous ensemble approach), feature selection is performed on several subsets of the samples, each analysis generating potentially different feature subsets. In this case the same feature selection method is used to analyze all subsets. The resulting feature subsets are then aggregated into one final feature subset [8]–[12]. Pes et al. showed that data perturbation can improve the stability of the original feature selection method [12].

Function perturbation (also referred to as the heterogeneous ensemble approach) combines the outputs from several feature selection methods - to free the user from having to choose one selection method and to benefit from the strengths of a set of methods [11], [13]–[15]. In this approach, a set of selected feature selection methods are all applied on the same training set. According to the literature, function perturbation can maintain or improve classification performance.

However, we have not been able to find in the literature any study of the stability of function-perturbation based methods for feature selection.

The concern is that each feature selection method makes different sets of assumptions and rationale for choosing the important features; combining selected features from across different selectors may give inferior performance including decreased stability. Especially in the field of biomedicine or genomics, where the feature dimension is very high but the sample number is comparably low, such as microarray data, a small change in the dataset may produce large change in the resulting features. Therefore, we find it highly relevant to investigate the issue of stability in ensemble feature selection and especially in context of function perturbation approaches. Through our preliminary experiments, we found that function perturbation could indeed result in low stability. Since data perturbation has been shown to improve stability, we propose a framework to combine these two strategies to solve the stability issue of function perturbation.

The framework includes two phases. In the first phase, data perturbation is applied to generate a number of data subsets and each of these is given to a number of feature selectors (also referred to as rankers since they rank the features). For each ranker, the results across data subsets are aggregated to produce one ranked list of features. In addition, for each ranker, a statistic reflecting its stability is calculated. In the second phase which is function perturbation, the results from each ranker are aggregated - using the estimated stability of each ranker to weigh their votes - to produce a final ranking and a final feature subset. The framework is named Ensemble Feature Selection Integrating Stability (EFSIS) and the source code is available on GitHub (https://github.com/zhxiaokang/EFSIS).

As benchmarks for our experiments, we tested our method on six cancer datasets coming from microarray experiments. To better understand its performance, we compared EFSIS with each of the methods aggregated in EFSIS and also with basic function perturbation. The result showed that the stability was highly improved by using EFSIS. Meanwhile, the prediction accuracy was also maintained well.

The rest of the article is organized as follows: Section II describes the proposed EFSIS framework, along with basic function perturbation, the individual feature selection methods, and the metrics applied to evaluate stability and prediction accuracy; Section III introduces the experimental study, including experimental settings and results; Section IV discusses the experiments and concludes the work.

II. Methods

A. Methodology of EFSIS

Our proposed ensemble feature selection framework includes two phases: data perturbation and function perturbation [6], [7]. The framework is illustrated in Figure 1.

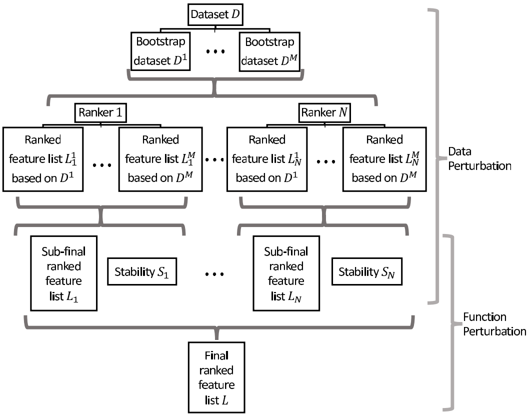

Dataset D
Bootstrap dataset D^1
Bootstrap dataset D^M
Ranker 1
Ranker N
Ranked feature list L_1^1 based on D^1
Ranked feature list L_1^M based on D^M
Ranked feature list L_N^1 based on D^1
Ranked feature list L_N^M based on D^M
Data Perturbation
Sub-final ranked feature list L_1
Stability S_1
Sub-final ranked feature list L_N
Stability S_N
Function Perturbation
Final ranked feature list L

Fig. 1. The framework of EFSIS

Given the original dataset D, we use bootstrapping [16] to get M perturbed variants of D ($\{D^1, ...D^m, ...D^M\}$) for the dataset D with p samples: we randomly draw p samples from D with replacement, allowing some samples to be picked multiple times while some samples may be absent in D^m. Each bootstrap dataset D^m is then passed to each of the included individual feature selection methods, each performing a ranking of all the features based on how well they distinguish samples from different groups. For simplicity, in the following, we call each feature selection method a *ranker*.

In the first phase which is data perturbation, let us take ranker n ($n \in \{1, ...N\}$) as a general representative to explain the idea of data perturbation. Ranker n will rank the features based on the bootstrap datasets. Corresponding to each bootstrap dataset, one ranked list will be generated. Therefore, each ranker will end up with M ranked lists $\{L_n^1, ...L_n^m, ...L_n^M\}$. With an aggregation strategy which we introduce in Subsection II.C, the M lists can then be combined into one list (L_n). In addition to L_n, a side product, the stability of ranker n, that we will denote as S_n, can be calculated using the stability definition described in Subsection II.B: with a pre-defined threshold t, the top t features in L_n^m will be picked to constitute a feature subset, and then the M feature subsets will be used to calculate the stability of ranker n. The data perturbation procedure above will be applied to all N rankers to generate N sub-final ranked feature lists $\{L_1, ...L_N\}$.

In the phase of function perturbation, another aggregation strategy which integrates the stability of the rankers (we introduce in Subsection II.C) combines those N sub-final ranked feature lists into one final list L. The top t features are kept as the selected important features by EFSIS.

B. Stability

A stable feature selection method should give similar feature subsets even given varying samples. We use the similarity between feature subsets derived from different sample sets to measure the stability of the corresponding feature selection method. We used the stability definition proposed by [8]:

$$S_n = \frac{\sum_{f \in F}(\omega(f)/M)}{|F|} \tag{1}$$

Where S_n is the stability of a given feature selection method n; M is the number of feature subsets analyzed; F is the set of features that appear in at least one of the M subsets and $|F|$ indicates the cardinality of F; $\omega(f)$ is the frequency of feature $f \in F$ that appears in those M subsets.

