SYSTEMS LEVEL ANALYSIS OF NON-MODEL ORGANISMS: A TOOL FOR UNDERSTANDING ENVIRONMENTAL STRESS

by

JAANIKA KRONBERG-GUZMAN

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
College of Life and Environmental Sciences
The University of Birmingham
July 2018

Abstract

Omics techniques are changing the focus of ecotoxicology. In addition to challenges resulting from large amounts of data, there are further difficulties for non-model species: from lack of annotation to limited number of additional databases for molecular interactions and functions. In this thesis, I demonstrate the use of systems biology to relate molecular measurements to physiological parameters in non-model species in the context of environmental stress. Firstly, I make dynamical data-driven model of how gene expression changes in relation to the nerve conductance in earthworm *Eisenia fetida* exposed to single chemicals in the laboratory. The model reveals that gene expression changes might reflect the recovery from nerve damage. Using a similar approach, I use blue mussel *Mytilus edulis* sampled from their natural environment to model their annual cycle, integrating ^{1}H-NMR metabolite levels with physiological and environmental parameters. I challenge this model created from data from a reference site to see site-effects for mussels sampled from an industrial harbour. Finally, I use systems biology to relate changing chemical concentrations and traditional toxicity assays in an effluent remediation system to stickleback gene expression and morphology. I demonstrate that data-driven systems biology can help the interpretation of complex problems.

Dedicated to my son Leo

for showing me what curiosity is.

ACKNOWLEDGEMENTS

Family and friends

First, I would like to thank my son Jorge Leonardo Guzman, who has been very patient during the stages of writing and has been motivating me by producing hundreds of drawings, which according to him are figures of his "thesis". His jokes, energy and presence in every moment have made my days better! I also thank Jorge Guzman Rojas for all good times together.

I am grateful for my parents Kersti and Aavo Kronberg for loving and supporting me and from early age helping me develop an interest in biology and science. I am also grateful for my grandparents for letting me help in gardening, woodwork and housebuilding and through these activities learn about science. I am thankful for my sister Jaana Kronberg for providing me company during our childhood in Estonia and also in Birmingham. I thank Urve Kronberg for helping to look after Leo.

I am very grateful to all my teachers who have helped me in my journey to being interested in science. Tiivu Matlep and Aire Narits for helping me learn about biology and Hannes Jukk for making mathematics interesting.

I would like to thank all my friends from Birmingham and Liverpool, especially the salsa and tango communities for making my time in the UK so special. From Birmingham, I would like to thank Dr. Riddhi Shah, who in addition to being a friend, guided me in the lab. I am also thankful to Dr. Fernanda Lembo, Dr. Sonia Martins and Dr. Patcharawarin Ruanto for our lunches and coffee breaks from the lab and regular discussions after we have all left the UK. I would like to thank my housemates and friends from all times in

the UK, especially Dr. Eszter Nagy, for hiking with me and providing me encouragement at difficult times and Di Wu and Cindy Ng for all fun times together.

In Estonia, I would like to thank my neighbour Dr. Helen Eenmaa-Dimitrieva for good chats. I am very grateful to Triinu Loomus, my friend, who has helped looking after Leo. I thank Kristi Kalm for a long friendship.

Academic

Professionally, I thank my supervisor Prof. Francesco Falciani for supporting and encouraging me throughout my PhD studies and for all scientific and artistic discussions and allowing me to be involved in interesting projects.

I am also thankful to Dr. Peter Lund for his guidance at the beginning of my PhD. I would like to thank the people who attended the Systems Ecotoxicology workshop in Birmingham in January 2011 – this event sparked my interest in ecotoxicology and resulted in future collaborations. I thank Prof. Mark Viant for providing the mussel data and for useful comments about the mussel project. I am very thankful to Dr. Timothy Williams for useful input to various projects, his insight and always being very helpful. I am also thankful to Prof. Kevin Chipman, the Head of School during the time when I started working on the topics of my thesis, for allowing me be involved in the WIPE project. I also thank Dr. Ron van der Oost for his collaboration in the WIPE project. I thank Dr. Edward Perkins for providing the earthworm data and for discussions about this project. I thank Dr Natalia Garcia-Reyero for her insight and ideas.

I would like to thank Dr. Neil Hotchin, Prof. Chris Bunce and Dr. Scott White for believing in me and allowing me finish the thesis.

I thank all my friends and colleagues from Prof. Falciani's group in Birmingham and Liverpool, who made all days in the office enjoyable and provided good discussions about science and inventions. I thank Dr. Philipp Antczak for helping me with all scientific and technical questions, interesting discussions in the office and helpful advice during my time in Estonia. I thank Dr. Wazeer Varsally for introducing me to salsa. To Dr. Nil

Turan-Jurdzinski I am thankful for discussions and good coffee and to Dr. Anna Stincone for keeping me company on Saturdays and teaching the basics of laboratory work. I am thankful to Dr. Rita Gupta for challenging me mathematically and to Dr. John Herbert for always helping with Perl and Linux. I would also like to thank Dr. Kim Clarke and Dr. Peter Davidsen for good company in the office. I am thankful to Dr. Helani Munasinghe, for the encouragement that it is possible to finish the thesis while being a mother. I also thank the rest of the group: Dr. Peter Li, Dr. Danilo Basili, Michaela Bonomo, Dr. Xiaoliang Sun and Dr. John Ankers.

Finally, I would like to thank all people from the Estonian Genome Centre (Institute of Genomics, University of Tartu, Estonia), especially Prof. Andres Metspalu and Dr. Tõnu Esko, for being supportive for my PhD work and providing many new challenges and ideas. From my colleagues, I would like to thank Prof. Krista Fischer, Dr. Reedik Mägi, Maris Alver, Nele Taba, Kristi Läll and Merli Mändul for lunchtime chats. I also thank Mart Kals, Tõnis Tasa and Dr. Toomas Haller for good discussions. I am thankful to Dr. Silva Kasela and Dr. Natalia Pervjakova for inspiration and encouragements.

Financial

I acknowledge the support from the Biotechnology and Biological Sciences Research Council (BBRSC) for the funding of my PhD studies.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

GENERAL INTRODUCTION

1.1 Introduction

1.1.1 Problem setting

Increase in population and technology development represent a serious challenge for the environment. For example, personal care products, pharmaceuticals and industrial chemicals enter the aquatic environment through wastewater treatment plants. Plant protection products such as pesticides and fertilisers can enter soil and water directly from the sites of use. In addition to aquatic and terrestrial pollution, air pollution generated by traffic, heating and manufacturing is also a significant environmental issue. Chemicals can affect organism's health in the whole biosphere and affect ecosystem functions. For example they affect reproduction in aquatic organisms [310, 129], reduce microbial biomass in soil [14, 340], affect health, reproduction and behaviour of birds ([253], reviewed in [100]), [330] and affect various mammals, from mice [246] to polar bears [293]. Pollution has also been linked to human health ([245], reviewed in [162]). For example, pollutant exposure has been associated with cancer [17] and Alzheimer's disease [263, 108]. Exposure to some persistent organic pollutants (POPs) has been shown to increase the risk of type 2 diabetes [187, 188] and cardiovascular disease (reviewed in [42, 143, 79]).

The problem of pollution can be approached from the legal perspective, enforcing

countries to adhere to norms of chemical release and requiring to achieve certain standards. For example, in the European Union, the European Water Framework Directive [77] has been implemented and one of the aims for 2015 was to achieve "good status" of surface waters. However, legislation should stem from knowledge. To enforce norms for chemical concentrations or ban the use of certain chemicals completely, the science supporting legislation should be robust. For example, the ban of neonicotinoids in 2018 follows many scientific studies showing the effects of these chemicals [350, 345, 222, 161].

Some chemicals have been banned in the past, such as tributyltin which was used as anti-fouling paint on ships, but the ban only took place in 2008 [116], many years after the effects were seen and published on organisms and population level, including imposex and population decline [215, 44]. Similarly, the ban of DDT starting with Sweden in 1970 [321] followed studies showing its effects on the environment [253, 139]. These three examples from different times are all similar: a chemical is used, its effects are seen and studied and the chemical is banned.

However, many chemicals are useful or necessary and for example, in June 2018, there were 21551 chemicals registered to be used in the European Union [86]. Pesticides from agriculture, plasticisers from home appliances, toys and packaging, flame retardants from furniture and microplastics from shopping bags and cosmetics, medicines from both domestic effluent and hospitals, industrial chemicals – these are just some examples of chemicals entering the environment from anthropogenic sources and have the potential to affect different organisms, including humans.

Therefore, in addition to knowing, what chemicals are in the environment and trying to reduce the quantities by reducing the use of chemicals, replacing chemical-based methods with alternatives, and treating both landfill run-off and wastewater effluent, it is important to understand their effects.

If a chemical represents a hazard for the ecosystem and for human health, it should be strictly regulated or ideally, should not be used at all. However, in many cases, first effects start appearing and later a chemical is linked to these effects via studies performed,

like was the case of DDT, neonicotinoids and tributyltin. With the current large number of chemicals in use, there is a large potential for unwanted effects, including effects on non-target organisms and general mixture effects for different chemicals. These examples have shown that:

1. It is necessary to understand the effects of chemicals to support legislation about their use and permitted concentrations

2. It is important to understand the mechanism of action of a chemical across multiple species.

3. Genomics can be important to elucidate a mechanism of action or to identify potential biomarkers. To make sense of these data, it is necessary to link the molecular response of an organism to exposure to higher levels of biological organisation (organism health and physiology, population changes)

4. It is necessary to develop tools that are able to detect possibly harmful environmental events as early as possible (molecular biomarkers, rather than population collapse)

The first point is regulated through REACH [93] which is in force since 2007 and requires a company to identify and manage risks associated with manufactured or imported chemicals before they can be registered in the European Union. European Chemicals Agency (ECHA) is responsible for further evaluation of chemicals. ECHA also provides test guidelines which are approved by both OECD and the EU, including testing for effects on human health, ecotoxicity and environmental fate [87]. Currently, a hazard assessment of a chemical is performed for human health, physicochemical and environmental hazards and also assessed for being persistent, bioaccumulative and toxic (PBT) and very persistent and very bioaccumulative (vPvB) [93]. In the US, the risk assessment guidelines are provided by the US Environmental Protection Agency [324].

However, it is not possible to evaluate the effects for all species and often the effects are assessed for acute exposure, which is not environmentally relevant. The importance of

extrapolating effects on lab-species on which the tests are performed to different species in the environment has been highlighted [40]. The second and third points refers to the need of understanding how molecular changes relate to other endpoints. Molecular changes seen in omics studies and traditional assays show changes at organism level, but especially for the interpretation of molecular measurements, it might be useful to relate these to more understandable endpoints. The fourth point of developing tools for detecting early molecular changes is an extension of the previous one, as when there is understanding of how chemical stressor affects the organism at molecular level and how these molecular changes relate to physiology and behaviour, it might be possible to use the molecular markers as early warning signs for unwanted changes at individual health, population or even ecosystem level.

1.1.2 Road map for the thesis

In this thesis, systems biology methods are used for the analysis of omics datasets from non-model species as a tool for understanding environmental stress. By using several non-model species to study different types of environmental stress, I show that system biology is a powerful tool for relating molecular changes to physiology and environment.

In the first chapter, I wish to develop dynamical models linking earthworm transcriptional response to chemicals to nerve conduction to identify mechanisms of neurotoxicity. More specifically, earthworms (*Eisenia fetida*) are exposed to two neurotoxic chemicals, an explosive that is released in the environment during army training exercises, and an insecticide. Gene expression is measured before and during exposure and after the removal of the chemical. In addition to this, the conduction velocity of the medial giant nerve fibre is measured, as an indicator of nerve damage. The aim in this chapter is to explore, whether it is possible to integrate gene expression analysis to a physiological parameter and whether the results are biologically meaningful. Indeed, the results indicate that systems analysis of gene expression can be informative of potential biological pathways affected during the chemical exposure. Moreover, the modelling approach con-

nects transcriptomic changes with a physiological change and provides hypothesis that the transcriptomic changes might primarily reflect the recovery from nerve damage.

In the second chapter, I model the metabolic profile of blue mussel *Mytilus edulis* over the annual cycle in relation to environmental and physiological parameters. This model is then used to study how the metabolic profile differs in a more polluted site. The analysis of the second chapter is building on the experience gained in the first chapter, particularly in the approaches used for reducing the dimensionality of data. The resulting model of the annual cycle reveals that metabolites upstream of gonadal stage are affected between a reference and a more polluted site. A more detailed machine learning approach suggests that some male mussels in the more polluted site have a female-like metabolic profile.

The third chapter is about using systems biology to gain insight into the process of wastewater remediation. In this study, gene expression of the three-spined stickleback (*Gasterosteus aculeatus*) was used to learn about the effects of remediation. The ultimate aim was to understand how changing water quality as indicated by chemical concentrations affects the gene expression. To address this aim, first the components of chemical concentrations and gene expression were analysed separately. Secondly, all available data, including chemical concentrations, gene expression, toxicity tests and physiological parameters was integrated into static similarity networks. Thirdly, a Bayesian model was used for modelling how gene expression changes in each of the remediation stages, also taking into account the chemical concentrations. The results of this chapter show that especially polycyclic aromatic hydrocarbons (PAHs) are associated with gene expression in the stickleback liver and that PAHs are a group of chemicals which decreases during the remediation in most sites. The Bayesian model further shows that chrysene (a PAH) is particularly important in driving the gene expression.

Thesis chapters are ordered in a logical order from most simple lab exposure to the complex wastewater remediation system with many chemicals. Although the scenarios and even omics platforms are different and they have specific aims, their overall aim is to interpret the effects of environmental stress to the organism as described by molecular

level and how these omics measurements relate to physiological parameters.

1.2 Relevant species

Lab species can be used for studying the effect of chemicals, as for these species, many pathways are known, making them especially suitable for understanding specific molecular responses However, it is also important to understand how various pollutants affect organisms in the environment.

In this thesis, three species are used: earthworm *Eisenia fetida*, blue mussel *Mytilus edulis* and three-spined stickleback *Gasterosteus aculeatus*. Two of them (earthworm and mussel) are invertebrates and the stickleback is a vertebrate. All of these species are relevant for studying the response of chemicals. The first reason for this is that they are all present in a wide range of locations. The second is that all of them have a specific characteristic making them particularly relevant for studies in ecotoxicology.

Earthworm is an annelid that lives in soil, is exposed to many chemicals in terrestrial ecosystems and have been used as bioindicators [239]. They have been called "soil engineers" because of their burrowing activity [54]. While various survival, reproduction and growth parameters [295, 230, 328, 267] in earthworms have been used as indicators of chemical exposure, there are also molecular markers indicative of chemical exposure [51, 303, 206, 11].

Mussels are sessile, filter-feeding molluscs which can filter large amounts of water [264], which is the reason it is suitable for studying the effects of aquatic pollution. The use in mussels in environmental monitoring has long history since they have been used in the Mussel Watch Program [118]. In mussels, physiological parameters, such as scope for growth and survival stress tests have been used [135, 115]. Despite some studies attempting to use mussels for studying the effects of chemicals which act as endocrine disruptors in other species, there is currently no evidence that these chemicals have effect on mussels [283, 281, 282] and the presence of vertebrate sex stereoids have been suggested

to be due to uptake from water [280].

Stickleback is a vertebrate, a fish that has been suggested to be suitable for studies in ecotoxicology and particularly for studying endocrine disruption, as it can be used to detect the effects of both estrogens and androgens [158, 131]. Specifically, in response to androgens, spiggin (a glue protein used for nest building) is produced in the kidney of male sticklebacks and while females normally do not produce this protein, it is made in response to androgen exposure, making it suitable as a biomarker for environmental androgens [131, 158]. Vitellogenin (which is a female-specific yolk protein induced by estrogens) can also be measured in sticklebacks and in males, this is a biomarker for estrogen exposure [131].

1.3 Omics for understanding environmental stress

Traditionally, environmental stress and effects of chemicals have been assessed using endpoints at organism of population level, including growth effects, mortality or changes in population structure. As technology developed, changes could be observed at histological level and also single molecules could be measured. However, with the advancement of high-throughput omics technologies, it is possible to measure changes in thousands of features in many layers of biological organisation, from DNA to mRNA to proteins and metabolites. The ultimate aim in using molecular measurements for monitoring or risk assessment would be to use the earliest changes as a warning or prediction of what would happen at later timepoints, often seen at tissue, organism or population level.

1.3.1 Transcriptomics

Transcriptomics is the high-throughput measurement of mRNA levels. The basic mechanism of mRNA measurement relies on binding of labelled single-stranded oligonucleotides to a complementary oligonucleotide strand on a glass slide of many oligonucleotides, each with specific known sequence and location.

Initially, simultaneous expression profiling of many genes was done using cDNA arrays [278], where cDNA probes were spotted on glass slides by robots. Later, the cDNA technologies were replaced by oligonucleotide arrays [200], GeneChip (Until 2016, Affymetrix) [154] and Agilent [4] being the most common. These technologies both use chips with thousands of oligonucleotides in specific locations, but the manufacturing processes of these arrays are different. GeneChip uses oligonucleotides that are 25bp long (25-mer probes) which are synthesized on the chip by photolithography, a process using UV light and mask, allowing the "deprotection" of each DNA strand that is to be added a new nucleotide in each step. For the addition of each new nucleotide, a new mask is used, "deprotecting" only the strands which are to be added the relevant nucleotide in this specific step. Solution containing the relevant nucleotide is then added – the nucleotide can only bind to the "deprotected" strands. The synthesis of all 25-mers on the chip can take up to 100 such steps [311]. Overall, the GeneChip technology represents each gene by 8-16 pairs of (perfect match and mismatch) 25-mers.

The process of Agilent is different, using SurePrint Technology, a technique similar to an ink-jet printer. Oligonucleotides are synthesized in situ by printing each individual nucleotide at a time directly on a glass slide (reviewed in [111]). This technology has enabled customers to easily design custom arrays and has made an impact especially for non-model species.

The processes of mRNA profiling are similar. In the case of Agilent, after mRNA isolation, cDNA is synthesised from mRNA by reverse transcriptase and this is amplified by T7 RNA polymerase with Cy-3 or Cy-5 in the solution, resulting in labelled cRNA which can then be hybridised to the chip. During incubation, cRNA is hybridised to probes on chip with complementary strands and excess unbound cRNA is washed away. The chip can then be scanned and the intensity of light emitted by fluorophores measured. The GeneChip process consists of the synthesis of cDNA which is then labelled with fluorescent tags and hybridised to probes on the chip. In both technologies, there is possibility to use dual-colour or single colour labelling of different samples (pairs of disease-control or

treated-control) which are then hybridised on the same chip.

RNA-seq is a newer technology that can also be used for transcriptomics profiling. The use of RNA-seq in publications has increased exponentially since 2010, reaching the same number of articles using this technology as RNA microarray in 2014. However, the use of gene expression microarrays has been decreasing since 2013 [203]. For RNAseq, mRNA is fragmented and using a reverse transcriptase, double-stranded cDNA is made, which is then sequenced by high-throughput methodologies. The read length differs depending on the sequencing technology used, but can be in the range of 30-10000bp (reviewed in [203]). RNA-seq has the advantage over gene expression arrays, that it can quantify genes with low or high expression more accurately. Compared to expression arrays, it also has the advantage of being able to detect the expression of new genes and splice variants [225, 228]. Compared to expression microarray, where the microarray has to be designed first, RNA-seq has the advantage that the results are not dependant on the design. This is the main advantage of RNA-seq for environmental studies, in non-model organisms [89]. However, as RNA-seq was more expensive at the time of the gene expression experiments in this thesis, cost was one of the major disadvantages, especially as many of the non-model species had not been sequenced yet and even in the case of de novo assembly or the transcriptome, many transcripts would have had low quality annotation, if any.