C. Aggregation strategies

There are two aggregations in the EFSIS paradigm shown in Figure 1. A very recent study [17] used intersection and union operations to aggregate the lists. But there is an extreme case where there is no intersection of the sub-lists. So we used another more robust strategy, Rank Products (RP) [18], to score each feature: the product of ranking positions of one feature in different ranked lists is used as its aggregated ranking score. This strategy was applied in data perturbation. The ranking score of a feature f from ranker n can be calculated as follows:

$$R_{f,n} = \prod_{m=1}^{M} R_{f,n}^m \tag{2}$$

where $R^m_{f,n}$ is the rank of feature f from ranker n on bootstrap set m. Based on this score, an aggregated ranked feature list L_n for ranker n can be obtained.

In function perturbation, to address the issue of stability, we extended the RP aggregation strategy so that the rankers with higher stability have greater weights. We refer to this strategy as stability-weighted RP utilizing the following equation for aggregating the ranks:

$$R_f = \prod_{n=1}^{N} (R_{f,n})^{(1-S_n)} \tag{3}$$

where $1 - S_n$ is defined as the weight of ranker n, so that a more stable ranker is assigned a higher weight. Ranking the features based on this score, we get the final ranked list.

In fact, the basic function perturbation is a special case of the second phase in EFSIS: each ranker ranks the features based on the original dataset D, afterwards, it will apply the RP aggregation strategy, aggregating the rankings from different rankers in a similar way as EFSIS does in the second phase, except that there is no weight for each ranker ($S_n = 0$ in Equation (3)).

D. Individual feature selection methods

In general, there are three categories of feature selection methods: filter methods which rank the features only based on their correlation with the targeted classes, wrapper methods which use an objective function (can be the prediction accuracy obtained by the classifier using the selected features) to evaluate features, and embedded methods where the classifier itself performs feature selection.

Since one motivation of the ensemble framework is to make it as general as possible, we want to make it classifier-independent. Therefore, we consider only filter methods in this context.

In our experiment, we used four very diverse feature selection methods which are based on different sets of assumptions, to demonstrate the generality of the proposed framework. In particular, we employed both univariate techniques which treat the features as independent from each other and multivariate techniques which take the interaction between features into consideration.

As representatives of univariate techniques, we used:

- Significance Analysis of Microarrays (SAM) that was originally designed to identify genes with significantly differential expression in microarray experiments [19]. It assigns a score to each gene based on the change in gene expression relative to the standard deviation of repeated experiments.
- Information gain which is one of the most popular univariate methods [20]. It evaluates each feature based on the entropy concept from information theory.

As representatives of multivariate techniques, we applied:

- The Characteristic Direction method (GeoDE) which is a geometrical multivariate approach [21]. It defines a separating hyperplane using linear discriminant analysis to characterize the differential expression of microarray or RNA-Seq data.
- ReliefF [22] is an extension of the original Relief algorithm [23], [24] that evaluates a feature according to how well it can distinguish among instances that are near to each other. Compared to Relief, ReliefF is more robust to noisy and incomplete datasets.

E. Classification algorithm

In evaluating the predictive performance of the selected feature subsets, we applied the classification algorithm Support Vector Machine (SVM) [25] to learn a classifier based on the selected feature subsets. Provided with a training dataset of samples marked with group labels (samples are characterized by the selected features), SVM will learn an optimal hyperplane separating the samples from different groups. And the optimal hyperplane will be used to predict the labels of the samples from test set. A prediction accuracy can be calculated comparing the predicted labels with the true labels. A better feature subset will enable the SVM to achieve a higher prediction accuracy. For simplicity, we chose a linear kernel for SVM and we used Area Under Curve (AUC) [26] to summarize the obtained prediction accuracy.

III. Experimental study

A. Datasets

EFSIS was tested on six gene expression datasets produced using microarrays to study different forms of cancer (datasets were collected by [27]). The main characteristics of the datasets, including numbers of features and samples, are given in Table I. Feature selection can provide valuable information in such applications. The selected features can be regarded as biomarkers and they reflect characteristics of the studied cancer forms and can help to classify the patients. Feature selection can allow the cancer researcher or clinician to focus on a small number of biomarkers instead of thousands of features, which can save lots of money and time for further studies. Biomarkers can also help to improve the understanding of the cancer forms on a molecular level.

TABLE I
Datasets used in the experiments

Name	Features	Samples	Refs
AML	12 625	54	[28]
CNS	7 129	60	[29]
DLBCL	7 129	77	[30]
Prostate	12 600	102	[31]
Leukemia	7 129	72	[32]
ColonBreast	22 283	52	[33]

B. Experimental procedure and settings

To evaluate the performance of EFSIS, it was compared with the aggregated individual rankers and the corresponding basic function perturbation aggregating the same four rankers. The performance was evaluated in two aspects: stability and prediction accuracy. Both stability and prediction accuracy depend on how many features are to be selected and used for classification (denoted t), hence we performed the assessment with a range of values for t.

In order to obtain an unbiased estimation of performance, we performed the experiments using a ten-fold cross-validation scheme [34], [35]. Thus, we obtained 10 selected feature subsets for each pre-defined threshold t, for each dataset and for each ranker. By doing classification analysis with those 10 feature subsets, we obtained 10 prediction accuracy scores. At the same time, by calculating the similarity of those 10 feature subsets using Equation (1), we obtained an estimate of the stability of the corresponding ranker.

Considering the highly variable number of features in each dataset (as shown in Table I), instead of using an absolute number of features t, we used a percentage of the original number of features. We explored a range of values from 0.3% to 5%.

The main parameters for EFSIS are the number of bootstrap datasets M and number of rankers N. M was chosen based on the recommendation in [12] ($M = 50$). In our analysis, $N = 4$, the rankers are described in Subsection II.D. The competitors of EFSIS would therefore be the four individual rankers and the basic function perturbation of the same four rankers.

To speed up the computation, parallel computing can be applied to EFSIS taking advantage of its structure. The parallelization can be done in multiple ways. What we have tried was to split the jobs by bootstrap datasets so that the job corresponding to one dataset was performed by one node.

C. *Experimental results of stability performance*

In this section, we will study the stability of the rankers. The stability was tested on 6 datasets with 9 different percentages of selected features.

Figure 2 shows the performance of the four individual rankers and the two ensemble rankers. Let us firstly look at the individual rankers. GeoDE has the same problem as in the previous section: it achieves a very high stability in the CNS dataset but a very low one in the DLBCL dataset. ReliefF seems to be a very unstable method with the lowest stability score across all datasets, even in the dataset DLBCL where it showed great predictive performance (as mentioned in the previous section).

When we compare basic function perturbation with the four individual rankers across the 6 datasets as shown in Figure 2, we can find that basic function perturbation is either the second or the third worst one. In comparison, the performance of EFSIS is much more satisfactory: it is the second best one in the first 3 datasets (Figure 2 A-C), and it performs consistently better than all the individual rankers in the latter 3 datasets (Figure 2 D-F). If we compare between basic function perturbation and EFSIS, Figure 2 shows clearly that EFSIS performs always better than basic function perturbation. The box plot in Figure 3 shows the comparison between these two ensemble rankers on 6 datasets with the star ($*$) indicating the significance of difference (P-value was calculated using Wilcoxon Signed-Ranks Test [36]). We can see that the stability of EFSIS is significantly higher than basic function perturbation in all 6 datasets.

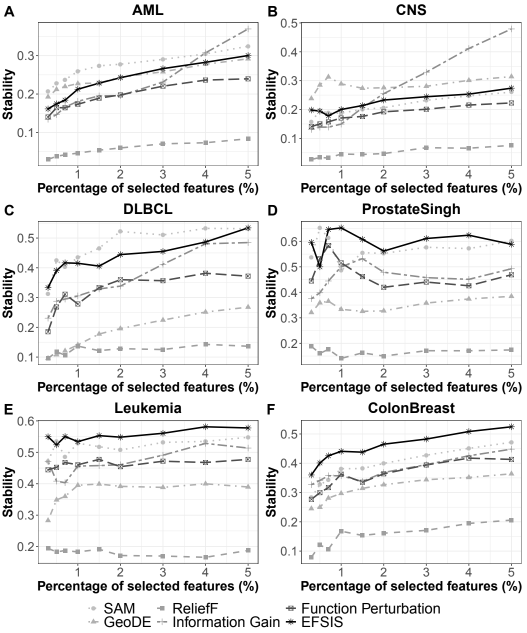


Fig. 2. Stability performance of six rankers on six datasets, tested with different percentages of selected features. For each dataset, four individual rankers (SAM, GeoDE, ReliefF, Information Gain), basic Function Perturbation, and EFSIS are considered.

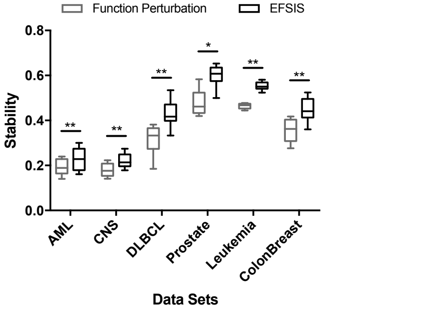


Fig. 3. Comparison of basic Function Perturbation and EFSIS in stability performance on six datasets. ** $=$ P-value < 0.005, * $=$ P-value < 0.01.

D. *Experimental results of predictive performance*

Even though stability is important, prediction accuracy cannot be ignored. The mean AUC (averaging the AUCs from ten-fold cross-validation) and associated standard deviation of four individual rankers and two ensemble ones (basic function perturbation and EFSIS) tested on datasets CNS and

TABLE II

Predictive performance of six rankers on dataset CNS with different percentages of selected features: mean AUC and standard deviation.

Ranker	**Percentage of selected features (%)**								
	0.3	0.5	0.7	1	1.5	2	3	4	5
SAM	0.72 ± 0.22	$\mathbf{0.69 \pm 0.22}^*$	0.71 ± 0.20	0.72 ± 0.17	$\mathbf{0.71 \pm 0.21}^*$	$\mathbf{0.72 \pm 0.17}^*$	$\mathbf{0.73 \pm 0.18}^*$	$\mathbf{0.69 \pm 0.19}^*$	$\mathbf{0.73 \pm 0.18}^*$
GeoDE	0.63 ± 0.16	0.76 ± 0.08	$0.81 \pm 0.16^\dagger$	$0.82 \pm 0.13^\dagger$	$0.82 \pm 0.18^\dagger$	$0.88 \pm 0.17^\dagger$	$0.89 \pm 0.16^\dagger$	$0.88 \pm 0.14^\dagger$	$0.90 \pm 0.14^\dagger$
ReliefF	0.69 ± 0.18	0.72 ± 0.14	0.75 ± 0.16	$\mathbf{0.68 \pm 0.15}^*$	0.74 ± 0.19	$\mathbf{0.70 \pm 0.19}^*$	$\mathbf{0.75 \pm 0.16}^*$	0.79 ± 0.21	$\mathbf{0.73 \pm 0.15}^*$
Info_Gain	0.69 ± 0.17	$0.78 \pm 0.18^\dagger$	0.76 ± 0.19	0.71 ± 0.19	$\mathbf{0.65 \pm 0.17}^*$	$\mathbf{0.70 \pm 0.13}^*$	$\mathbf{0.66 \pm 0.18}^*$	$\mathbf{0.71 \pm 0.16}^*$	$\mathbf{0.78 \pm 0.15}^*$
Func_Pert	0.72 ± 0.12	0.77 ± 0.18	0.68 ± 0.21	$\mathbf{0.68 \pm 0.21}^*$	0.77 ± 0.15	0.80 ± 0.21	$\mathbf{0.80 \pm 0.11}^*$	0.80 ± 0.13	$\mathbf{0.80 \pm 0.16}^*$
EFSIS	$0.74 \pm 0.22^\dagger$	0.69 ± 0.16	0.68 ± 0.19	0.75 ± 0.15	0.79 ± 0.16	0.83 ± 0.14	0.82 ± 0.14	$\mathbf{0.78 \pm 0.11}^*$	$\mathbf{0.79 \pm 0.15}^*$