1.3.2 Metabolomics

Metabolomics is the characterisation of small molecules in the organism in a high-throughput manner. Main methods used for the characterisation of metabolite levels in an organism are nuclear magnetic resonance (NMR) of mass spectrometry (MS) (reviewed in [262, 3]).

^{1}H-NMR is a technique that uses the properties of protons when an external magnetic field is applied. With the application of strong external magnetic field, the spinning is either in the same or opposite direction of the magnetic field, resulting in two energy states for each proton, lower and higher. The energy difference depends on the strength of the external magnetic field and also the magnetic fields of other protons. When radio

frequency is applied, the proton enters a state of high-energy spin and when relaxed, returns to the original spin and emits radiation.

Resonance is achieved when the protons are entering the high-energy spin continuously, meaning that energy of the applied radio frequency is equal to the energy difference between spins [132]. Chemical shift frequency of resonance compared to a standard, which for ^{1}H-NMR is tetramethylsilane [84]. Chemicals shifts, after further processing are normally used in ^{1}H-NMR analysis by either unknown bins or annotated metabolites based on existing libraries of single metabolite standards. For non-model species the main advantage of ^{1}H-NMR has been that the technique does not rely on genome sequence of annotation and existing metabolite libraries can be used for identification of single metabolites in many species. One of the main disadvantages is the low number of annotated metabolites. However, chemical shift bins have been used for machine learning methods and also have been integrated with other omics methods, making the interpretation easier.

Another frequently-used method for metabolomics is mass spectrometry. For this, metabolites need to be ionised and in the mass-spectrometer, the ions are sorted based on their mass-charge ratio. In the mass-spectrometer, the ions are accelerated and either electric or magnetic field is applied, the ions' trajectory changes, which is called deflection. The deflection depends on the mass of the ion and also charge and by detecting the deflection, the mass-charge ratio can be calculated. The mass-charge ratios with relative abundance of each ratio are the output of mass spectrometer. Frequently, the process of preceded by another separation step, for example gas chromatography or liquid chromatography [75].

The main advantage of mass spectrometry is the ability to identify more metabolites than ^{1}H-NMR.

1.3.3 Other high-throughput technologies used

Although transcriptomics and metabolomics are often used in environmental studies, other methods have also been used. Microbiome as the characterisation of different microbial species living in the gut, skin or other place or an organism, often characterised by operational taxonomic units or different types of diversity. There are many studies in human health relating microbiome to various diseases, environment and diets, but it is also used in environmental studies. For example, microbiome differences have been studied in different fish in locations of varying pollution [136]. Lipidomics [344] is the study of lipids and has also been studied in response to chemical exposure. For example, bisphenol S has been shown to disrupt lipid metabolism [360]. Proteomics has also been used in environmental studies [315, 8, 355].

1.3.4 Comparison of different omics methods

Different omics methods provide different types of data. All of them have their advantages and disadvantages. Transcriptomics and proteomics, for example are easier to interpret, but require species-level sequence information, either to design the microarray, or to align short reads from next-generation sequencing, unless de-novo transcriptome assembly is performed in the case of RNA-seq. If knowledge exists about the function of genes or proteins, or they can be aligned to other species, for example model species, then there are many methods for interpretation, for example by calculating the enrichment of Gene Ontology or KEGG terms in a set of genes or proteins.

Metabolomics data, on the other hand does not require species-level sequence or metabolite data to be available and if for example ^{1}H-NMR metabolomics has been performed on a number of single metabolite standards, they can be predicted in a mixture of metabolites from differen species. However, the number of metabolites to be annotated from ^{1}H-NMR data is relatively small. Mass spectrometry can provide annotation of larger number of metabolites, but with increased cost. In Table 1.1, main differences

between transcriptomics, proteomics and metabolomics methods have been highlighted.

1.3.5 Omics methods used in this thesis

In this thesis, transcriptomics was used for second and fourth chapter and metabolomics for the third chapter. Although in the case of this thesis, data for the third chapter was already generated by the time I started the PhD, transcriptomics is suitable for the biological problem, as it allows more thorough interpretation of biological pathways involved than metabolomics. Although full genome was not available for the earthworm at the time this project started, there were ESTs, which could be annotated against more well-annotated species. Gene expression microarray, as opposed to RNA-seq, was used as the experiments had been performed in 2010, and the first published articles describing RNA-seq in model organisms were from 2008. In comparison, gene expression microarrays were still widely used for non-model species at this time. The array design for *Eisenia fetida* was published in 2010, using data from both previous Sanger sequencing and author's own unpublished 454 high-throughput experiment. Moreover, the authors state that high-throughput sequencing would be used more in the future, when they become more affordable and acceptable, implying that at that time they were still too expensive [125].

For the fourth chapter, microarrays were used, as for stickleback, sequences could be annotated against zebrafish and other vertebrates, which would provide opportunities for interpretation. As the experiments started in 2011, it was still more cost-effective to use gene expression microarray, as opposed to RNA-seq. Moreover, for non-model species, the main advantages of RNA-seq, the ability to detect novel genes and splice variants, although interesting, were not as relevant, as it might not have been possible to annotate novel genes.

For the third chapter, metabolomics was used. Although for data for this chapter had also been generated before the start of my PhD project (mussels were collected in 2004-2005, and descriptive metabolomics were part of Adam Hines's PhD thesis submitted in

Table 1.1: Comparison of transcriptomics, proteomics and metabolomics

Method	Transcriptomics		Proteomics		Metabolomics	
	Microarray	RNA-seq	Mass spectrometry	Olink	^{1}H-NMR	Mass spectrometry
What information?	mRNA levels	mRNA levels	protein levels	protein levels	metabolite levels	metabolite levels
Cover range	Only transcripts on chip (depends on chip design and previous knowledge)	All transcripts, not dependant on previous design	All proteins or subset	Subset, depends on chip design	Metabolic bins of whole metabolome of small molecules	Whole metabolome or subset (for targeted analysis)
Annotation	For many transcripts; for non-model species, depends on sequence homology	For many transcripts; for non-model species, depends on sequence homology	For many proteins; for non-model species, depends on sequence homology	For many proteins; for non-model species, depends on sequence homology	Small number of annotated metabolites; same quality for both model and non-model organisms	Larger number of annotated metabolites; same quality for model and non-model species
Need for sequence annotation for the species to be studied	+	+	+	+	-	-
Need for method-specific annotation	-	-	-	-	+	+

2008, ^{1}H-NMR metabolomics, despite its disadvantage of limited metabolite annotations, was a suitable choice, as this technology does not require designing species-specific microarray. Moreover, as multiple species were used, designing the microarray would have had to take into account both populations and different species of mussels present. In this light, RNA-seq would have been preferable over microarray, but this was not possible at the time of the project. Of course, at the time of submission of the thesis, it is quite common for published studies to combine multiple types of omics data, and for example, the inclusion of transcriptomics could have improved interpretation in the mussel chapter.

1.3.6 Systems biology as a key to analysing omics datasets

Systems biology is a field of biology aiming to describe biology at the systems level [165]. It relies on using advanced technologies for measuring biological molecules and computational analysis for generating biological knowledge [165]. The aim of systems biology is to analyse all parts of a system together, as opposed to by analysing differences in a single molecule. The term "systems biology", although mentioned even before [145] the completion of the Human Genome Project [63]), started increasing to appear in scientific publications after the completion of the Human Genome Project and peaked in 2013, reaching 45000 publications with this keyword (Figure 1.1 B), following the peak of genomics (Figure 1.1 A).

Systems biology is a wide term, and can be understood in multiple ways. In the broadest term, systems biology should incorporate data-driven data analysis. Depending on field, this can be done using statistical modelling, networks, machine learning: using data to describe the system. Ideally, systems biology should incorporate phenotypical outcomes with omics data [152]. It has been suggested, that in the future, integration of data from multiple omics layers will help understanding disease mechanisms and is able to account for both genomics and environmental factors [356]) and the importance of multi-omics methods in scientific publications reflects the importance of multiple omics layers (Figure 1.1 C).

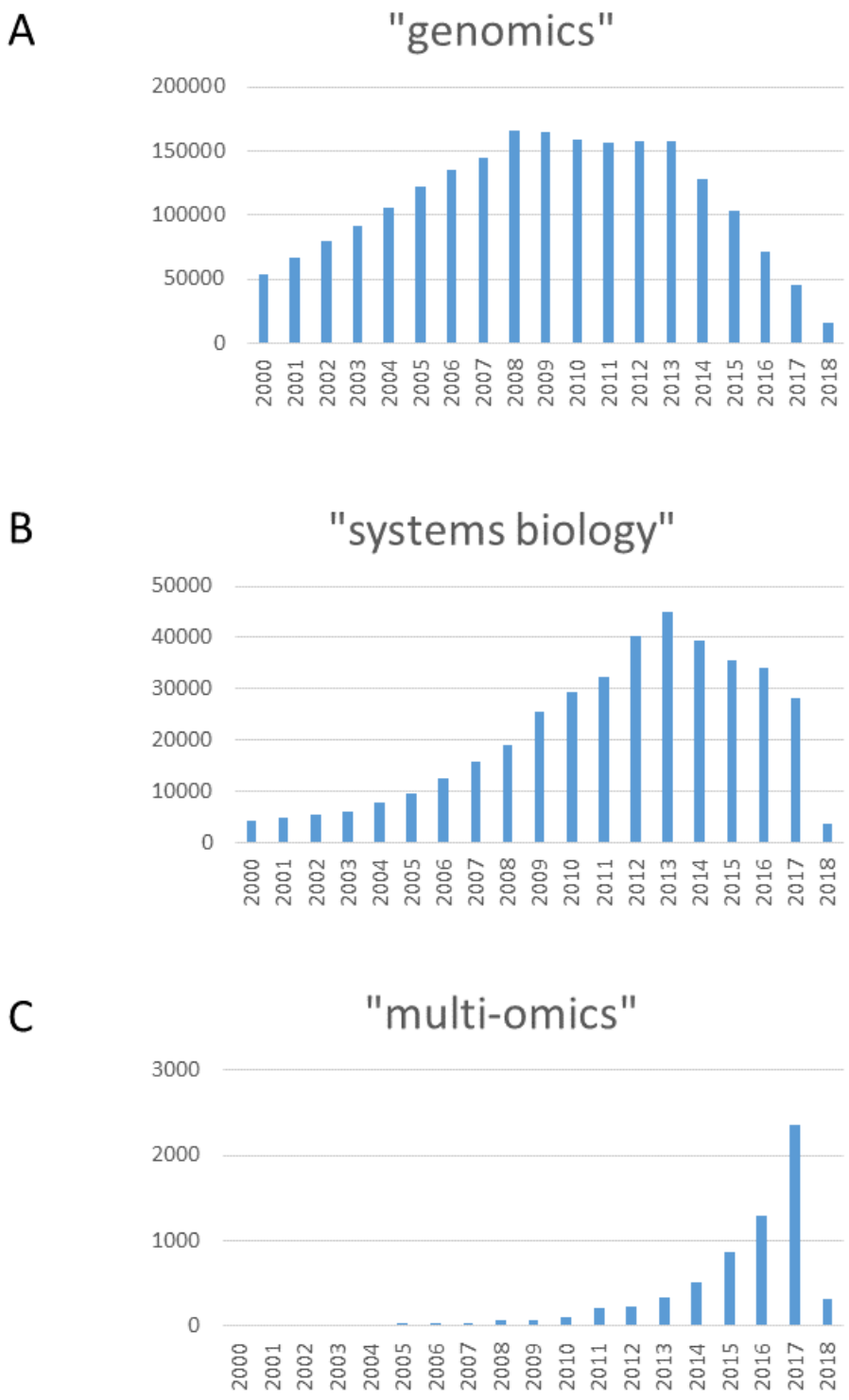


Figure 1.1: Number of publications for each of the search terms in Google Scholar. A: search term "genomics", B: search term "systems biology" and C: search term "multi-omics". Raw data for these plots was extracted from Google Scholar PLotR [209] on February 6th 2018.

As has been said before the term systems biology existed: "Essentially, all models are wrong, but some are useful" (George Box), systems biology is just a first step in understanding the problem and data and as real life is always more complex than possible models, models resulting from different methods can and should be different. However, they should lead to a better understanding of the system and a testable hypothesis.

Systems biology has shown potential to help addressing many complex biological problems, including drug discovery (reviewed in [46, 146]) and repurposing [119]. Systems biology has also been used to discovering markers associated with diseases, such as type 2 diabetes mellitus [216] or coronary heart disease [5]. Currently, with large amounts of data generated, hopes are high for systems biology to also advance medicine [242, 146, 231].

As systems biology developed as a field, many methods were published and mostly used in model organisms. For example, Cytoscape [285] made it easy for biologists to create and visualise biological networks, for example connecting genes or proteins that have been shown to interact or are co-expressed. Databases of biological interactions, such as STRING [335, 309] further contributed to creating of biological networks based on existing knowledge in public databases. Additionally algorithms allowing the reconstruction of biological networks based on gene expression data (such as ARACNE [212]) brought new opportunities. All these developments made it possible to address most problems in a data-driven manner. Instead of concentrating on a set of molecules, looking at specific pathways, it is possible to create network of differentially expressed genes and interpret these using for example Gene Ontology enrichment [62]. As the methods were developed and initially used for model species, their potential started to be utilised for non-model species of environmental relevance a well. These studies faced additional challenges, especially for the interpretation of results, as for non-model organisms, the annotation was not of the same quality as for model species. However, the difficulties could be overcome by mapping transcripts of non-model species to their orthologs in other species, for example using BLAST2GO [59].

With omics technologies maturing, it is possible to integrate multiple layers of bio-

logical information. This has already been demonstrated in both human health specific studies and several tools have been proposed and compared for such integration [313]. Multiple omics layers have also been integrated for environmental scenarios. For example, in flounder, the integration of metabolomics and transcriptomics has allowed prediction of sampling sites [348]. In largemouth bass, the integration of transcriptomics with physiological endpoints has allowed to identify novel endocrine disruptors for this species [25].

1.4 Data analysis

1.4.1 Finding differentially expressed genes

Univariate statistics is commonly used as the starting point for the analysis of omics data. Very often, t-test is used to find out which of the variables are different between two groups, usually treatment and control. The advantages of t-test are its simplicity to understand and being implemented in various statistical tools. As more advanced methods to identify genes or metabolites differing between groups, SAMR [312] could be used. SAMR is non-parametric, permutation-based method that was developed specifically for finding differentially expressed genes. In addition to two-group analysis, SAMR allows the analysis of multiple groups and also time-course data. An alternative, limma [289, 265], takes advantage of replicate spots within array, and uses Bayesian statistics. Although at the time of submitting the thesis, limma [289, 265] is more widely used, at the time of starting my PhD, SAMR was more well-known (in 2010, SAM method which is used in the SAMR package had been cited 7180 times, while limma had been cited 959 times). Limma and SAMR also work with RNA-seq data, and for RNA-seq data, edgeR [268] is an alternative also using Bayesian statistics.

Table 1.2: Comparison of different methods for finding differentially expressed genes

	t-test	Wilcoxon signed-rank test	SAMR	limma
Availability in R	R core (t.test) [260]	R core (wilcox.test) [260]	samr package [312]	limma package [288]
Normal distribution assumed?	yes	no	no	no
What does it do for 2 related samples	Do their means differ statistically	Do the means of the ranks differ statistically	Do their means differ statistically (using permutations)	Do their means differ statistically (using within-array replicate spots and Bayesian statistics)

1.4.2 Exploratory data analysis

For further analysis of data, after finding significantly different genes or metabolites, various multivariate methods can be used. Omics data is multi-dimensional, each gene, protein or metabolite being a dimension. It is often useful to reduce this dimensionality and explore the main characteristics of the data. Principal Component Analysis (PCA) is an option often used for this. PCA works by transforming multi-dimensional data into a new set of orthogonal basis so that the first base describes the most of the variability of the dataset and the next basis follow to describe most of the remaining variability so that all basis are orthogonal.

Principal components describing most of the variability can then be visualised to see whether sample groups, for example treatment and control differ based on their gene expression. Outliers can also be detected by PCA visualisation.

As part of exploratory analysis, clustering can also be used. The simplest clustering approach is to produce a dendrogram describing the similarities in the dataset by grouping the variables to be clustered hierarchically. This kind of clustering can for example be produced by hclust algorithm in the Stats package in [260]. Clustering groups samples

or omics measurements (genes or metabolites) based on a given distance. For example, correlation or Euclidean distance can be used. When clustering both samples and omics measurement, sometimes a heatmap visualisation is also created. Hierarchical clustering is an unsupervised clustering method. A dendrogram can be used for exploratory purposes, but it can also be cut at some level to produce clusters to be used in further analysis.

A special hierarchical clustering method, HOPACH [327, 251] algorithm is also available. It works by recursively splitting or merging clusters at each level and uses Silhouette function for defining the optimal number of clusters.

1.4.3 Biological networks

Network is a framework that allows the analysis of multiple types of biological data. It can also be used for analysing correlation matrices in a more complex way than clustering. In a network (also called a graph by mathematicians), entities are represented as nodes (also called vertices) which are connected by edges (Figure 2 A and B). Networks can be used in different fields, for example in a social network, all nodes are people and edges represent whether they are friends and in an air traffic network, all airports are nodes and the connecting flight are edges.

The study of networks for biology and other fields and also their topological properties were established in the works of A-L. Barabasi and R. Albert [21, 20, 22]. Since then, and especially after Cytoscape [285] was developed, networks have been used extensively in biological studies, as suggested by the number of times the Cytoscape article [285] has been cited (11058 citations in June 2018). This is half the number of citations the article of the human genome [63] has (22073 in June 2018). Comparing these with the number of citations the Sanger sequencing article [273] has since 1977 (72884 in June 2018), network biology definitely has potential to play an important role in current and future research.