†The best ranker in one experiment (one specific percentage of selected features).

*The rankers that are significantly worse than the best one.

TABLE III

Predictive performance of six rankers on dataset DLBCL with different percentages of selected features: mean AUC and standard deviation.

Ranker	**Percentage of selected features (%)**								
	0.3	0.5	0.7	1	1.5	2	3	4	5
SAM	0.91 ± 0.13	0.90 ± 0.12	0.96 ± 0.08	0.96 ± 0.06	0.97 ± 0.07	0.97 ± 0.07	0.95 ± 0.11	0.94 ± 0.11	$0.97 \pm 0.07^\dagger$
GeoDE	$\mathbf{0.86 \pm 0.10}^*$	$\mathbf{0.87 \pm 0.10}^*$	$\mathbf{0.86 \pm 0.12}^*$	$\mathbf{0.89 \pm 0.10}^*$	$\mathbf{0.88 \pm 0.16}^*$	$\mathbf{0.86 \pm 0.22}^*$	$\mathbf{0.85 \pm 0.22}^*$	$\mathbf{0.89 \pm 0.11}^*$	0.92 ± 0.10
ReliefF	$0.96 \pm 0.08^\dagger$	0.94 ± 0.11	$0.99 \pm 0.03^\dagger$	0.96 ± 0.09	0.98 ± 0.06	0.94 ± 0.10	$0.99 \pm 0.03^\dagger$	$0.99 \pm 0.03^\dagger$	$0.97 \pm 0.07^\dagger$
Info_Gain	0.95 ± 0.11	$0.95 \pm 0.09^\dagger$	0.95 ± 0.11	0.96 ± 0.08	0.96 ± 0.08	0.96 ± 0.08	0.96 ± 0.08	0.97 ± 0.08	0.96 ± 0.06
Func_Pert	0.91 ± 0.12	0.92 ± 0.09	0.96 ± 0.08	0.96 ± 0.06	$0.98 \pm 0.05^\dagger$	$0.98 \pm 0.05^\dagger$	0.96 ± 0.07	0.94 ± 0.10	0.93 ± 0.11
EFSIS	0.92 ± 0.10	0.94 ± 0.08	$\mathbf{0.93 \pm 0.10}^*$	$0.97 \pm 0.06^\dagger$	0.97 ± 0.06	0.97 ± 0.06	0.96 ± 0.07	0.97 ± 0.07	$0.97 \pm 0.07^\dagger$

†The best ranker in one experiment (one specific percentage of selected features).

*The rankers that are significantly worse than the best one.

DLBCL with 9 different percentages of selected features are shown in Table II and Table III. For each experiment with a specific percentage of selected features, the best ranker (the one with the highest mean AUC and lowest standard deviation) is marked with dagger, and the ones that are significantly worse than the best one are marked with star and are in bold font (P-value < 0.05, Wilcoxon Signed-Ranks Test [36]). It shows a problem of the individual rankers: some individual rankers perform quite well in some datasets but poorly in some others. For example, GeoDE performs quite well in dataset CNS (it achieves the highest prediction accuracy among all rankers 7 times out of 9), but performs unsatisfactorily in dataset DLBCL (it achieves a significantly lower prediction accuracy than the best one 8 times out of 9, which makes it the worst for this dataset). But ReliefF performs contrarily to GeoDE in these two datasets. Since the performance of feature selection methods varies from dataset to dataset, it is difficult for researchers to choose an adequate one for their dataset. That problem is actually a big motivation for function perturbation since it can free researchers from that difficult decision. Ensemble rankers (function perturbation and EFSIS) combine the results from all candidate rankers.

The results of the other four datasets are given in the Appendix Table IV. Table II-IV show that the predictive performance of ensemble rankers is more stable across the different datasets analyzed. Function perturbation and EFSIS are slightly better than the individual rankers: they are significantly worse than the best ranker in 4 out of the 54 experiments (6 datasets × 9 percentages of selected features), while four individual rankers are worse in 8, 10, 6, 7 experiments, respectively.

IV. Conclusion

We have described a new framework for ensemble feature selection, which combines data perturbation and function perturbation and utilizes the stability of the individual methods as weights. The new framework utilizes data perturbation's ability to improve stability to solve the low-stability issue of function perturbation. It possesses the advantages of both function perturbation and data perturbation: it combines the results from different individual feature selection methods and shows robust predictive performance, and it also provides more stable selected feature subsets. Therefore, it frees the researchers from choosing the most suitable feature selection method for their datasets. Also, compared to basic function perturbation, it provides higher stability. To be noted, EFSIS is a framework, meaning that researchers can put whatever they like in the framework. For example, they can replace the individual rankers with some specific ones that are commonly used in their research field or add new ones as more and more feature selection methods are being proposed.