In a biological network, genes or proteins are usually represented by nodes and interactions between nodes are represented by edges (Figure 1.2 A and B). Static networks, which are networks where edges do not have a direction (Figure 1.2 C) can be made based

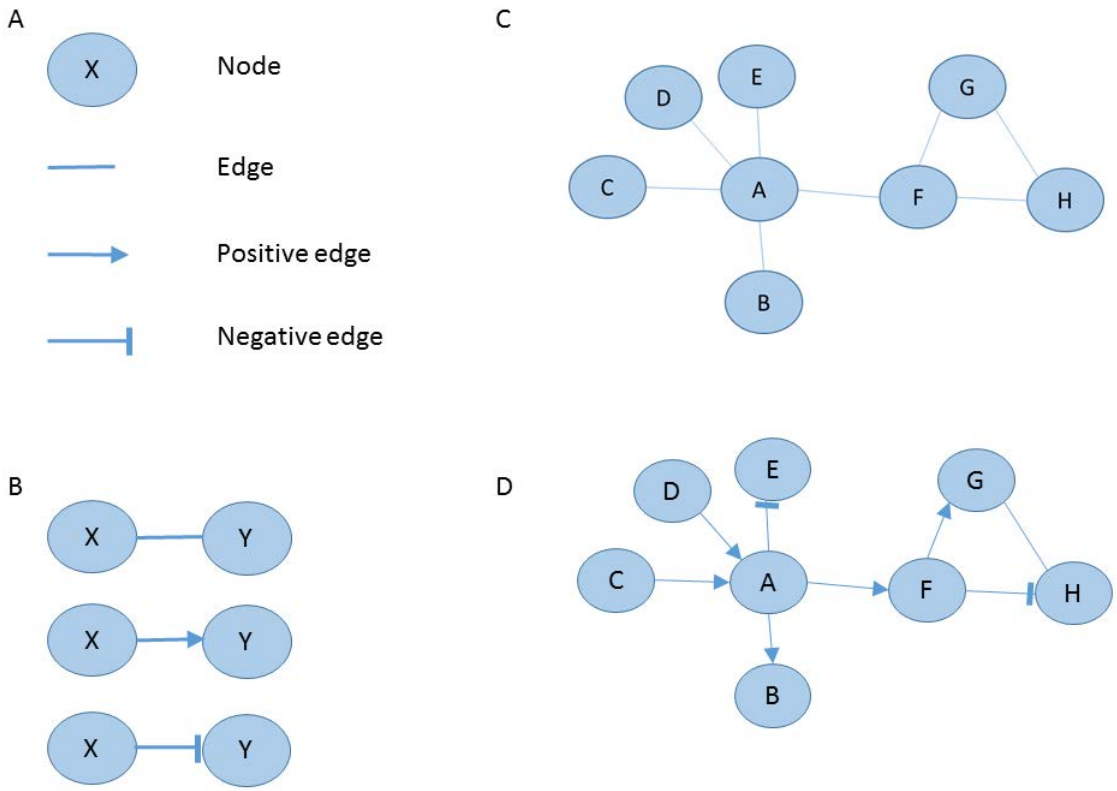


Figure 1.2: Networks for the representation of data. Panel A shows components of a network. Panel B shows different types of interactions between nodes: X and Y are associated, X has positive effect on Y (for example, activation), X has negative effect on Y (for example, inhibition). Panel C shows a static network. Panel D shows a dynamical network.

on known interactions, such as using a database containing protein-protein interactions or co-expression of genes based on public data. One of such databases is the STRING database [335, 309]. Although containing data for several species, it is more suitable for model organisms. An option for creating a network for a set of genes or proteins for non-model species would be to first map these to a model species and then use interactions of this model species for creating interaction networks.

An alternative is to create the network from data. The process of inferring the network from biological data is called network inference. In 2010, it was estimated, that the rate of the number of network inference methods was doubling every 2 years [211]. Comparison of network inference methods is shown in Table 1.3.

However, despite the number of methods available in 2010, the only gold standards to compare the performace of an inference method against were in silico networks [211], firstly due to nonexistence of a biological network with enough knowledge, and secondly, the recognisability of such network if it existed [301, 211]. However DREAM3 challenge provided an in silico gold standard [211] that was used for evaluating the performance of 29 methods used in the DREAM3 network inference challenge. In the challenge, 29 participating methods could be grouped by their predominant approach: correlation-based methods, information-theoretical methods, Bayesian methods and methods based on dynamical models [211]. Interestingly, the top performers included methods from each of the classes, but representatives of each of these main classes were also amongst lower-performing entries. The best performed methods included knock-out and perturbation data [354], and several best-performing teams included all available data, including time-course. Interestingly, one of the main conclusions from this challenge was that in addition to the importance of integrating different types of data, community predictions consisting of the consensus of predictions from different methods can also outperform single methods [301, 211, 210].

Table 1.3: Comparison of different methods for network inference

	Correlation	Mutual information	Methods with time-delay	Bayesian methods	Dynamical models
Specific examples	Pearson correlation, Spearman correlation	ARACNE, CLR	Time-delay Spearman, TDARACNE	ARTIVA	NIMOO
Type of data used	perturbation matrix	perturbation matrix	time-course	perturbation matrix (Bayesian Network) or time-course (Dynamic Bayesian Network)	time-course
Network type	undirected	undirected	directed	directed	directed
Signed	yes	no	yes	yes	yes
Network size	very large	very large	large	moderate	moderate
Main advantage	simplicity, speed	simplicity, speed	speed, directed graph	directed graph	directed graph

1.4.4 Static network creation

The simplest way for network inference is to calculate correlations between genes. An example is network creation from a gene expression correlation matrix is shown in Figure 1.4. Correlation is a statistical measure, that can describe the association between two variables. Two commonly used correlation methods for network inference are Pearson and Spearman correlation. Pearson correlation describes the linear relationship between two variables, and can be described by the following formula:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$$

where $x_1 ... x_i$ and $y_1 ... y_i$ are variables for which correlation is calculated and $\overline{x}$ and $\overline{y}$ are means of x and y.

Pearson correlation is suitable for normally distributed linear relationships. Another correlation measure often used is Spearman correlation, r_s, which is Pearson correlation of ranks of two variables. Spearman correlation is non-parametric and suitable for data for which statistical distribution is not knows. It can also capture non-linear relationships [69].

Examples of correlation are shown in Figure 1.3. For example, on Figure 1.3, panel A shows perfect correlation between a and b where all points are on a line, and on panel B, variables are not correlated at all. Panels C and D show high positive and negative correlation. Panel E on Figure 1.3 shows non-linear relationship between two variables, where Pearson correlation $r = 0.82$ and Spearman correlation $r_s = 0.89$.

After correlations have been calculated between all genes, it is necessary to decide on a threshold value of which edges are statistically significant, for example by finding a correlation coefficient that is significantly different from correlations calculated for a matrix of resampled data of the data matrix, *i.e.* correlations based on random data. For all significant relationships, a .sif file can be created, where on each line, there are the two variables, followed by the correlation coefficent. This file, where two variables on each

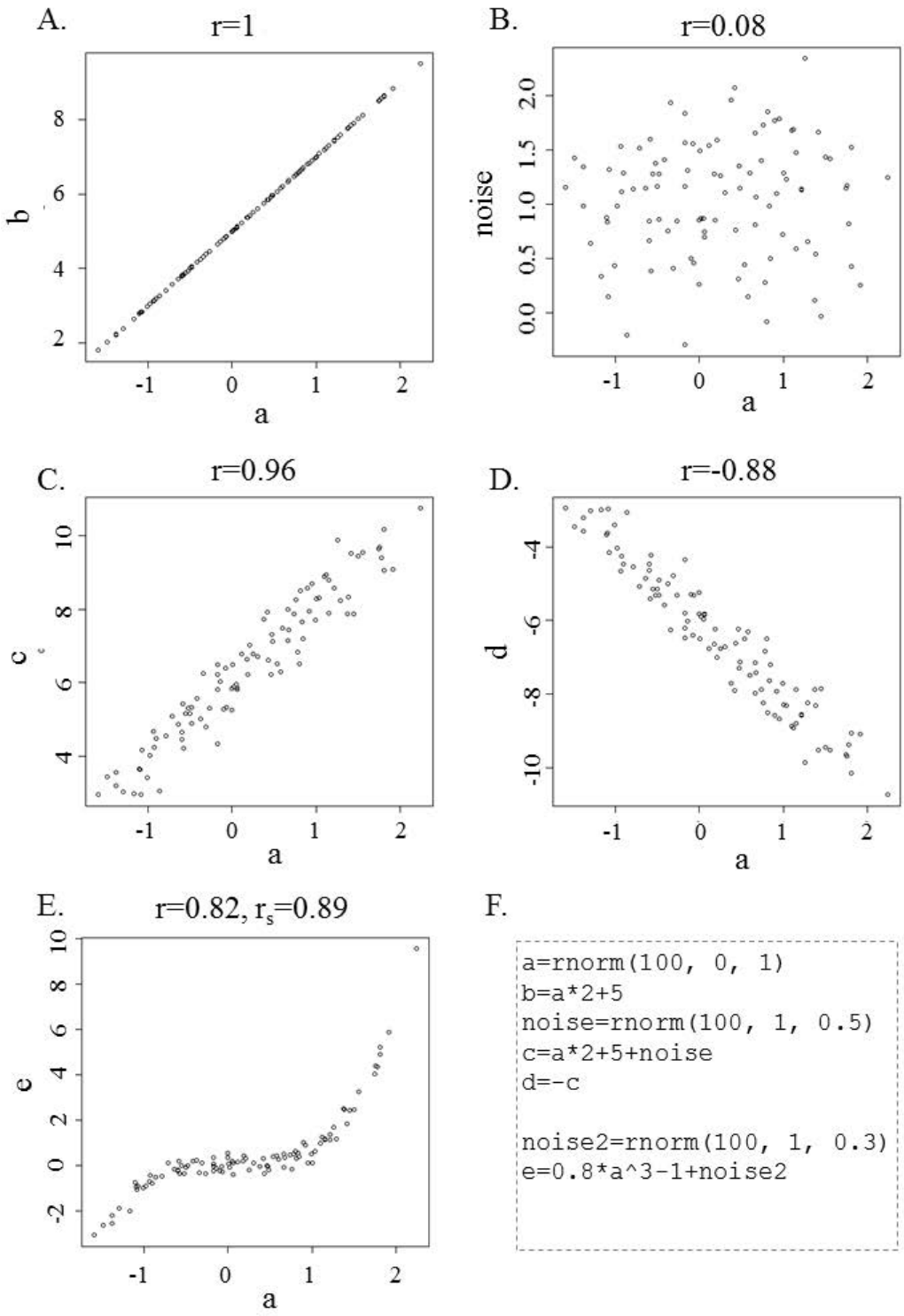


Figure 1.3: Examples of correlation. Panel A shows Pearson Correlation where $a = b$ (correlation 1), panel B shown two non-correlated variables, panel C shown positive and panel D negative Pearson correlation. Panel E shows two variables a and e with nonlinear relationship, where Pearson correlation $r = 0.82$ and Spearman correlation $r_s = 0.89$. Panel F shows what R commands were used to create the variables.

line are nodes, connected by an edge, can be read into Cytoscape [285]. Edge strength is the correlation coefficient. Of course, in addition or instead of the correlation coefficient, FDR can also be stored as an edge attribute. This collection of nodes, connected by statistically significant edges will form a network, which can be used as starting point of further analysis. Overview of steps involved in network creation are shown in Figure 1.4, with an example of making correlation-based network from gene expression data.

The main advantage of correlation-based methods is their simplicity and speed and for this reason, they have been used to calculate correlation-based networks using large public gene expression databases, such as ArrayExpress [170] of GEO [88].

A more advanced method, designed specifically for the inference of biological networks, having the advantage of capturing also non-linear relationships, calculates mutual information instead of correlation. One of the earliest tools for creating such mutual information networks from gene expression data was ARACNE [212]. Like in the case of the correlation-based network, for each pair of variables, a value is calculated, and in this, case, this value is mutual information. The network is again thresholded by a p-value given by the user, and resulting network can be visualized with Cytoscape [285].

In addition to ARACNE, other tools based on mutual information exist, such as CLR [95] as implemented in Minet [219] package or R [260].

1.4.5 Dynamical networks

Static networks are easy to make and can be useful for interpreting large datasets or integrating many types of data. In a static network, edges do not have a direction. In dynamical networks they do and in addition to direction, they interaction can also be described as positive or negative (Figure 1.2 D).

A very intuitive dynamical model to describe biological systems is a Boolean network, where nodes have values of 1 and 0 and each node can be described as a Boolean function of its in-nodes. For example, for a gene A, there might be a repressor, B and activator, C, in which case node B has negative edge towards node A and node C has a positive

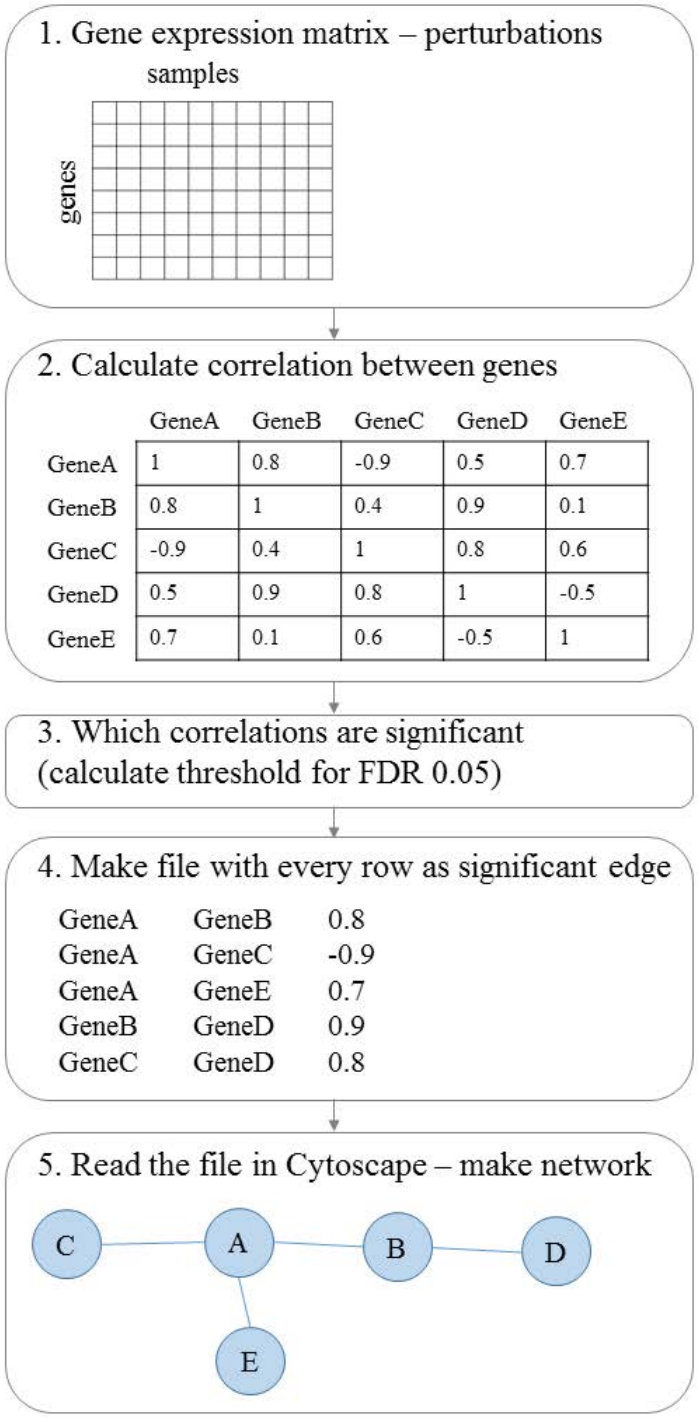


Figure 1.4: Steps involved in network creation, shown with a simple example of correlation-based network. In step 1, gene expression matrix is made, where rows are genes and columns different samples. Then in step 2, correlations are calculated between all genes. In step 3, it is determined, which correlation threshold is statistically significant. In step 4, a file is made where each line contains two genes with significant correlation. In step 5, the file from previous step is opened in Cytoscape to be further analysed

edge towards node A. The overall function can be described by Boolean logic. One of earlier examples Boolean network models was developed for ABA signalling in *Arabidopsis thaliana* and enables predicting network component most important for stomatal closure [195]. Although simple and intuitive, this kind of networks require experimental knowledge of the molecular interactions and are therefore more useful for prioritising nodes based on their effect on certain outcome, than for initial data-driven analysis of omics data for non-model species.

For non-model species with no molecular interactions known, a purely data-based method would be more appropriate. As with static networks, the simplest way to describe statistical associations, but with direction, would be to use a correlation-based method, for example time-delay Spearman correlation [279], where two temporal profiles are shifted step-by-step and correlation calculated, noting the largest correlation and the corresponding time-delay. A network can then be constructed connecting nodes (genes or gene clusters) with largest correlations, with edge length proportional to time-delay.

Time-delay ARACNE [362] as implemented in the TDARACNE R package uses a similar approach, but instead of correlation, mutual information is calculated, allowing to capture also non-linear interactions. The procedure as implemented in the TDARACNE R package [362] computes the mutual information between every combination of variables given a time delay. More precisely, for a given combination of features X and Y the algorithm computes the mutual information between datapoints of feature $X_{1..(\text{P-delta})}$ with feature $Y_{(1+\text{delta})..(\text{P})}$, where P is the number of time-points and delta is the current delta value. This highest mutual information score across the different deltas is chosen and the delta value noted. Positive deltas then denote that feature Y is affecting feature X. Indirect connections are eliminated using the inequality principal (DPI).

Correlation and mutual information-based methods infer statistical and information-theoretical associations in the network, but are not able to describe the network as a model which can be simulated.

Ordinary differential equations create deterministic models, where each gene is mod-

elled as sum of all of the other genes and the resulting network is a signed, directed graph. Ordinary differential equations can deal with cycles and use as input both stedy-state and time-course data. Their main limitation for network inference is that the larger the network, the more parameters need to be estimated and for this reason, they can best work with relatively small network sizes, as opposed to tens of thousands as is feasible with easier methods for undirected networks. A more complex method ARTIVA is based on Dynamic Bayesian Networks, which create stochastic models, and allows to learn the time-varying structure of a network [186].

Time-delay Spearman, time-delay ARACNE and ARTIVA create directed networks from one type of data (time-course omics data), but often it is desirable to integrate multiple sources of data. For example, in addition to time-course data, knock-out data can be used, or known interactions from other databases. A framework allowing this, called NIMOO, is based on multi-objective-optimisation and models the system as a set of ordinary differential equations [130]. In the NIMOO framework, gene expression of gene i depends on the expression of all other genes plus the external perturbation:

$$\dot{x}_i = \sum_{j=1}^{N} w_{ij} x_j + b_i x_i$$

where w is the unknown parameter matrix, x_i and x_j are the expression of genes i and j. b_j is the external perturbation. $\dot{x}_i$ represents the first derivative of x_i.

The first objective is to find parameters matrix w so that the square error between measured and modelled expression is minimised:

$$E^{SQE} = \sum_{i=1}^{N} \sum_{t} (x_i^{measured} - x_i)^2$$

The second objective is to minimise E_{Object}:

$$E^{Object} = \sum_{j} \sum_{t} (o_{ij} - w_{ij})^2$$

where o_{ij} is the second objective (for example, time-delay Spearman correlation or knock-out data)

The result is a parameter matrix w where element w_{ij} shows the effect of gene j on gene i.

Dynamical networks allow inferring the network structure taking into account time, but as these methods are complex, especially the more advanced ones using complex modelling or optimisation, they can only be used with a limited number of variables. As one option, they can be used for a small gene set selected after other methods have been applied. For example, genes to model can be known to be important for a particular phenotype from previous experiments, they can also be small subnetworks or hubs from large static networks. As an alternative, the modelled variables can be representatives of clusters of similar genes or metabolites, cluster medians or principal components of particular pathways.

1.4.6 Methods for non-model species in this thesis

In this thesis, I used TDARACNE for the third chapter (mussel). The main reason was that I only had one type of data available and this was a published method at the time, based on the ARACNE algorithm. Alternative to be used was Time-Delay Spearman correlation. As these was no benchmark to be compared against, I chose TDARACNE because of its ability to infer non-linear interactions.

For the third chapter, I used NIMOO, a newly-developed method in Francesco Falciani's group [130]. The reason for this is the knowledge, that by combining different methods and datasets, network inference accuracy increases. Although I did not have multiple types of data, I included time-delay Spearman correlation in addition to the

time-course data.

At the time of doing analysis, there was hope that different methods submitted to DREAM challenges would be included in an easily usable too to run all of them at once on the same dataset. However, to this date, such tool does not exist and I decided to use available already implemented and easy-to-use tools, keeping in mind that although existing methods might not be mechanistically realistic, they can still give useful knowledge [301].

The reason for choosing ARACNE for static network inference in the 4th chapter was that it was an already established method, preferable over simpler correlation because of the ability to infer non-linear relationships. This method was chosen over an alternative mutual-information method, CLR, because the original publication of CLR was based on prokaryotic data.

For the dynamical model in the 4th chapter, Bayesian method as implemented and published by collaborators [53], was used because they had already demonstrated the use of this method on a lab-simulation of wastewater purification, *i.e.* by using consecutive stages of water tanks where chemical concentrations are known and decreasing and gene expression is measured in Daphnia living in these tanks.

1.4.7 Analysis and functional annotation of networks

When a network is created, it needs to be analysed. Cytoscape [285] provides several tools for network analysis and many graph metrics can be calculated. Examples of two most common measures are shown on Figure 1.5. For example, degree describes the number of connections a node has and hubs are nodes with high degree. Other metrics, like betweenness centrality, on the other hand is calculated by finding shortest paths between every node pair in the network and finding the counting number of shortest paths through every node, divided by the number of nodes. This measure is also high for hubs. There are also bottlenecks, which are nodes connecting 2 parts of the network through which the connections are possible and these nodes have low degree and high betweenness centrality.

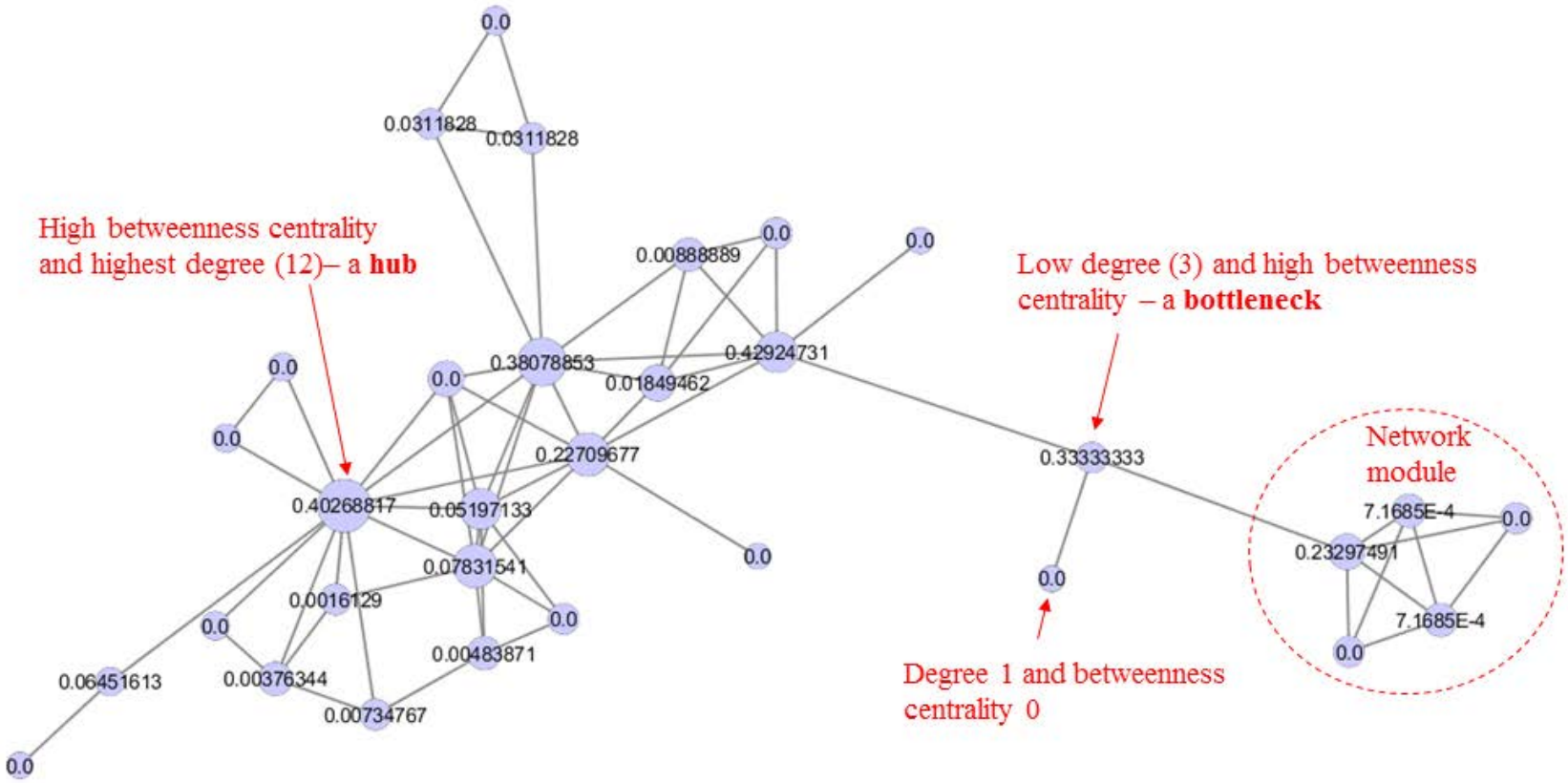


Figure 1.5: Most common network measures. Node size is proportional to the node degree and numbers on nodes indicate the node betweenness centrality measure. Examples of a hub and a bottleneck are indicated with red arrows and an example of a network module is shown inside a dashed circle.

As networks can be quite large, it is useful to functionally annotate them. One way for doing this is to divide the network into parts that have more connections among the nodes of each part than with nodes in other parts. This is called modularisation and can be done by several algorithms. One of the earliest examples is Markov Cluster Algorithm (MCL) [92], which simulates flow through the network with the paths with more frequent flow getting stronger. Another algorithm often used is community clustering such as GLay [304] as implemented in the clusterMaker package [224] in Cytoscape.

Once the modules are found, they can be annotated. For model organisms, there are packages in Cytoscape [285] enabling enrichment for Gene Ontology [62] terms or KEGG pathways [235]. However, these resources have not been implemented for all non-model species. At the moment, the way to perform enrichment analysis for non-model species would be to have a custom-made gene-term list, which can for example be made in BLAST2GO [59, 127] and for each term, find 4 values: count of this term in the module to be annotated, count of this term in the whole annotation list, size of the module and size of the annotation list. Based on these 4 values, a modified Fisher test, known as EASE score [147] can be calculated, which should then be corrected for multiple testing. This custom workflow takes into account the annotated transcriptome size of the specific organism. The custom annotation allows calculating enrichment for different data types. For example, in addition to commonly used Gene Ontology [62] and KEGG [235], data can be downloaded for example from the CTD database [72, 71] and enrichment performed for chemical targets. An easier alternative would be to map transcripts to a model species and use a tool implemented for model species. However, as normally annotation is not complete for non-model species, the full transcript list should also be uploaded as background, otherwise the significance values do not reflect the true proportions of different functions in the set of annotated genes.

A proprietary option for data interpretation is Ingenuity [172, 259]. Ingenity Pathway Analysis Tool uses both external (KEGG and Gene Ontology) and internal databases (Ingenuity Knowledge base, containing curated information from the literature; now In-

genuity belongs to Qiagen and the Knowledge Base is called Qiagen Knowledge Base [259]) databases to calculate enrichment for given gene lists. For every gene list, for every KEGG, GO or Canonical Pathway term, a p-value is calculated based on the number of associations for the specific term in the list and in the whole database and number of genes in the list and in the database.

1.4.8 Machine learning methods used in systems biology

In addition to all exploratory analysis and network methods, machine learning approaches can also be used for analysing omics data for both model species and non-model species. Machine learning methods can be classified into unsupervised and supervised: for the unsupervised, the there is no known structure about the data. One of the most-used scenarios of unsupervised learning is clustering, where data are divided into similar groups based on some characteristics. Clustering was described in the section of exploratory analysis.

In the case of supervised learning, the structure of the data is known. For example, these could be cases and controls of a disease, or treatment. The aim of supervised learning is to build a model that can discriminate between these 2 classes. Generally, the model is built with one part of the dataset (training data), using cross-validation, and then tested on the test set which has not been used for training.

Two of the most widely-used algorithms for supervised machine learning are Support Vector Machines (SVM) [27] and Random Forests (RF) [39]. SVM works by separating two classes by a hyperplane so that the distance from both classes is as large as possible. An illustration of SVM for 2-dimensional data is shown on Figure 1.6. For real biological data, the number of dimensions is the number of genes or other biological measurements and the separator between two classes does not have to be linear.

Random Forests are an ensemble learning method based on decision trees. The training set data is sampled repeatedly with replacement for a set of features, which are then each used to build a decision tree, which are then used on the test data for the prediction of

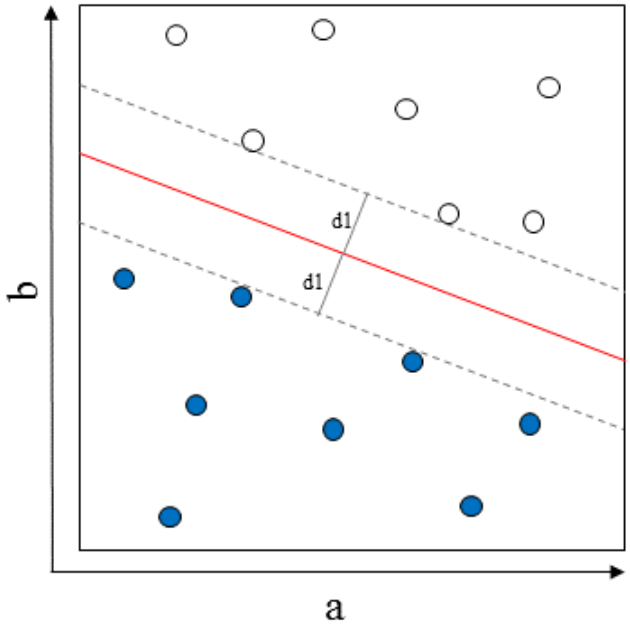


Figure 1.6: Example of Support Vector Machine in 2-dimensional space. Two classes are represented by blue and clear circles. The aim of the SVM here is to find a line (here red) that separates two classes so that the distance (d1 and d2) from the separating line to the nearest member of both classes is as large as possible

outcome, by taking the majority vote.

Although both Support Vector Machine and Random Forest are widely used in biology, SVM is generally faster. In some studies, Random Forest has shown better performance [156], but in others, SVM has outperformed Random Forests [296, 2]. In one comparison between machine learning methods for gene expression data, although SVM performed the best on the dataset of all genes, the results showed that performance depends on feature selection [248]. SVM has also been used for metabolomics data, performing better over Partial Least Squares Discriminant Analysis, PLS-DA, which is another method [208]. In this thesis, I used SVM: because the desired accuracy was achieved, this was sufficient not to try alternatives.

Genetic algorithms can also be used for finding markers predictive of different groups. Genetic algoritm is a heuristic solution to a NP-complete problem of finding a small set of variables from a large variable space that are best able to predict different classes. Genetic algorithm is inspired by biology and works by creating many sets of variables, which are

called chromosomes and are each evaluated for their predictive ability, using any machine learning algoritm, including SVM and Random Forest. The chromosomes with better predictive ability give "progeny" and also participate in the process of "cross-over".

A genetic algorithm initially implemented for gene expression data is GALGO [317] and has been used for biological data and also in other fields. The main advantage of genetic algorithm over using SVM or Random Forest on their own is that in addition to classification, the best-predicting features can be ranked and the optimal set of predictive features selected.

1.4.9 Integration of data from additional resources

In addition to data generated for a specific study, the existing data that has been previously generated can also be used. For this to be successful, there should be standards for data generation, processing and deposition. For example, Minimum Information About a Microarray Experiment (MIAME) criteria apply to microarray datasets [38] and there is also the Metabolomics Standards Initiative (MSI) [97].

Large datasets are generated and public repositories, such as ArrayExpress [170] and Gene Expression Omnibus [88] for gene expression studies and MetaboLights [134] and Metabolomics Workbench [306] for metabolomics data allow the re-analysis and integration of data from previous studies. For the integration of human or mammalian data, there are tools which allow the integration of various resources, such as Genemania [226], STRING [335] and Cytoscape [285]. In addition to freely accessible open-source tools, there are also proprietary manually-curated databases which make the interpretation of omics studies relatively easy [172]. For the interpretation of gene of chemical lists, tools also exist, such as the CTD database [72, 71]. Cytoscape [285] also allows the integration of chemical data with gene or protein networks, for example via CyTargetLinker app [177]. The AOP-DB [250] also integrates data from several external databases, making it possible to connect AOP-s with chemicals from the CTD database [71], provide links with the KEGG pathways [235] and STRING database of protein-protein interactions [309].

1.5 Can systems ecotoxicology help in risk assessment?

Risk assessment is a process of assessing various risks associated with each chemical. In the European Union, this is regulated by REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) since 2007 [85] and as described before, consists of assessment of hazard for human health, for physicochemical hazard, for environmental hazard and also for the assessment of whether the chemical is persistent, bioaccumulative or toxic [93].

Adverse Outcome Pathway is a concept that connects the effects of a chemical to Molecular Initiating Event and through a series of Key Events, to an Adverse Outcome [6, 331]. The Adverse Outcome Pathway (AOP) framework has been proposed to have potential for the risk assessment of both individuals and populations [173]. Moreover, the AOP framework might also be able to draw knowledge from multiple taxa where a mechanism of toxicity might be conserved [207]. For the advancement of the Adverse Outcome Pathway framework, there must be experimental data and most importantly, insightful interpretation and analysis of such data. Systems biology has been proposed to be "leading the revolution" in understanding and interpreting the effects of various chemicals [106]. Moreover, systems toxicology consisting of the integration of omics measurements, chemical concentrations and responses at cellular, organ and organism level and resulting in computational models which are then used for adverse outcome prediction has been proposed lead to a new paradigm in risk assessment [302].

I have shown the need for risk assessment and reviewed the methods commonly used for generating and analysing omics data. The main question that needs to be addressed is whether systems biology approaches can really help in identifying putative AOPs and markers that are useful in risk assessment.

In a review of omics and systems biology for the Adverse outcome Pathways [40], the authors propose that the AOP Framework [6, 332] could potentially help in integrating omics data into risk assessment, as it allows connecting chemical perturbations leading

to Molecular Initiating Events, and through Key events, to the Adverse outcomes. For environmental risk assessment, there are already several published studies about effects of chemicals which have the potential to help in risk assessment [7, 67, 179] and it has been suggested that the AOP framework might be the platform via which the results of various omics studies can be incorporated in risk assessment [40]. It has also been suggested by others, that omics datasets used in ecotoxicological research can results in mechanistic understanding of chemical effects which can be incorporated in the AOP framework [191], which in turn might be used in chemical risk assessment. For example, some AOP-s have already been used in official test guidelines, such as OECD/EU test guideline for human health, specifically OECD442E for addressing skin sensitisation [234]. An example of omics data analysed with a systems biology approach has led to a putative AOP has been demonstrated in *Daphnia magna* [7].

To develop the best possible understanding of how chemical perturbation leads to Molecular Initiating Event and through key events to an Adverse Outcome, the results of systems biology analysis should be validated experimentally. To use this information for risk assessment, a complicated computational framework addressing several challenges must be developed. These include species extrapolation [181], and also chemical-specific properties as described by absorption, distribution, metabolism and excretion (ADME) [49]. This is a huge challenge which are also being investigated in the area of drug development and even in this area, there is currently no integrated framework that is able to integrate many types of available information. However, for predicting molecular effects, especially for drug repurposing, the extrapolation across species for safety concerns and the ADME properties of chemicals all have to be taken into account. For this, systems biology has been proposed as a suitable tool [90, 257, 244]. Additionally, the importance of the AOP framework has been acknowledged by the drug development community [184].

In conclusion, there is a potential for systems biology to be used for the development of AOPs and for risk assessment, by providing hypotheses that can be tested further, and certainly the utility of these approaches is advancing, as shown by recent example of the

development of an alternative assay for predicting chemical effects [300]. Moreover, as the tools for data integration are developing, such as the AOP-DB [250], omics datasets and systems biology analysis can contribute to the development AOPs and provided the regulatory acceptance, also used for environmental risk assessment.

1.6 Aims and objectives

The aim of this thesis is learn about the effects of environmental stress in non-model species, using systems biology. The aim connecting all chapters is to use data-driven analysis and biological networks created from data to connect omics measurements with other measurements, such as physiological and environmental measurements. In all chapters, the ultimate aim is to interpret results biologically.

Specifically, in the second chapter, objectives are:

1. To learn whether two neurotoxic chemicals RDX and carbaryl affect gene expression in earthworm *Eisenia fetida*.

2. To learn whether genes affected by two neurotoxic chemicals RDX and carbaryl overlap.

3. To learn whether differentially expressed genes are connected to the conduction velocity of the medial giant nerve in a directed network?

4. To interpret the results biologically.

In the 3rd chapter, objectives are:

1. To learn whether metabolites in blue mussel *Mytilus edulis* are associated with annual cycle, sex and site.

2. To learn how metabolite levels change during the annual cycle.

3. To learn whether metabolites can be associated with environmental and physiological parameters in a directed biological network.

4. To interpret the results biologically.

In the fourth chapter, objectives are:

1. To study whether gene expression changes during water remediation in three-spined stickleback *Gasterosteus aculeatus* living in various stages of remediated water.