In the EFSIS framework, we have chosen to perform data perturbation in the first phase so that each ranker (feature selection method) is performed on all bootstrap datasets to produce one ranking that is next combined with rankings from the other rankers. In this way we can obtain the stability of each individual ranker based on the same subsets of samples, enabling us to use the stability estimates when combining results across the rankers. However, it would be interesting to explore an alternative approach where function perturbation is applied to each bootstrap dataset, which will produce M ranked lists. In the next step, these M lists will be combined (using for example Rank Products) to obtain the final ranked list. The idea behind this strategy is to make use of data perturbation's ability to improve the stability of function perturbation. Future studies will include this and other directions.

Appendix

TABLE IV
Predictive performance of six rankers on four datasets with different percentages of selected features: mean AUC and standard deviation.

Dataset	Ranker	Percentage of selected features (%)								
		0.3	0.5	0.7	1	1.5	2	3	4	5
AML	SAM	**0.69 ± 0.17***	0.73 ± 0.16	0.73 ± 0.20	0.76 ± 0.14	0.78 ± 0.17	0.75 ± 0.18	**0.74 ± 0.20***	0.76 ± 0.17	0.77 ± 0.16
	GeoDE	0.74 ± 0.16	0.69 ± 0.25	0.76 ± 0.20†	**0.76 ± 0.16***	0.80 ± 0.16†	0.80 ± 0.18†	0.84 ± 0.15†	0.79 ± 0.18	0.79 ± 0.22
	ReliefF	0.76 ± 0.21	0.73 ± 0.15	0.68 ± 0.21	**0.69 ± 0.14***	0.75 ± 0.15	0.76 ± 0.18	**0.72 ± 0.14***	0.74 ± 0.18	0.76 ± 0.15
	Info_Gain	0.81 ± 0.16†	0.75 ± 0.18†	0.74 ± 0.16	**0.73 ± 0.17***	0.79 ± 0.14	0.76 ± 0.17	0.77 ± 0.17	0.79 ± 0.16	0.80 ± 0.17
	Func_Pert	0.75 ± 0.16	0.73 ± 0.23	0.74 ± 0.15	0.81 ± 0.14†	0.79 ± 0.13	0.79 ± 0.17	**0.75 ± 0.21***	0.80 ± 0.18†	0.81 ± 0.14†
	EFSIS	0.73 ± 0.20	0.74 ± 0.17	0.72 ± 0.22	0.75 ± 0.17	0.72 ± 0.18	0.73 ± 0.19	**0.77 ± 0.16***	0.75 ± 0.19	0.76 ± 0.17
Prostate	SAM	0.95 ± 0.08†	0.95 ± 0.08	0.95 ± 0.08	0.96 ± 0.06†	0.96 ± 0.07	0.95 ± 0.07	0.96 ± 0.07	0.96 ± 0.07	0.96 ± 0.07†
	GeoDE	0.90 ± 0.15	0.93 ± 0.09	0.94 ± 0.09	0.95 ± 0.08	0.95 ± 0.09	**0.94 ± 0.08***	0.95 ± 0.06	0.95 ± 0.06	0.96 ± 0.06
	ReliefF	0.94 ± 0.08	0.96 ± 0.08†	0.96 ± 0.06†	0.94 ± 0.10	0.97 ± 0.04	0.96 ± 0.06	0.94 ± 0.08	0.96 ± 0.07	0.94 ± 0.09
	Info_Gain	0.94 ± 0.11	0.94 ± 0.10	0.94 ± 0.10	0.95 ± 0.09	**0.95 ± 0.09***	0.96 ± 0.07	0.97 ± 0.06†	0.97 ± 0.06†	0.96 ± 0.08
	Func_Pert	0.95 ± 0.09	0.94 ± 0.10	0.95 ± 0.10	0.95 ± 0.09	0.96 ± 0.06	0.96 ± 0.07	0.96 ± 0.06	0.95 ± 0.08	0.95 ± 0.09
	EFSIS	0.95 ± 0.09	0.94 ± 0.10	0.95 ± 0.09	0.95 ± 0.08	0.97 ± 0.07†	0.97 ± 0.07†	0.95 ± 0.08	0.94 ± 0.09	0.94 ± 0.09
Leukemia	SAM	0.99 ± 0.04	0.98 ± 0.05	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†
	GeoDE	0.98 ± 0.05	0.99 ± 0.02†	0.99 ± 0.04	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†
	ReliefF	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.04	0.99 ± 0.04	0.99 ± 0.02†	0.99 ± 0.04	0.98 ± 0.04	0.98 ± 0.05	0.98 ± 0.04
	Info_Gain	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†
	Func_Pert	0.97 ± 0.08	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†
	EFSIS	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†	0.99 ± 0.02†
ColonBreast	SAM	0.98 ± 0.08	0.98 ± 0.08	0.97 ± 0.06	0.98 ± 0.05†	0.99 ± 0.03†	0.97 ± 0.07	0.97 ± 0.06	0.97 ± 0.06	0.97 ± 0.06
	GeoDE	0.99 ± 0.04†	0.99 ± 0.04†	0.99 ± 0.04†	0.98 ± 0.05	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.98 ± 0.05
	ReliefF	0.95 ± 0.08	0.95 ± 0.12	0.95 ± 0.11	0.94 ± 0.12	0.98 ± 0.05	1.00 ± 0.00†	0.97 ± 0.08	0.96 ± 0.08	0.97 ± 0.05
	Info_Gain	0.98 ± 0.05	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.98 ± 0.08	0.98 ± 0.08	0.99 ± 0.04	0.98 ± 0.08	0.95 ± 0.12
	Func_Pert	0.98 ± 0.08	0.99 ± 0.04†	0.98 ± 0.05	0.98 ± 0.05	0.97 ± 0.06	0.99 ± 0.03	0.99 ± 0.03†	0.98 ± 0.05	0.98 ± 0.05
	EFSIS	0.98 ± 0.08	0.98 ± 0.08	0.99 ± 0.04†	0.98 ± 0.05	0.96 ± 0.07	0.98 ± 0.05	0.98 ± 0.05	0.99 ± 0.04†	0.99 ± 0.04†

†The best ranker in one experiment (one specific percentage of selected features).
*The rankers that are significantly worse than the best one.

ACKNOWLEDGMENT

We would like to thank Computational Biology Unit at University of Bergen, where the work was carried out. We would like to thank the colleagues in Jonassen Group for helpful discussions, especially Dr. F. Z. Ghavidel. The research has been done within the project dCod 1.0, funded by the Digital Life Norway initiative of the BIOTEK 2021 program of the Research Council of Norway (project no. 248840).

REFERENCES

[1] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” Neurocomputing, vol. 300, pp. 70-79, Jul. 2018.

[2] A. Jovi, K. Brki, and N. Bogunovi, “A review of feature selection methods with applications,” in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings, 2015, pp. 1200-1205.

[3] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” Comput. Biol. Med., vol. 112, p. 103375, Sep. 2019.

[4] A. Kalousis, J. Prados, and M. Hilario, “Stability of feature selection algorithms: A study on high-dimensional spaces,” Knowl. Inf. Syst., vol. 12, no. 1, pp. 95-116, May 2007.

[5] O. Sagi and L. Rokach, “Ensemble learning: A survey,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, p. e1249, Jul. 2018.

[6] Z. He and W. Yu, “Stable feature selection for biomarker discovery,” Comput. Biol. Chem., vol. 34, no. 4, pp. 215-225, Aug. 2010.

[7] V. Boln-Canedo and A. Alonso-Betanzos, “Ensembles for feature selection: A review and future trends,” Inf. Fusion, vol. 52, no. May 2018, pp. 1-12, 2019.

[8] C. A. Davis et al., “Reliable gene signatures for microarray classification: assessment of stability and performance,” Bioinformatics, vol. 22, no. 19, pp. 2356-2363, Oct. 2006.

[9] F. R. Bach and F. R., “Bolasso: model consistent Lasso estimation through the bootstrap,” in Proceedings of the 25th international conference on Machine learning - ICML 08, 2008, pp. 33-40.

[10] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, “Robust biomarker identification for cancer diagnosis with ensemble feature selection methods,” Bioinformatics, vol. 26, no. 3, pp. 392-398, Feb. 2010.

[11] B. Seijo-Pardo, I. Porto-Daz, V. Boln-Canedo, and A. Alonso-Betanzos, “Ensemble feature selection: Homogeneous and heterogeneous approaches,” Knowledge-Based Syst., vol. 118, pp. 124-139, Feb. 2017.

[12] B. Pes, N. Dess, and M. Angioni, “Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data,” Inf. Fusion, vol. 35, pp. 132-147, 2017.

[13] N. C. Tan, W. G. Fisher, K. P. Rosenblatt, and H. R. Garner, “Application of multiple statistical tests to enhance mass spectrometry-based biomarker discovery,” BMC Bioinformatics, vol. 10, no. 1, p. 144, May 2009.

[14] A. Ben Brahim and M. Limam, “Robust ensemble feature selection for high dimensional data sets,” 2013 Int. Conf. High Perform. Comput. Simul., pp. 151-157, 2013.

[15] S. Ahmed, M. Zhang, and L. Peng, “Improving feature ranking for biomarker discovery in proteomics mass spectrometry data using genetic programming Improving feature ranking for biomarker discovery in proteomics mass spectrometry data using genetic programming,” Conn. Sci., vol. 26, no. 3, pp. 215-243, 2014.

[16] B. Efron, R. J. Tibshirani, and R. J. Tibshirani, An Introduction to the Bootstrap. CRC press, 1994.

[17] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, “A new hybrid ensemble feature selection framework for machine learning-based phishing detection system,” Inf. Sci. (Ny)., vol. 484, pp. 153-166, May 2019.

[18] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, “Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments,” FEBS Lett., vol. 573, no. 1-3, pp. 83-92, Aug. 2004.

[19] V. Goss Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” Proc. Natl. Acad. Sci., vol. 98, no. 9, pp. 5116-5121, 2001.

[20] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[21] N. R. Clark et al., “The characteristic direction: a geometrical approach to identify differentially expressed genes,” BMC Bioinformatics, vol. 15, no. 1, p. 79, Mar. 2014.

[22] I. Kononenko, “Estimating attributes: Analysis and extensions of RELIEF,” in European conference on machine learning, Springer, Berlin, Heidelberg, 1994, pp. 171-182.

[23] K. Kira and L. A. Rendell, “The feature selection problem: traditional methods and a new algorithm,” in Proceedings of the tenth national conference on Artificial intelligence, 1992, pp. 129-134.

[24] K. Kira and L. A. Rendell, “A Practical Approach to Feature Selection,” Mach. Learn. Proc. 1992, pp. 249-256, Jan. 1992.

[25] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273-297, Sep. 1995.

[26] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit. Lett., vol. 27, no. 8, pp. 861-874, Jun. 2006.

[27] N. Lazzarini and J. Bacardit, “RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers,” Lazzarini Bacardit BMC Bioinforma., vol. 18, 2017.

[28] T. Yagi et al., “Identification of a gene expression signature associated with pediatric AML prognosis,” Blood, vol. 102, pp. 1849-1856, 2003.

[29] S. L. Pomeroy et al., “Prediction of central nervous system embryonal tumour outcome based on gene expression,” Nature, vol. 415, no. 6870, pp. 436-442, Jan. 2002.

[30] M. A. Shipp et al., “Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning,” Nat. Med., vol. 8, no. 1, pp. 68-74, Jan. 2002.

[31] D. Singh et al., “Gene expression correlates of clinical prostate cancer behavior,” Cancer Cell, vol. 1, no. 2, pp. 203-209, Mar. 2002.

[32] T. R. Golub et al., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” Science, vol. 286, no. 5439, pp. 531-537, Oct. 1999.