2. To compare differential expression of genes between sites.

3. Perform exploratory analysis of chemical concentrations in different sites.

4. To explore whether gene expression and chemical concentrations can be analysed using a static biological network.

5. To interpret results of a collaborator about Bayesian model using the same data that I used in this chapter.

CHAPTER 2

A DATA-DRIVEN APPROACH TO UNDERSTAND THE TRANSCRIPTIONAL RESPONSE TO CHEMICAL EXPOSURE AND ITS EFFECT ON NERVE CONDUCTION VELOCITY IN THE EARTHWORM *EISENIA FETIDA*

2.1 Contributions

- Jaanika Kronberg-Guzman[1,2,5] performed systems biology analysis (differentially expressed genes, clustering, heatmaps, PCA, network pruning, analysis and interpretation, wrote this chapter (details of the experimental methods were written based on information provided by Edward Perkins' lab)
- Ping Gong[3] did experimental work (exposure, microarrays and medial giant nerve fibre measurements)
- Rita Gupta[1] ran the ODE framework with her method [130] that was unpublished at the time of work and not annotated to be easily used by others. She provided the results in the form of matrix which Jaanika Kronberg-Guzman analysed
- Tim Williams[1] made improved annotation with BLAST2GO and found human orthologs to be used for Ingenuity annotation

- Natalia Garcia-Reyero[4] oversaw the experimental work, participated in the discussions of computational work
- Edward Perkins[3] initiated and organised this study, oversaw the experimental work, participated in the discussions of the computational work
- Francesco Falciani[1,5] oversaw the computational work, provided scientific ideas, participated in the computational discussions

Affiliations

1. School of Biosciences, University of Birmingham, UK
2. Institute of Genomics, University of Tartu, Estonia
3. U.S. Army Corps of Engineers, USA
4. Institute for Genomics, Mississippi State University, USA
5. Institute of Integrative Biology, University of Liverpool

2.2 Introduction

Anthropogenic chemicals released into the environment can enter both terrestrial and aquatic ecosystems. Within the terrestrial environment, chemical pollutants can affect the health of organisms living in soil either by direct exposure [57] and/or through the ingestion of contaminated food [232, 199]. Since soil health is important for the health of the whole ecosystem (reviewed in [82, 81]) and can also impact human health (reviewed in [297]), understanding mechanisms of toxicity in earth organisms is of paramount importance in ecotoxicology. Earthworm species have been called "soil engineers" because of their burrowing activity which is important for maintenance of soil structure [54]. Due to the widespread presence of earthworms in soils, they can be used for soil monitoring, for example by measuring species number or biomass (reviewed in [239, 243]). Together with other causes, chemicals in soil can affect the health of earthworms and this can have a profound impact on biodiversity and biomass. Therefore, it is important to understand how different chemicals affect earthworms. So far, this has been studied in various earthworm species, such as *Lumbricus rubellus*, *Lumbricus terrestris* and *Eisenia fetida* [54, 26], reviewed in [249, 325].

Survival, reproduction and growth parameters [295, 230, 328, 267], avoidance behaviour and burrowing activity tests [322, 107, 183], have been used to assess the effects of various chemicals. Wound healing rates indicating the ratio of exposed worms healed within a given time compared to control worms [56] have also been shown to indicate chemical exposure. All these above-mentioned markers indicate changes at the level of organism or tissue. At the cellular level, neutral red uptake has been shown as indicator of chemical exposure [284, 11]. At the molecular level, various biomarkers have been shown to indicate the presence of chemicals [51, 303, 206, 11].

The choice of a specific test or biomarker can set limits on the types of effects seen. For example, if reproductive tests are used, only effects on the reproductive system can be observed. These tests have been used extensively for assessing specific hypotheses. In the last 15 years, the development of high-throughput functional genomics methods

has provided new ways of investigating molecular networks linked to toxicity mechanisms, independently of any specific hypothesis. More specifically, omics methodologies can bring to a better understanding of environmental stress and have been used for many species of environmental relevance [7, 276, 277, 80, 106]. Earthworms have also been used for transcriptomics studies [121, 122, 47, 238, 308] and these species has been described as suitable for toxicogenomics [124].

In *Eisenia fetida*, gene expression changes and the conduction velocity of the medial giant nerve fibre (MGF) have been studied in the context of exposure to a neurotoxic chemical [122]. In this chapter, I present a systems-level analysis of earthworm *Eisenia fetida* exposed to two neurotoxic chemicals in the laboratory. The main aim was to develop a strategy to analyse high-dimensional time course data and relate it to a physiological measurement. The work in this chapter with lab-exposed *Eisenia fetida* exposed to single chemicals establishes ground for using similar approaches with natural populations of non-model species to understanding the effect of multiple environmental stressors.

More specifically, *Eisenia fetida* was exposed to two neurotoxic chemicals, RDX (1,3,5-Trinitro-1,3,5-triazinane) and carbaryl (1-naphthyl methylcarbamate) in the laboratory. In addition to gene expression, the conduction velocity of the medial giant nerve fibre (MGF) was measured as a proxy for possible nerve damage. RDX and carbaryl are especially important to understand as RDX is a neurotoxic explosive that enters the environment through military training events and carbaryl is an insecticide. Both chemicals are toxic to humans and affect the nervous system RDX has been shown to cause seizures in rats and humans [117, 347].

Carbaryl is a cholinesterase inhibitor [37, 189], a chemical that has the potential to inhibit the breakdown of acetylcholine by acetylcholinesterase. The activity of acetylcholinesterase has been shown to be indicative of chemical poisoning or exposure in various different species, for example humans [13], lizards [271] and honeybees [16]. By the use of another acetylcholinesterase inhibitor (chlorpyrifos; however, this is irreversible and carbaryl is reversible), the effects of the acetylcholinesterase inhibition have been

demonstrated in *Eisenia fetida*, starting with constriction and swelling in the initial 12 hours of exposure and progressing to sluggishness and unresponsiveness to stimuli [261]. In *Eisenia andrei* (another earthworm species), the acetylcholinesterase activity has been shown to be affected by carbaryl: maximum inhibitory effects can be observed after 1 day of exposure on filter paper and after 3 days in soil [103]. Earthworm *Eisenia fetida* has been shown to tolerate carbaryl better as compared to other earthworms: instead of lethality, high doses affected its burrowing activity [298]. Carbaryl also affects growth and reproduction of *Eisenia fetida* [230].

RDX (also called cyclonite) is an explosive that has been shown to cause seizures in rats and humans [117, 347]. It has been shown in rats that RDX binds to the GABA receptor A, reducing synaptic transmission mediated by $GABA_A$ receptor [347]. The mode of action of RDX has been demonstrated to be conserved between several species, including earthworm *Eisenia fetida* [105]. In earthworms, RDX has been shown to reduce fecundity [266] and cause neurophysiological symptoms (rigidity and shrinking) [123].

In this thesis, an ordinary differential equation (ODE) modelling was used to model the relationship between molecular and physiological response to the two chemicals with neurotoxic effect. Our approach is data-driven and aims at discovering putative mechanisms of action linked to neurotoxic effects. I show that the dynamic gene expression profile resulting from exposure to one of the chemicals, RDX, is linked to nerve conduction velocity in the medial giant nerve fibre. The model also suggests that the changes in gene expression might be involved in the recovery from nerve damage.

2.3 Methods

Overview of both experimental and computational analysis is shown in Figure 2.1. Part A describes the lab experiments and part B shows the summary of the computational workflow.

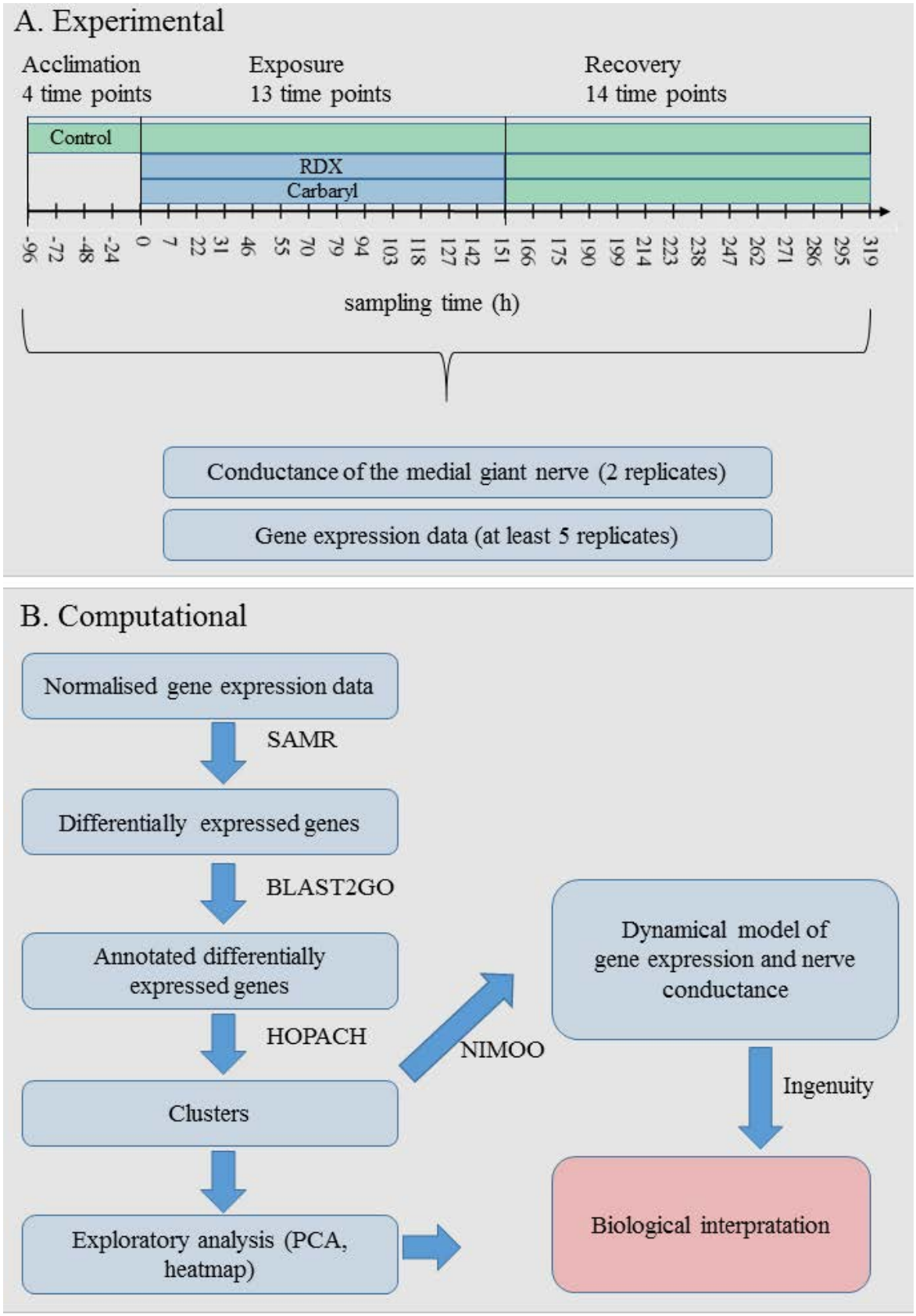


Figure 2.1: Overview of experimental (A) and computational (B) workflow.

2.3.1 Laboratory work (done by Ping Gong in Edward Perkins's lab)

Experimental setup (Edward Perkins' lab, experimental details written based on the thesis of Haoni Li)

A synchronised *Eisenia fetida* culture was started from cocoons and maintained in the labs as described previously [123]. Mature worms of 0.4-0.6 g with clitellum were selected for the experiment. Worms were transferred to individual glass vials (115ml) and acclimated for 4 days prior to exposure. Worms were then exposed to carbaryl (20ng/cm^2), RDX (2μg/cm^2) or acetone (solvent control) on moistened filter paper, as described in [192]. These concentrations were chosen so that they are sub-lethal. The experiment consisted of 3 phases: acclimation (4 days), exposure (6 days) and recovery (7 days). During acclimation, worms were sampled daily. During exposure, worms were sampled 13 times (from 0 to 142 hours, Figure 2.1 A). At 151 hours (the end of the exposure phase), all worms to be sampled during recovery were placed in new glass vials containing non-spiked filter paper. During the recovery phase, worms were sampled 14 times (from 151 to 319 hours, Figure 2.1 A) for all treatments. At each sampling, during acclimation, exposure and recovery, the conduction velocity of the medial giant nerve fibre (MGF) was measured as described before [123]. Worms were sacrificed by snap-freezing in liquid nitrogen, fixed in RNAlater-ICE and stored in -80°C (Figure 2.1 A). RNA was extracted from at least 5 worms per time point per treatment similarly to a previous study from the same lab [247]. One sample was eliminated due to poor RNA quality in the 10th timepoint of RDX exposure. RNA was hybridised to custom-designed Agilent array [125] using Agilent's one-colour Low RNA Input Liner Amplification Kit. The arrays were from 3 manufacturing batches, so from each sample of 5, replicates were distributed on arrays from each of the batches. After scanning, gene expression data acquired with Agilent's Feature Extraction Software Version 9.1.3. Multi-dimensional scaling was used to check for batch effects, showing no significant effects [192].

Pre-processing of microarray data (Edward Perkins' lab)

Pre-processing consisted of feature filtering, conversion of signal intensity into relative RNA concentration, normalisation and gene filtering and were done as described in [192]. After these procedures, 43000 genes remained and were used for subsequent analysis.

2.3.2 Systems biology analysis

Statistical processing (Jaanika Kronberg-Guzman)

Differentially expressed genes were identified using SAMR R packege (FDR < 0.05) [312] between treatment and control, separately for both chemicals. Time-course consisted of gene expression values (5 replicates for each time point, except recovery point 10 for RDX-treated worms where there were 4 replicates, and recovery points 5,6,8,10,11 and 12 where there were 6 replicates for carbaryl-treated worms) for each of the 27 time points in exposure and recovery phases, plus control samples from same time-points. Using two-class time-course (resp.type="Two class unpaired timecourse", time.summary.type="slope"), differentially expressed genes were found that differ between control and treatment. SAMR first computes the standard error of the slope and then uses this slope-summarised data as regular two-class data. The resulting positive d-statistic shows that the in the treatment group, gene expression increases in time compared to the control group, and negative d-statistic shows that in the treatment group, the expression decreases in time compared to the control group.

For further analysis, transcriptional measurements were expressed as ratio between exposure and recovery (in each time-point) over the corresponding control measurements, *i.e.* the mean of logged control values were subtracted from the exposure log values. The significance of the overlap between the two gene sets was determined by the R function phyper ($1 - phyper(q, m, n, k)$) where $q = 4158$ (overlap), $m = 12061$ (number of differentially expressed genes in the carbaryl experiment), $n = 63542$ (total genes) and $k = 8852$ (number of differentially expressed genes in the RDX experiment).

Data annotation (Tim Williams)

The *E. fetida* microarray had been designed to target 63,542 sequences, the majority of which had been derived from 454 sequencing of *E. fetida* cDNA [125]. In the current study, 16,905 sequences were identified as differentially expressed, but only 2,385 of these had originally been annotated to known proteins by BLASTx at E value $< 1E-06$. Two further annotation strategies were employed; the 16,905 sequences were directly annotated by Blast2GO [127], and indirectly annotated by first employing BLASTn at $< 1E-06$ versus 630,000 Annelid ESTs (Genbank EST) then annotating the matching ESTs by Blast2GO. Combining significant matches from all approaches resulted in 4,703 sequences identifiable with known proteins. These protein sequences were used to search the human proteome [Ensembl Genes 62; *Homo sapiens* genes (GRCh37.p3)], resulting in 3712 matches at $< 1E-06$. Therefore, the annotation of *E. fetida* transcripts was improved and identification of putative human orthologs allowed association of *E. fetida* genes with the extensive mammalian gene annotation. Of the significantly differentially expressed genes, only those with any annotation were used in further analyses.

Clustering and exploratory analysis (Jaanika Kronberg-Guzman)

Expression profiles of annotated differentially expressed genes for both chemicals were clustered separately with HOPACH [327, 251], a clustering method that automatically identifies the optimal number of clusters. As a distance matrix, Spearman correlation was used. The centroid of each cluster was then interpolated using polynomial interpolation to obtain a high density time series (in this case 143 time points were generated for the worm by using polynomial interpolation with root 8; this root was chosen based on visual inspection), ready for the ODE modelling technology NIMOO.

The interpolated gene cluster medians and MGF were visualised as heatmaps using the heatmap function in R [260]. Principal component analysis was done in R [260], using interpolated gene cluster medians. R function prcomp was used, and first 2 principal components visualised, using R plot function [260].

NIMOO (Jaanika Kronberg-Guzman, Rita Gupta)

Here, a method called NIMOO [130] was used to integrate conventional parameter estimation procedures with a correlation-based method for network inference. Specifically, two matrices were given as input to NIMOO: a time-course data matrix of interpolated clusters and MGF and a time-delay Spearman correlation of interpolated clusters and MGF (a custom script provided by Kim Clarke). The output was a matrix describing the expression level of each cluster median by all other cluster medians. Since every interaction has weight, and for visualisation and interpretation it is easier to analyse only the largest effects (strongest edges in the network), both networks were thresholded for the network visualisation to give similar average connectivity of the network (thresholds were 0.7 for carbaryl and 0.4 for RDX).

Biological interpretation (Jaanika Kronberg-Guzman)

Gene lists of putative human orthologs corresponding to each of the annotated worm transcripts were input into Ingenuity [172]. Reports of enriched Canonical Pathways and KEGG pathways [235] were generated based on the Ingenuity Knowledge base [172].

2.4 Results

2.4.1 Two neurotoxic chemicals, carbaryl and RDX, affect the expression of different genes in *Eisenia fetida*

I first tested whether these neurotoxic chemicals have any effect on the transcriptional state of earthworms. There were 12061 differentially expressed genes after exposure to carbaryl and 8852 after exposure to RDX (FDR 0.05) (Figure 2.2 A) that differ between control and treatment time-courses, *i.e* by comparing slopes of each gene between the two classes. The overlap between the two gene lists was 4158 (significantly different from the overlap of randomly sampled genes from the same set, p-value=0). Previously,

from all differentially expressed genes, only 2385 had been annotated to known proteins by BLASTx at E-value $< 1E-06$. With the new Blast2GO annotation, the number of annotated genes increased to 2885. The number of differentially expressed annotated genes for carbaryl was 2100 and for RDX it was 1301 and the overlap of annotated differentially expressed genes between the two chemicals was 516 (Figure 2.2 B). Although the overlap was significant, there are also many differentially expressed genes which are specific for each of the chemicals, therefore the two sets of genes were analysed separately.