[33] D. Chowdary et al., “Prognostic Gene Expression Signatures Can Be Measured in Tissues Collected in RNAlater Preservative,” J. Mol. Diagnostics, vol. 8, no. 1, pp. 31-39, Feb. 2006.

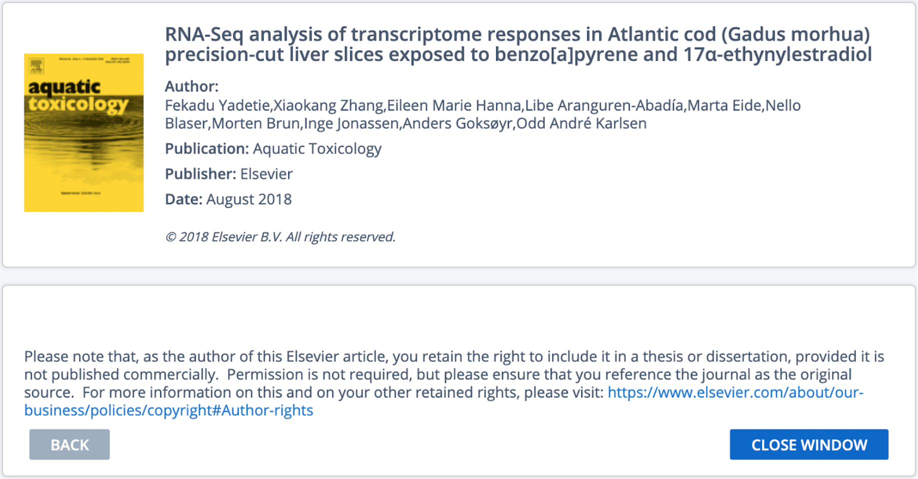
[34] D. D. Jensen and P. R. Cohen, “Multiple Comparisons in Induction Algorithms,” Mach. Learn., vol. 38, no. 3, pp. 309-338, 2000.

[35] I. Tsamardinos, E. Greasidou, and G. Borboudakis, “Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation,” Mach. Learn., vol. 107, no. 12, pp. 1895-1922, Dec. 2018.

[36] J. Demšar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” J. Mach. Learn. Res., vol. 7, pp. 1-30, 2006.

Permissions to reuse the publications in the thesis

Granted permission for Paper I



RNA-Seq analysis of transcriptome responses in Atlantic cod (Gadus morhua) precision-cut liver slices exposed to benzo[a]pyrene and 17α-ethynylestradiol

Author:
Fekadu Yadetie,Xiaokang Zhang,Eileen Marie Hanna,Libe Aranguren-Abadía,Marta Eide,Nello Blaser,Morten Brun,Inge Jonassen,Anders Goksøyr,Odd André Karlsen