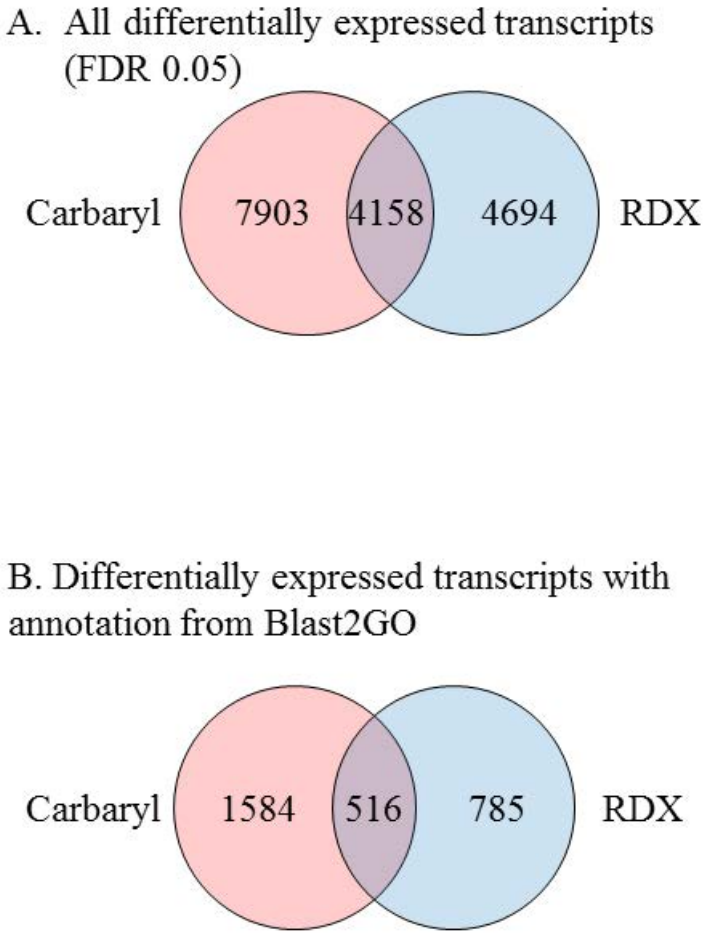


Figure 2.2: A. Overlap of all differentially expressed transcripts between two experiments. B. Overlap of differentially expressed transcripts with annotation

In order to reduce the complexity of the dataset and visualise the dynamics of gene expression, a clustering procedure (HOPACH) [327, 251], which determines the number of

clusters automatically, was used on the gene expression time-course profiles, and resulted in 111 gene clusters for carbaryl (31 gene clusters >10 genes) and 15 gene clusters for RDX (11 gene clusters >10 genes). To obtain an overview of the dynamics of clusters during

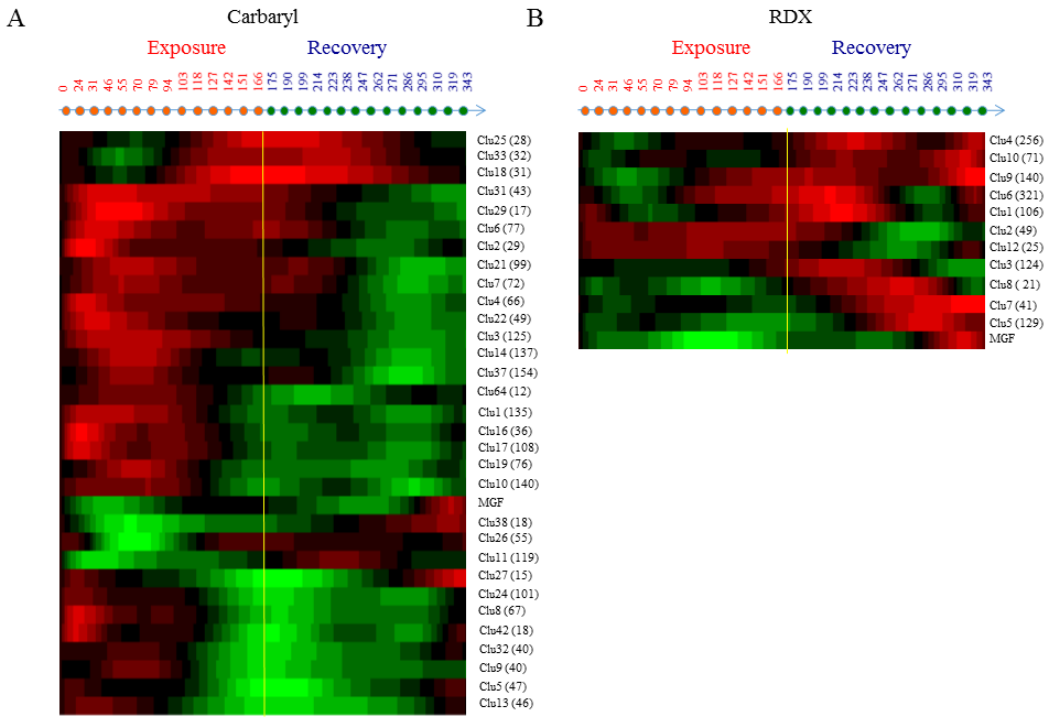


Figure 2.3: Heatmap of interpolated clusters for the exposure to carbaryl (left) and for the exposure to RDX (right). Green shows low expression and red high expression for normalised, standardised cluster medians. Number in brackets indicates number of annotated transcripts in each cluster.

exposure and recovery, cluster medians were visualised as heatmaps (Figure 2.3). Visual inspection identified several clusters whose average profile displayed a marked modulation in response to exposure and recovery. In order to assess the overall dynamics of response to exposure in the two experiments, I performed principal component analysis using the averaged profiles as input data (Figure 2.4). Panel A clearly shows that response to carbaryl exposure follows a trajectory across both components. Interestingly, immediately after the removal of the chemical (recovery phase) the organism responds by a trajectory across the second component. On the contrary, after RDX removal, *E. fetida* gene expression continues on a trajectory established in the exposure phase for 3 days before it

shows any sign of responding to the removal of the chemical (Figure 2.4 B).

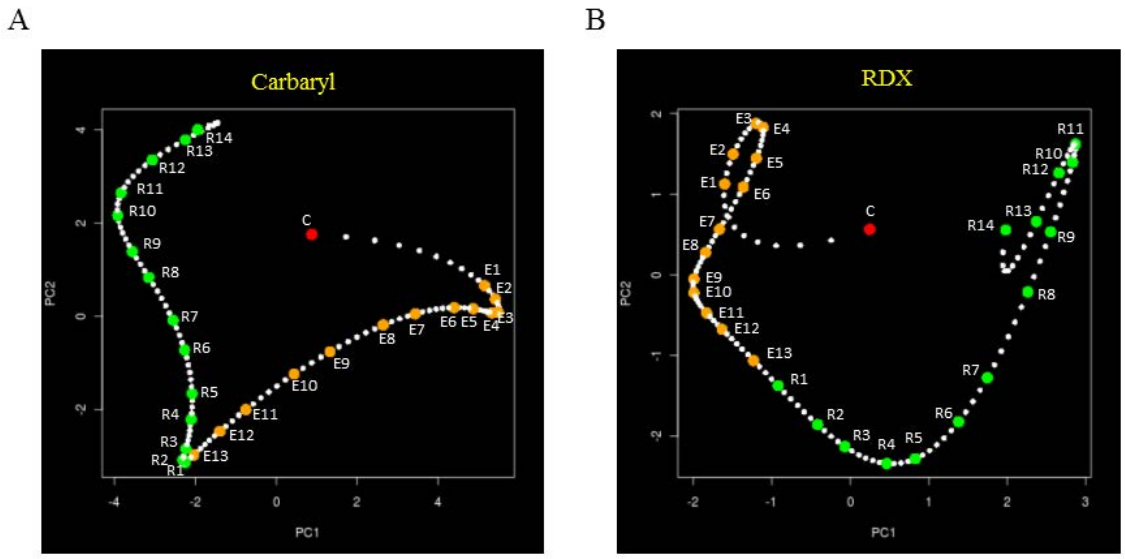


Figure 2.4: PCA of interpolated clusters for the two experiments. The trajectory of the gene expression alters immediately after removal of carbaryl, but a similar alteration takes longer after the removal of RDX. The red dot represents the average of unexposed controls (C-control), orange dots (E1-E13) are exposure and green dots (R1-R14) recovery. White dots represent interpolated timepoints.

2.4.2 Advanced computational modelling connects the gene expression changes with the conduction velocity of the medial giant nerve fibre

As one of the aims was to test whether gene expression trajectories are related to the dynamics of the conduction velocity of the medial giant nerve fibre (MGF, Figure 2.5), I decided to use an approach where each biological entity can be described as a function of all other biological entities to integrate temporal changes in gene expression and MGF. Here the entities are clusters of interpolated gene expression time-course profiles and the interpolated MGF time-course. To accomplish this aim, I used an ODE-based approach as implemented in NIMOO [130] which is a multi-objective optimization framework that has been shown to be able to work well with different datasets and network inference methods. NIMOO has the advantage of being able to integrate multiple types of data. Here I used NIMOO to integrate time-course data and time-delay Spearman correlation of the same

data. An overview of the approach is shown in Figure 2.6. Specifically, NIMOO models the expression of each biological entity as a weighted linear sum of all other entities. During the optimization process, aim can be set for squared error between inferred and measured expression value. Here the aim for squared error was set for $< 10^{-5}$ and both models ran successfully. NIMOO resulted in a matrix of interaction strengths, *i.e.* how strong if the effect of one cluster on other ones. For an intuitive interpretation of the NIMOO model, the matrix can be visualized as a network, where nodes are gene clusters and MGF, edges the strength parameters from the NIMOO matrix. Although models could be developed for both chemicals, the network linking gene expression dynamics with the phenotypic endpoint could be identified only in the case of RDX (Figure 2.7, Figure 2.8).

2.4.3 Further analysis of network model of RDX exposure suggests that the gene expression changes might be the result of nerve damage

As the aim of this work was to discover molecular mechanisms controlling the toxicity mechanisms and network analysis of gene expression linked to the conduction velocity might enable to understand more of the hierarchy of molecular events, the RDX network (Figure 2.7) was analysed further. In the RDX network, there are 3 upstream nodes (clusters 4, 10 and 12) which change before other nodes. Two of these (cluster 4 and cluster 10) are also directly connected with the conduction velocity of the medial giant nerve fibre (MGF), which is the most downstream node of the network. Cluster 12 is also upstream of MGF, but through 3 intermediate nodes. To aid biological interpretation, functional annotation was performed for the clusters in the RDX network (Table 2.1 and Table 2.2). For each of the clusters, representative genes are shown for which sequence homology was confirmed by performing a sequence alignment (an example shown on Figure 2.9).

Interestingly, many of the clusters are enriched in terms related to signalling. Other

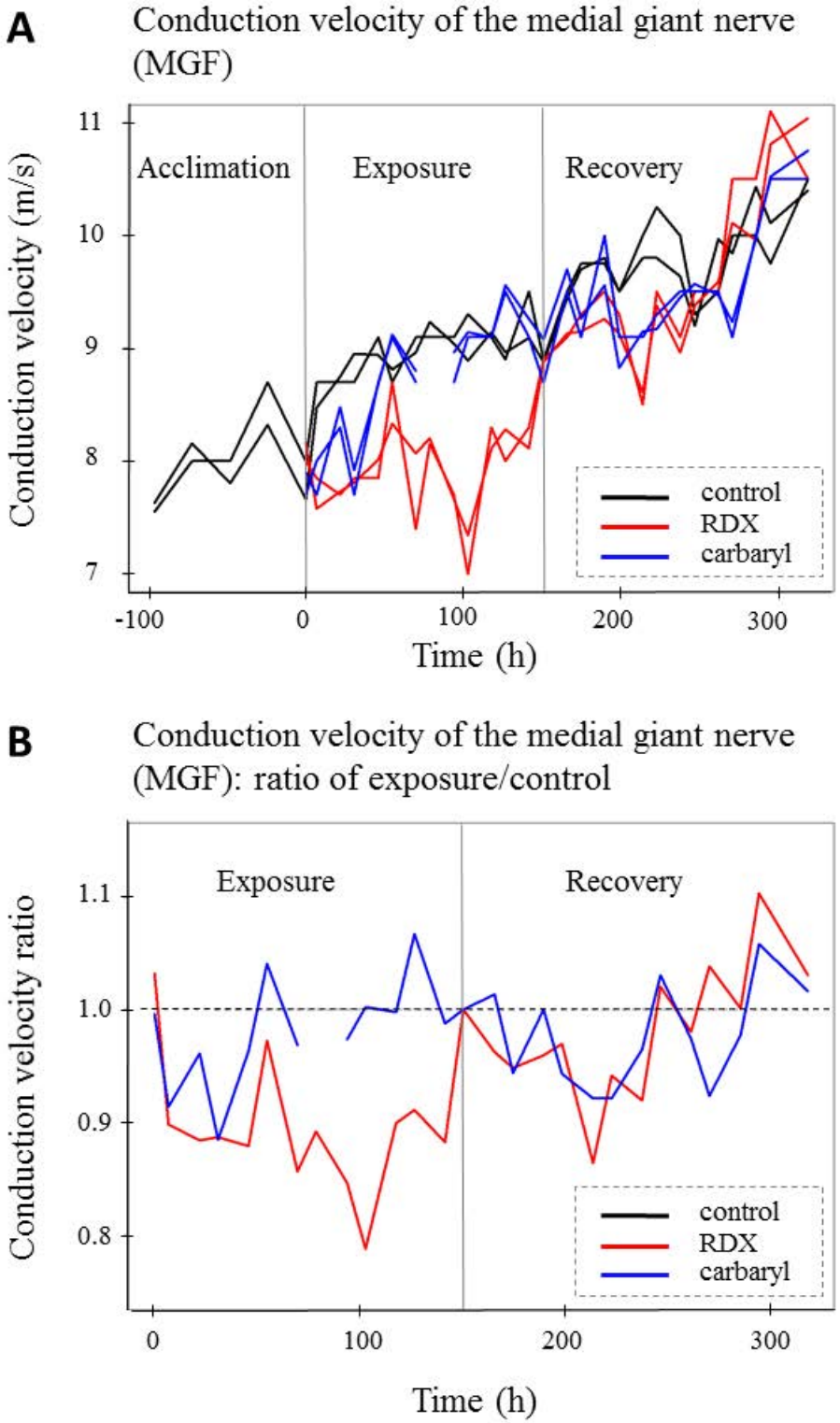


Figure 2.5: Conduction velocity of the medial giant nerve fibre. A. Raw conduction velocity measures of control, exposure and recovery from RDX and exposure and recovery from carbaryl. B. Ratios between exposure and recovery of RDX and carbaryl over control measurement of same time points.

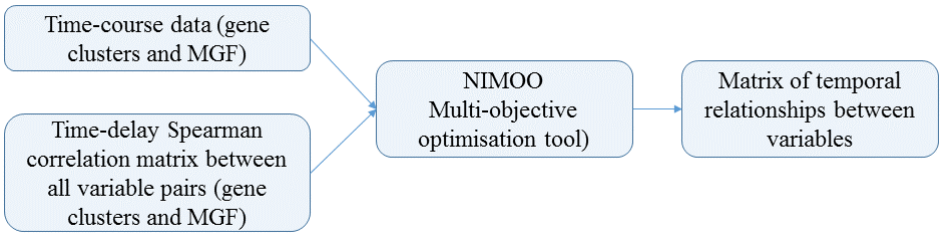


Figure 2.6: Overview of the multi-objective optimisation for the time course data

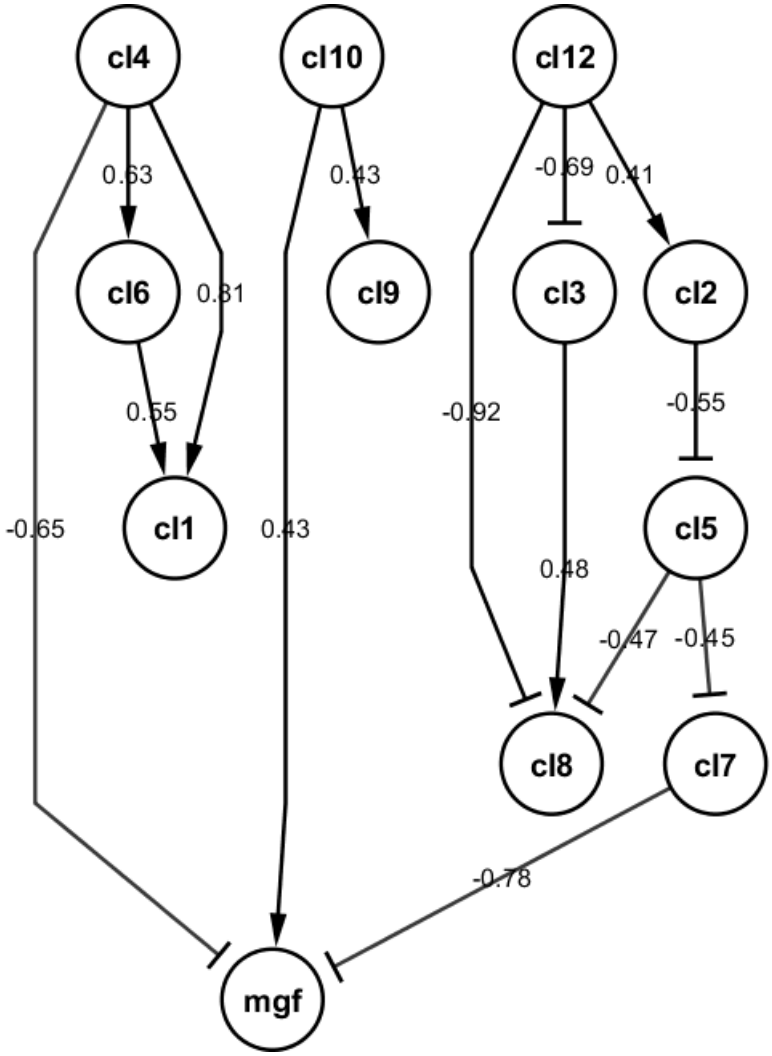


Figure 2.7: NIMOO model of the RDX time-course of both exposure and recovery. The model relates clusters of genes to the conduction velocity of the giant median nerve fibre (MGF). Nodes represent gene clusters and the MGF measurement, edges represent the connection strength (shown numerically) and direction of effect (as arrow). Arrowhead shows whether the predicted interaction is positive or negative.

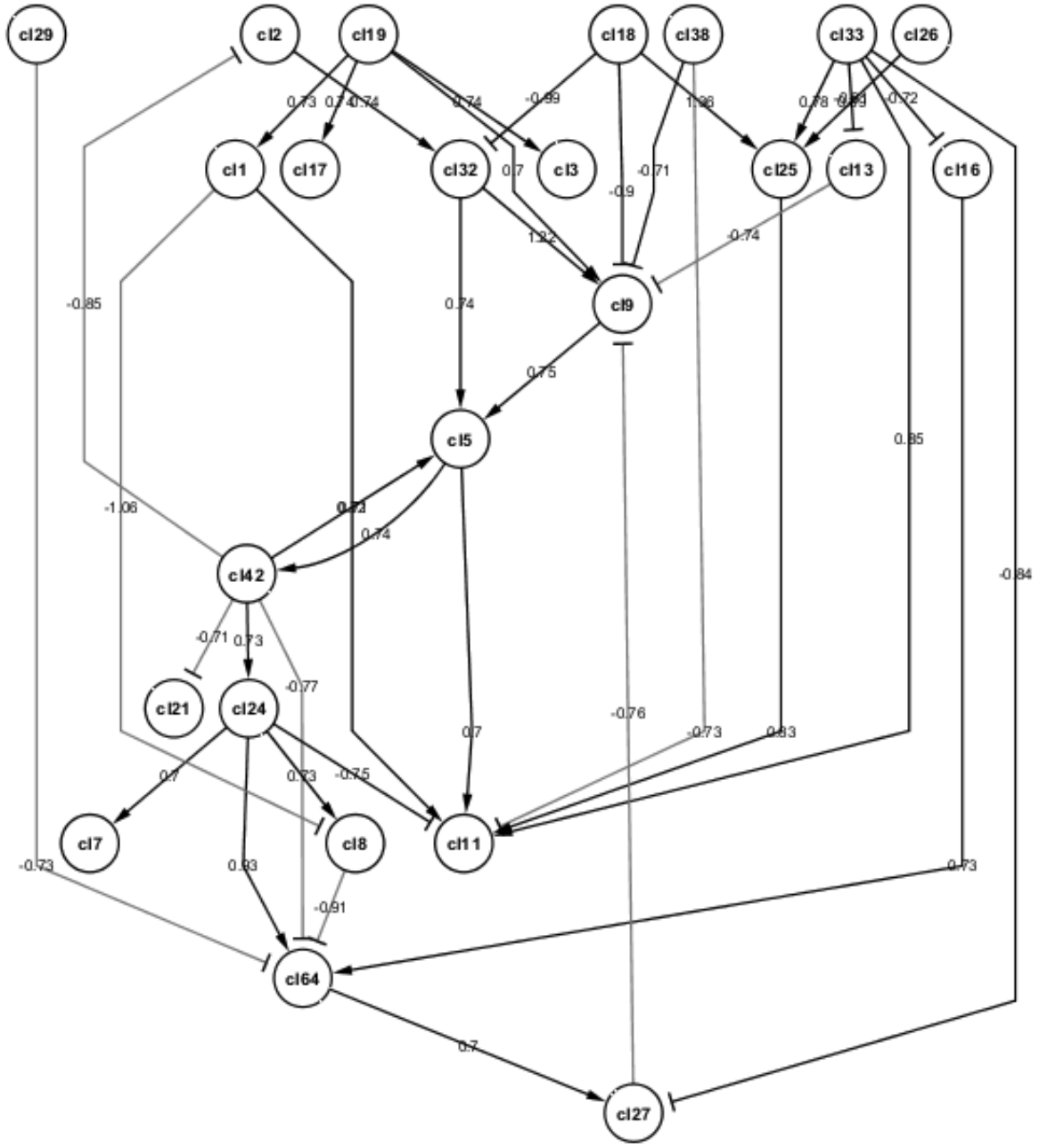


Figure 2.8: NIMOO model of the carbaryl time-course, of both exposure and recovery. Nodes represent gene clusters and the MGF measurement, edges represent the connection strength (shown numerically) and direction of effect (as arrow). Arrowheads show whether the predicted interaction is positive or negative.

functions that stand out are apoptosis and cell death, lipid metabolism, transport, energy, oxidative phosphorylation, chromatin remodelling and gene expression. The model described was attempted to describe the dynamics of both exposure and recovery. However, as MGF changes only during recovery and many clusters also do not change much during

A. Translated protein: Reading frame 4

```
>CHRNA7_4 Eisenia fetida;
HPSIKFVHQFVVGCFSVSISSIAASF*PTSYQPTLTLRMFYAKGRLFALLFVCATACLLY
VQGVSGRGSSAXDLLIQNLFNDYHSSVXPSGFIDLRXGLXIXCXRFDKDKAKLTTDAVES
YTWKDFRLSWNVKDFDGIKKIRVPASLVWKPDIDCYNSLEVEKRSDKNVIIESDGTVKWA
PRGQXKTLCAVEDGKPVCNLSFGSWTYDAWSMALELEGEGVELSNYMNKTCPRSIENFTS
HVKTSTYLCCKEPYNSLEIKLNLKPLPKQEEGVASD*MK*RRSRTNKRKNASVNERD*TN
*CXXXXXAFNQSINQSXKIRXVKKHNK*EILFSIYVTRYI*KCIISRV
```

B. Protein BLAST for reading frame 4

neuronal acetylcholine receptor subunit alpha-7-like precursor [Ciona intestinalis]

Sequence ID: NP_001265890.1 Length: 476 Number of Matches: 1

See 1 more title(s)

Range 1: 6 to 234 GenPept Graphics Next Match Prev

Score	Expect	Method	Identities	Positives	Gaps
107 bits(268)	5e-22	Compositional matrix adjust.	71/232(31%)	108/232(46%)	15/232(6%)

```
Query  45   RLFALLFVCATACLLYVQGVSGRGSSAXDLLIQNLFNDYHSSVXP----SGFIDLRXGLX  100
            LF  + +     +   QGV+G  S A   LIQ+L  +Y   V P     +  I++   +
Sbjct  6    ELFLRIVLPMLVLMTMAQGVNG--SQAEKDLIQDLLRNYDVMVRPIDKYNDIINVSFAVT  63

Query  101  IX-CXRFDKDKAKLTTDAVESYTWKDFRLSWNVKDFDGIKKIRVPASLVWKPDIDCYNSL  159
            +      D+    LTT     +TW D  L WN  ++ GI +IR+PA  VWKPDI  YNS
Sbjct  64   LQQIVDLDEKNQLLTTSMYMGWTWNDTYLKWNPDNYSGIVEIRLPAKKVWKPDILVYNSA  123

Query  160  --EVEKRSDKNVIIESDGTVKWAPRGQXKTLCAVED-----GKPVCNLSFGSWTYDAWSM  212
               ++    NV+I S G V+W P G  KT C V+       +  C + FG+WTY    +
Sbjct  124  VDSFDQMLQTNVVIWSTGAVEWLPPGLFKTTCDVDIRYFPFDEQRCTMKFGAWTYHGGMV  183

Query  213  ALELEGEGVELSNYMNKTCPRSIENFTSHVKTSTYLCCKEPYNSLEIKLNLK  264
             L L  E   L NY+  +    + +   H  +  Y CC  P+  +   ++++
Sbjct  184  DLALPDENAILDNYI-PSGEWDLISMKGHRTSVKYECCPHPFVDVTYTIHMR  234
```

C. Interpro scan for the reading frame 4

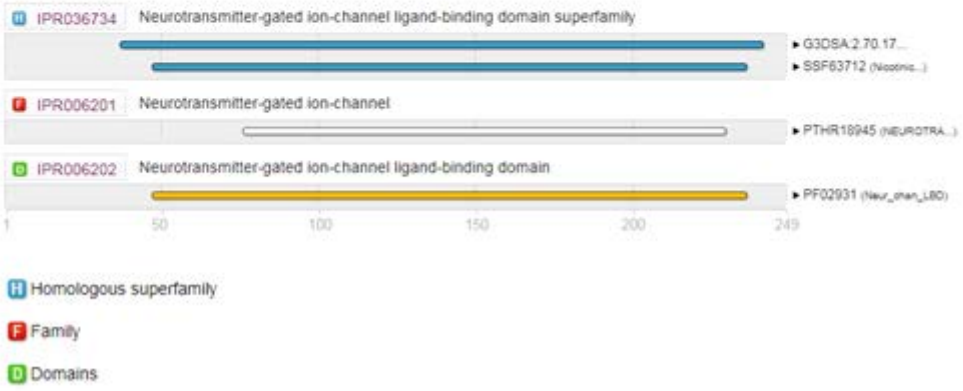


Figure 2.9: An example of sequence alignment of CHRNA7 to confirm homology. A: The *Eisenia fetida* sequence was translated in all 6 reading frames, here reading frame 4 is shown. B: example of a hit from sequence alignment of the translated nucleotide sequence. C: Interpro scan for the translated reading frame 4 shows a neurotransmitter-gated ion-channel domain.

exposure, it might be worth modelling the two parts separately. For these models, only nodes which change during the phase modelled (either exposure of recovery) were used.

For the exposure phase, the model was built with clusters 9, 6, 1 and 4. In the exposure network (Figure 2.11), most clusters are related to apoptosis: cluster 1 is enriched in cell death and other clusters contain genes associated with apoptosis. Clusters in this network are also enriched in signalling and the cluster hierarchically on top (cluster 6) is enriched in neurological disease.

The model of recovery phase (Figure 2.12) is larger, as many more clusters change after the removal of chemical. This model places MGF as the most downstream node, similarly to the model made for the full time-course. This indicates that the previous model made for the full time-course might have captured mainly changes associated with the recovery stage. As the clusters are same as used before, the annotations are same. However, to ease interpretation, separate visualisations were made for stress, signalling, apoptosis, response to calcium ion, chromatin regulation and endocytosis (based on Figure 2.10). These terms are visually highlighted in Figure 2.13. Overall, drug metabolic process is enriched in all clusters except clusters 8 and 12 (Figure 2.10). ATP metabolic process, oxidative phosphorylation, organonitrogen compound metabolic process and transport are significant in all clusters except clusters 2 and 8 (Figure 2.10). The most upstream node, cluster 3, which is also directly upstream of MGF, is enriched in terms related to stress, signalling and apoptosis (Figure 2.13). Different types of signalling are also enriched in many other clusters (clusters 10, 5, 4, 12, 1, 6) (Figure 2.10, Figure 2.13).

2.4.4 The analysis of transcriptional response to nerve damage in rat supports the prediction of the dynamical model

The exploratory analysis and the network modelling results suggests that the organisms recover from exposure. Moreover, the analysis of the model derived from the recovery-stage dataset places conduction velocity of the MGF as the most downstream node as in the model of full time course, suggesting that transcriptional changes associated to recovery may mechanistically control the conduction velocity of the MGF. In order to further test this hypothesis, a previously published proteomics study [155] of recovery after

Table 2.1: Functional annotation of clusters of the RDX model (KEGG terms). Only terms with FDR <0.05 are shown. For every cluster, representative genes from significant pathways are also shown. *Human orthologs annotated with *Italics* were predicted but worm sequences did not contain functional domains.

Cl.	**KEGG pathways**	**Representative genes (human gene ortholog name and description from the Genecards database)**
1	Lipid metabolism; Small molecule biochemistry; Cell to cell signalling; Cell death	DCTN6 – Dynactin subunit 6
2	Transport; DNA replication, recombination, repair; Energy; Cell death; Immune response	RCN2 – Reticulocalbin 2, EF-Hand Calcium Binding Domain
3	Cell death; Nervous system development	CHMP6 – Chromatin Modifying Protein 6; this protein is part of ESCRT-III; AIFM2 – Apoptosis Inducing Factor, Mitochondria Associated 2
4	Cell to cell signaling; Gene expression; Glucocorticoid receptor signaling; Assembly of RNA polymerase II complex	TRIAP1 – TP53 Regulated Inhibitor of Apoptosis CDK7 – Cyclin Dependent Kinase 7 KCNAB2 – Potassium Voltage-Gated Channel Subfamily A Regulatory Beta Subunit 2 SMOC1 – SPARC Related Modular Calcium Binding 1
5	SWI/SNF related protein	SMARCD1 – SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin; EFCAB1 – EF-Hand Calcium Binding Domain 1; SEC63 – SEC63 Homolog, Protein Translocation Regulator
6	Cell death; RNA trafficking; Neurological disease	CNBP – CCHC-Type Zinc Finger Nucleic Acid Binding Protein; CHMP1B – Chromatin Modifying Protein 1B; DBI – Diazepam Binding Inhibitor (GABA Receptor Modulator, Acyl-Coenzyme A Binding Protein); DDR2 – Discoidin Domain Receptor Tyrosine Kinase 2; PRKAA2 – Protein Kinase AMP-Activated Catalytic Subunit Alpha 2, Acetyl-CoA Carboxylase Kinase; LLPH – LLP Homolog, Long-Term Synaptic Facilitation
7	Drug metabolism; Lipid metabolism; Small molecule biochemistry;	FLNB – Filamin B, Beta (Actin Binding Protein 278); MYL12B – Myosin Light Chain 12B
8	Lipid metabolism; Small molecule biochemistry; Gene expression; Retinoic acid mediated apoptosis signaling	DAP3 – Death Associated Protein 3
9	Lipid transport; Cell to cell signaling	CARSHP1 – Calcium-Regulated Heat-Stable Protein 1; APOA1BP – APOA1 Binding Protein (also in rat nerve study)
10	Cell morphology; Developmental disorders; Cell cycle; Chromatin remodelling; Calcium signaling	CHRNA7 – Cholinergic Receptor Nicotinic Alpha 7 Subunit; ARF3– ADP Ribosylation Factor 3
12	Cell to cell interaction; Cell death; Oxidative phosphorylation	*CCT6 – Chaperonin Containing TCP1, Subunit 6A; *SYNCRIP – Synaptotagmin Binding Cytoplasmic RNA Interacting Protein

Table 2.2: Functional annotation of clusters of the RDX model – Ingenuity canonical pathways

Cluster	**Canonical pathways from Ingenuity**
1	Germ cell-sertoli junction signalling; 14-3-3 mediated signalling; Beta-alanine metabolism; Mitochondrian dysdunction; Oxidative phosphorylation
2	Valine leucine isoleucine degradation; Fatty acid metabolism; Tryptophan metabolism; Butaonate metabolism; Oxidative phosphorylation
3	Clathrin-mediated endocytosis signalling; VEGF signalling; Fatty acid metabolism; Mitochondrian dysdunction; Oxidative phosphorylation
4	Valine leucine isoleucine degradation; Purine metabolism; Fatty acid metabolism; Mitochondrian dysdunction; Oxidative phosphorylation
5	Tyrosine metabolism; Tryptophan metabolism; Arginine and proline metabolism; Mitochondrian dysdunction; Oxidative phosphorylation
6	Valine leucine isoleucine degradation; Beta-alanine metabolism; Propanoate metabolism; Fatty acid metabolism; Oxidative phosphorylation
7	Virus entry via endocytic pathways; Caveolar mediated endocytosis signalling; Protein kinase A signalling; Regulation of actin-based motility by rho
8	Germ cell-sertoli junction signalling; 14-3-3 mediated signalling; Clathrin-mediated endocytosis signalling; Virus entry via endocytic pathways; IGF-1 signalling
9	RhoA signalling; Glycolysis/gluconeogenesis; Bile acid biosynthesis; Mitochondrian dysdunction; Oxidative phosphorylation
10	Valine leucine isoleucine degradation; Purine metabolism; Glycolysis/gluconeogenesis; Mitochondrian dysdunction
12	Germ cell-sertoli junction signalling; Axonal guidance signalling; 14-3-3 mediated signalling; Oxidative phosphorylation

Figure 2.10: Gene Ontology Biological Process annotation of clusters of genes in response to RDX exposure. Only terms with FDR<0.05 are shown. Green represents smaller FDR values close to 0 and red larger FDR values closer to 0.05. Cells where FDR>0.05 are shown as white.

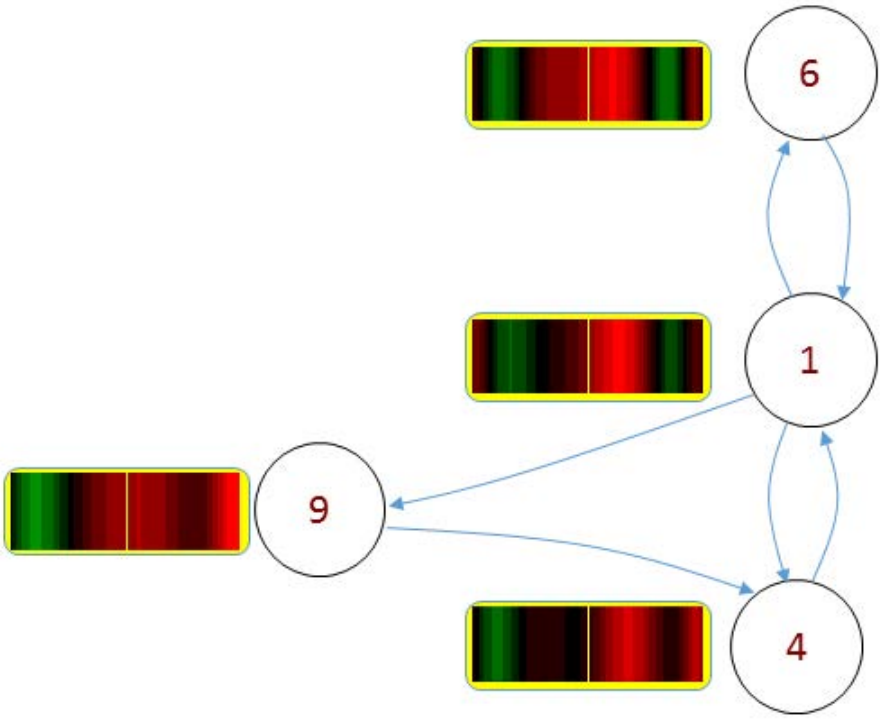


Figure 2.11: NIMOO model of clusters changing only during exposure. Red indicates high expression and green low expression.

nerve damage in rat was used. In this study, many axonal transport proteins were found to be significantly altered in abundance. Unfortunately, many of the probes representing worm axonal transport proteins were eliminated from the dataset because they failed the quality control. However, some genes related to axonal signalling remained in the dataset and axonal signalling was enriched in cluster 12, which is a central node in the network.

The proteins important for the nerve recovery in rat were compared with corresponding genes in the RDX dataset – there were 16 genes/proteins that overlap. These worm genes and corresponding proteins in rat are shown in (Table 2.3). Most interestingly, genes/proteins related to calcium signalling are among the 16. Moreover, although many tubulin genes were not used in further analysis due to being annotated in multiple clusters, there are some that were specific for only one cluster and overlap with proteins from the rat study. Interestingly, 5 of the genes corresponding to the significant rat proteins are in cluster 4 which is a central node in the network and also upstream of the MGF (Figure 2.14, Table 2.3).

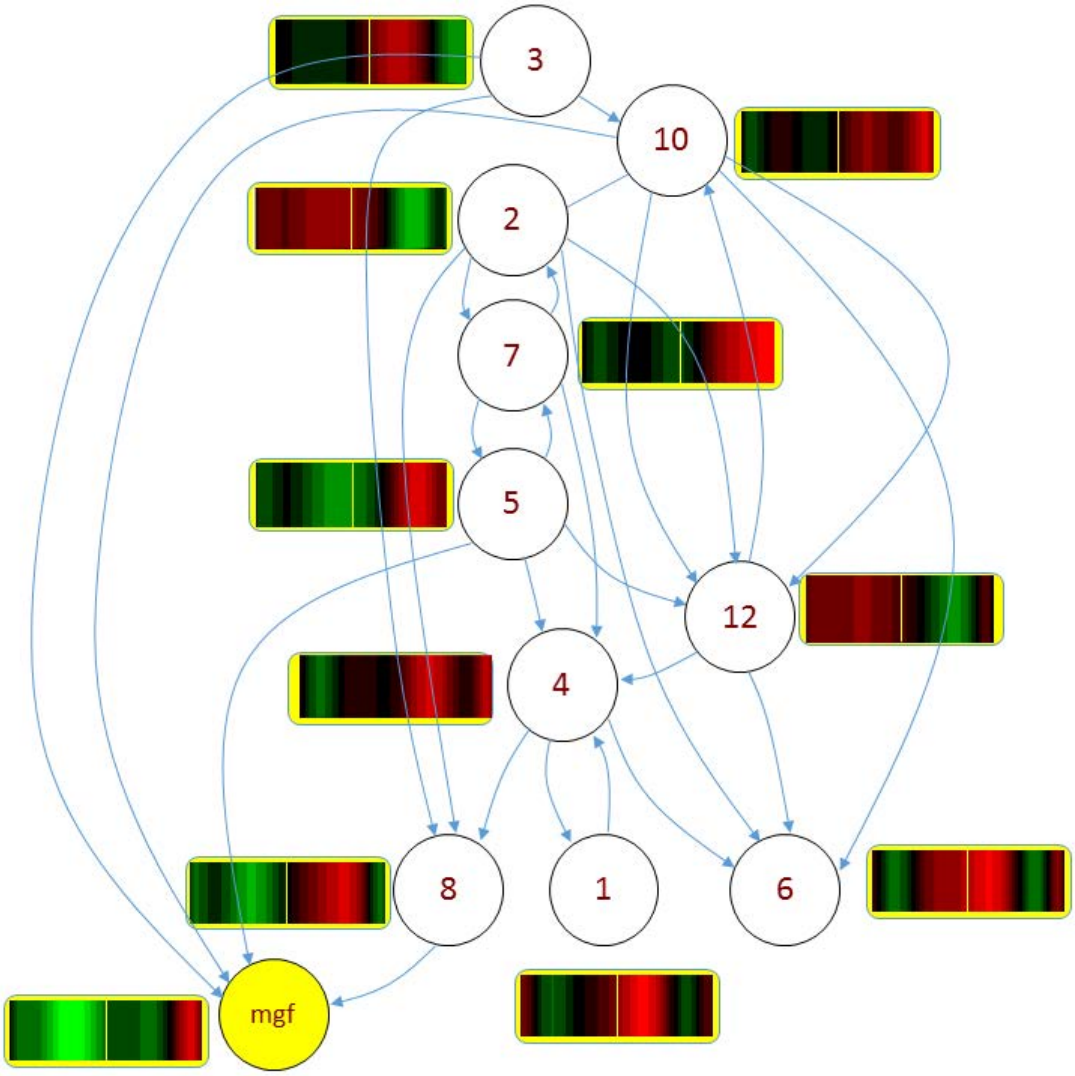


Figure 2.12: NIMOO model based on recovery phase. Red indicates high expression and green low expression.