Publication: Aquatic Toxicology

Publisher: Elsevier

Date: August 2018

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

Permission for Paper II

Paper II is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

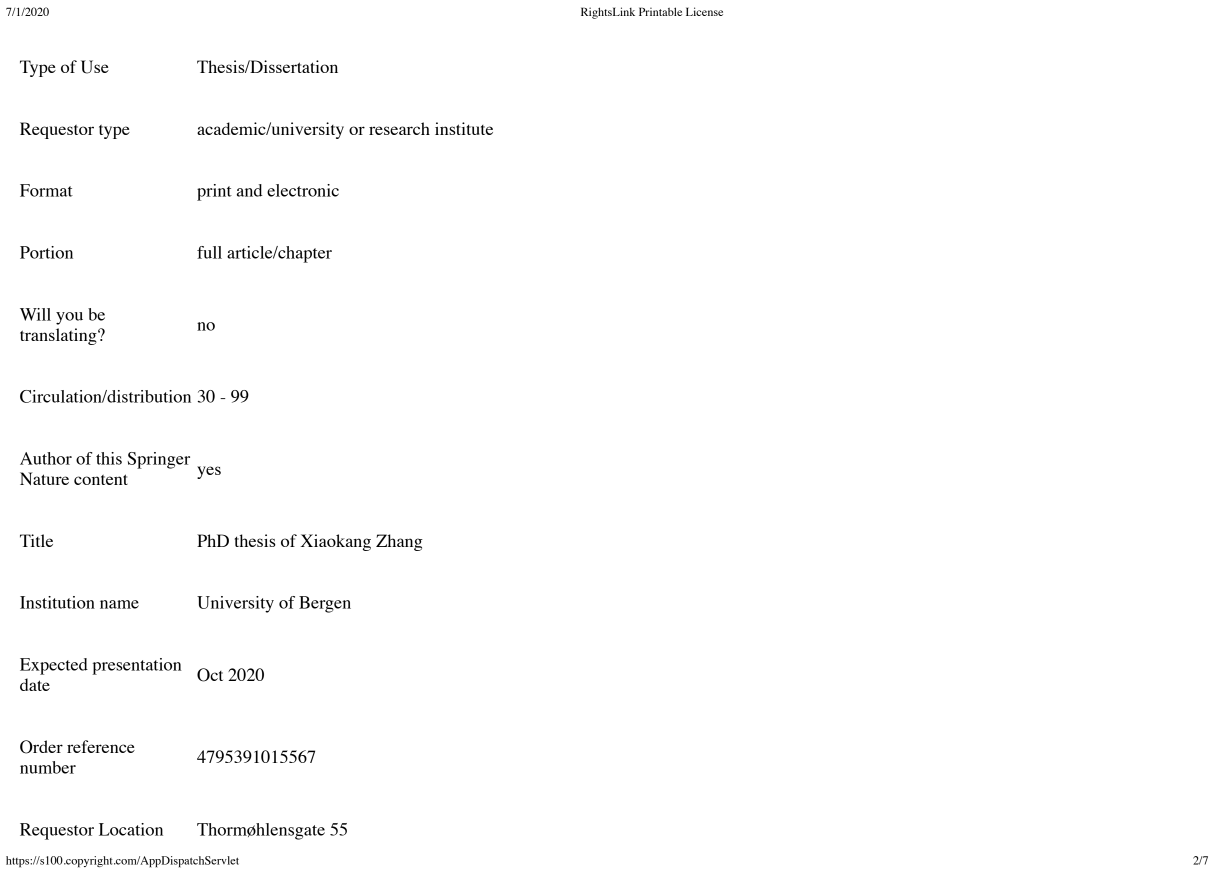
Granted permission for Paper III

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jul 01, 2020

This Agreement between Department of Informatics, UiB, Thormøhlensgate 55 ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4860230624430
License date	Jul 01, 2020
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	A Comparative Analysis of Feature Selection Methods for Biomarker Discovery in Study of Toxicant-Treated Atlantic Cod (Gadus Morhua) Liver
Licensed Content Author	Xiaokang Zhang, Inge Jonassen
Licensed Content Date	Jan 1, 2019



Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	30 - 99
Author of this Springer Nature content	yes
Title	PhD thesis of Xiaokang Zhang
Institution name	University of Bergen
Expected presentation date	Oct 2020
Order reference number	4795391015567
Requestor Location	Thormøhlensgate 55

Department of Informatics

Bergen, 5020
Norway
Attn: University of Bergen

Total 0.00 USD

Terms and Conditions

Springer Nature Customer Service Centre GmbH
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

1. Grant of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

1. 2. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

1. 3. If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Scope of Licence

2. 1. You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

2. 2. A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

2. 3. Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

2. 4. Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

2. 5. An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

3. Duration of Licence

3. 1. A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

Scope of Licence	Duration of Licence
Post on a website	12 months
Presentations	12 months
Books and journals	Lifetime of the edition in the language purchased

4. Acknowledgement

4. 1. The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

5. Restrictions on use

5. 1. Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

5. 2. You must not use any Licensed Material as part of any design or trademark.

5. 3. Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

6. Ownership of Rights

6. 1. Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

7. Warranty

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON

LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

8. Limitations

8. 1. ***BOOKS ONLY:***Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

9. Termination and Cancellation

9. 1. Licences will expire after the period shown in Clause 3 (above).

9. 2. Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

Appendix 1 — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), **[COPYRIGHT]** (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

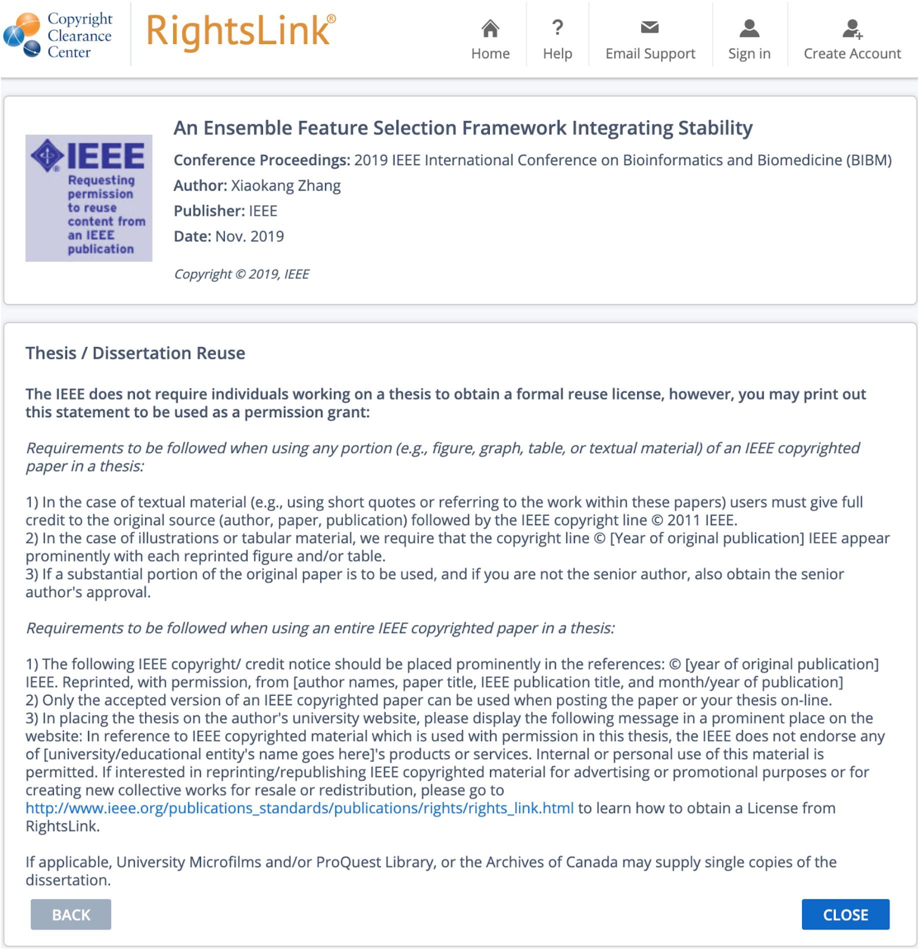
For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.2

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Granted permission for Paper IV


Copyright Clearance Center
RightsLink®
Home
Help
Email Support
Sign in
Create Account
IEEE
Requesting permission to reuse content from an IEEE publication

An Ensemble Feature Selection Framework Integrating Stability

Conference Proceedings: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
Author: Xiaokang Zhang
Publisher: IEEE
Date: Nov. 2019

Copyright © 2019, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

NORDIC ECOLABEL
241 731 PRINTED MATTER

Graphic design: Communication Division, UIB / Print: Skipnes Kommunikasjon AS

UNIVERSITAS BERGENSIS

uib.no

ISBN: 9788230860526 (print)
9788230857113 (PDF)