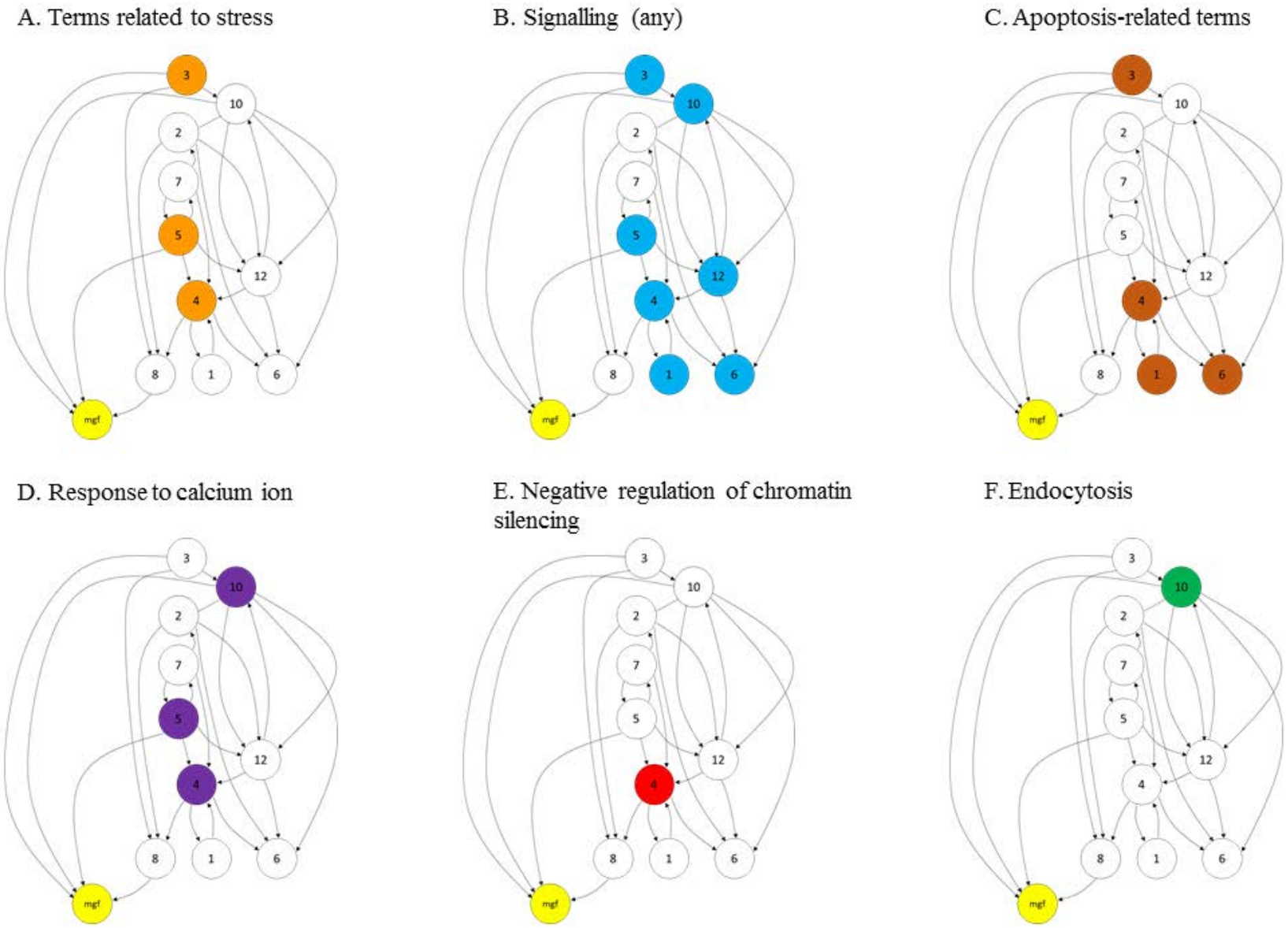


Figure 2.13: Examples of significant (FDR<0.05) Gene Ontology Biological Processes mapped onto network of the recovery phase. A. Terms related to stress (shown in orange). B. Any significant terms related to signalling (blue). C. Apoptosis-related terms (brown). D. Response to calcium ion (purple). E. Negative regulation of chromatin silencing (red). F. Endocytosis (green). Detailed Biological Processes are shown in Figure 2.10.

Table 2.3: Overlap of genes corresponding to proteins found significant in the rat proteomics study of nerve regeneration [155] and differentially expressed genes in response to RDX in the current study.

Protein name (rat)	**Name**	**Cluster**
CALU_RAT	Calumenin	2
AL1A3	Aldehyde dehydrogenase	4
AL1A7	Aldehyde dehydrogenase	4
ATPB	ATP subunit beta, mitochondrial precursor	4
RSSA_RAT	RPSA, 40S ribosome protein, laminine receptor 1	4
TBB2B_RAT	Tubulin beta 2B	4
ALDR	Aldehyde reductase	6
VIM	Vimentin	6
CALM_RAT	Calmodulin	7
CATB	Cathepsin B precursor	9
AAA	Aldolase dehydrogenase	9
FGB	Uncharacterised protein, similar to fibrinogen C-terminal domain	9
TBB2C	Tubulin beta 2C	10
ENOG	gamma-enolase	10
TBB5_RAT	Tubulin beta 5 chain	12
MMSA_RAT	Aldehyde dehydrogenase	13

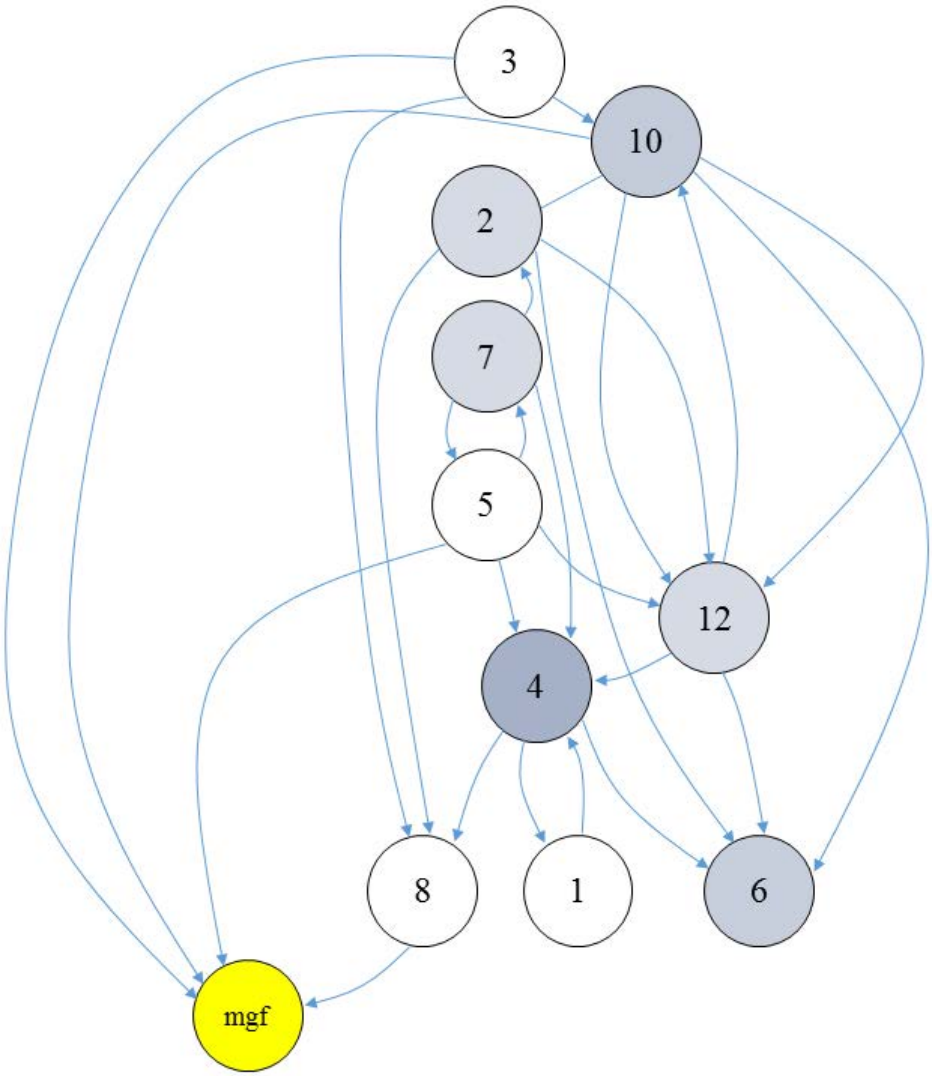


Figure 2.14: Mapping of genes corresponding to rat proteins into clusters of the earthworm model. Intensity of the node colour indicates the number of mapped genes.

2.5 Discussion

2.5.1 Temporal gene expression trajectories are consistent with the existence of chemical-removal mechanisms

Gene expression analysis showed that for two neurotoxic chemicals, the affected genes are very different. Moreover, the Principal Component Analysis suggests that carbaryl might be removed from the organism rapidly, allowing the worm to start recovering as soon as the chemical is not present in the environment. This is consistent with the fact that there is a removal mechanism for carbaryl in the earthworm ([299], reviewed in [314]). However, according to literature, RDX is eliminated rapidly as well [358], making in even more intriguing why the recovery does not start after the removal of the chemical from the environment. In the current study, the gene expression pattern continues after removal of RDX, suggesting that that if the chemical is removed, the continued gene expression changes might be caused by something else. Moreover, there were fewer clusters of genes changing during the exposure than during recovery. It is therefore possible that the network relating the gene expression changes to the conduction velocity of the medial giant nerve fibre (MGF) captures recovery from nerve damage during this process instead.

2.5.2 Calcium signalling, apoptosis and endocytosis in nerve damage recovery

Calcium signalling is an important process and has previously been proposed to be involved in the recovery from nerve damage [120]. Genes involved in cell to cell signalling and cell death have also been shown to be important by other omics studies of nerve injury recovery [353]. Another transcriptomics study of nerve recovery used clustering approaches and showed clusters enriched in calcium signalling [120]. Interestingly, in mentioned studies, immune functions were also highlighted, but were not enriched in the worm study. This might be because immune-related genes might have been lost in the annotation process, as the other studies were in rats (and as model species and mammals,

they have better annotation). The annotation also relies on conservation between rat and earthworm and even if immune functions were affected, genes involved in immune system processes might not have been annotated. Interestingly, the hierarchically most downstream node in the exposure network, cluster 4, which has very little dynamics during exposure but has higher expression levels during recovery, contains a gene related to inhibition of apoptosis (TRIAP1) [254] – and apoptosis is enriched in clusters of the recovery network. Cluster 4 is also a central cluster in the recovery model and moreover, it contains the largest overlap with the rat proteomics study of nerve damage recovery.

Another term enriched in several clusters, endocytosis is a process important for nerve functions [78] and plays a role in the recovery from nerve damage, especially in relation to membrane trafficking [318]. Additionally, endocytosis is initiated by Ca^{2+} and calmodulin [352] and calcium signalling plays a role in endocytosis in synaptic vesicles [66, 78]. One of the genes in the current study which overlapped with corresponding protein from the rat study [155] was vimentin. Moreover, a review about nerve injury signalling [1] argued that the expression of this gene is affected by Ca^{2+} and Na^{2+} perturbations following nerve injury.

In addition to biological functions, there are several genes in the clusters also functionally relating the enriched pathways. For example, it is known that calcium can upregulate calcineurin (cluster 2 in the model) and this might also play a role in endocytosis in synapses [307]. Calcineurin has been shown to be pro-apoptotic [10]. Another gene, calreticulin, which is also in cluster 2, is a calcium-binding protein, chaperone in the endoplasmic reticulum (ER), responding to ER stress [359, 65]. Additionally, Sec63, which part of the Sec62/Sec63 complex on the human ER is also involved in various signalling events (reviewed in [198]).

2.5.3 Is the model consistent with known effects of RDX?

It has been shown that RDX exposure affects GABA signalling by reducing $GABA_A$ currents [347]. The effects of RDX on GABA receptor have been shown to be conserved

between various species [105]. In the models of the current study, the ortholog of human diazepam binding inhibitor (DBI), a GABA modulator was in the top node in both the network based on the full time-course and in the network based on only exposure. In humans, RDX can cause seizures [220, 117] and a gene in one of the top nodes in the recovery network, CHRNA7, has been shown to be associated with encephalopathy and seizures in humans (Endris et al. 2010). As the modelling approach and additional analysis of rat nerve damage suggested, the network might capture recovery from nerve damage. This is also in concordance with previous studies in the earthworm, where RDX caused neurological symptoms [123]. Additionally, as the Principal Component Analysis of the gene expression indicated, the recovery is not immediate after placing the earthworms in clean environment, despite rapid removal of RDX as suggested by literature [358]. It has also been previously shown using a physiological endpoint, that after RDX exposure, the conduction velocity in the medial giant nerve fibre, although returning toward normal levels, never reached the original conduction velocity [123].

2.5.4 Conclusions and possible improvements

In this chapter, I have shown that by data reduction and ODE models using multi-objective optimisation, that it is possible to make biologically meaningful dynamical models relating high-dimensional gene-expression data with a physiological measurement. Moreover, the RDX model suggested that gene expression dynamics capture recovery from nerve damage and this hypothesis is supported by previous studies form literature. This shows the power of data-driven systems biology in non-model species. At the time of the study, the annotation of the *Eisenia fetida* transcripts was challenging. However, the nerve cord transcriptome annotation was made available in 2017 [252]. This makes it possible to interpret the transcripts further and potentially facilitate the interpretation of changes. It might also solve the problem assigning gene identities to multiple clusters. In our approach, such genes were removed from analysis, removing potentially important biological information. If the annotation was improved, these might give more insight

into the effects of chemical exposure and changes after the removal of chemical. For example, the overlap analysis with rat nerve damage study [155] could be more extensive. Another limitation is the reliance on human orthologs. It can give some functional insights, especially is the sequences between human and earthworm are conserved, but to understand specific mechanisms of nerve damage or recovery, more experiments or the use of earthworm-specific additional datasets would be needed. However, as the main focus on this chapter was to develop a strategy for analysing dynamical changes in non-model organisms, and also relate these to other types of measurements, such as physiological parameters, the current analysis with existing annotation serves the purpose.

CHAPTER 3

MODELING THE METABOLIC PROFILE OF *MYTILUS EDULIS* REVEALS MOLECULAR SIGNATURES LINKED TO GONADAL DEVELOPMENT, SEX AND ENVIRONMENTAL SITE

3.1 Contributions

Parts of the data used for this article and thesis chapter are already published by co-authors (A. Hines thesis and J. P. Bignell [31] using univariate and multivariate statistics. The focus of this chapter and paper is systems biology and analysing all types of data together in a data-driven manner. All the analyses based on previously published data in this thesis are novel.

- Jaanika Kronberg-Guzman[1,2,3] performed computational systems biology analysis from the metabolite bins (exploratory analysis, annual cycle dynamical models, sex-prediction), checked automatic metabolite annotations provided by Jonathan Byrne, did additional metabolite annotations based on literature, interpreted results, wrote the paper draft and this chapter

- Jonathan J. Byrne[2] did ^{1}H-NMR data processing and metabolite annotations according to the Birmingham Metabolite Library database. Provided comments on

the paper.

- Jeroen Jansen[5] did statistical processing of the metabolite bins before the systems biology analysis.

- Philipp Antczak[1] provided help and guidence in the systems biology analysis and also useful comments in the manuscript-writing.

- Adam Hines[2] did all the ^{1}H-NMR measurements and had analysed the same data with different methods for his thesis in the University of Birmingham. Provided comments for the paper.

- John P. Bignell[4] was involved in the sampling of the mussels and histopathology (this work on the same mussels is already published [31] – without ^{1}H-NMR or systems biology. Provided comments for the paper.

- Grant D. Stentiford[4] was the leader and initiator of this project especially in respect to design, sample collection and histology.

- Mark R. Viant[2*] was the leader and initiator of this project in the area of ^{1}H-NMR experiments and processing, provided comments for the paper.

- Francesco Falciani[1,2*] oversaw and guided the systems biology aspect of this study, which is the focus of this chapter. Participated in discussions of all systems biology analysis and provided comments for the chapter and paper.

Affiliations

1. Institute of Integrative Biology, University of Liverpool, UK

2. School of Biosciences, University of Birmingham, UK

3. Institute of Genomics, University of Tartu, Estonia

4. Centre for Environment, Fisheries and Aquaculture Science, UK

5. Radboud University, Netherlands

6. * joint corresponding author

3.2 Abstract

Continuous monitoring of anthropogenic pollution is essential for maintaining high water quality standards and to minimize the impact of chemicals on aquatic wildlife. In parallel with monitoring the concentrations of chemicals of concern, biosensor species are essential tools for environmental monitoring. Among these, mussels are filter-feeding and sessile, hence potentially a good model system for measuring localized pollution. Here we address the hypothesis that the metabolic state of the blue mussel, *Mytilus edulis*, characterized by ^{1}H-NMR spectroscopy, is correlated to organism physiology and that this relationship is affected by the environment. We approach this challenge by developing a computational model representing the reference site and integrating the metabolite seasonal dynamics with key physiological indicators and environmental parameters. The analysis of the model revealed that changes in metabolite levels during the annual cycle are potentially influenced by water temperature and are linked to gonadal development. Moreover, a statistical model trained in the reference site to predict sex from metabolite markers, forecast the presence of "molecular intersex" in a population of mussels sampled from Southampton. This work shows the power of data-driven metabolomics and its potential in environmental monitoring.

3.3 Introduction

Anthropogenic pollution affects water quality and consequently represents a threat for ecosystem functioning and human health. Therefore, with a constant increase in the release of chemicals, there is a need to improve environmental monitoring, especially to detect early biotic effects of chemical contamination. In the European Union, this monitoring is performed under the umbrella of the Water and Marine Water Framework Directives [77]. In the US, monitoring is handled by the Environmental Protection Agency′s (EPA) National Aquatic Resource Surveys according to the Clean Water Act [323]. Both aim at providing and supporting healthy biological communities in surface and groundwater

bodies.

Currently, monitoring is performed by measuring water physico-chemical parameters and biodiversity. In addition, sensitive indicator species are used to assess ecosystem health and provide additional information on potential environmental pollutants [142]. While absence of such species is an indicator in itself, molecular tools have been proposed as a more sensitive approach capable of detecting organism-level effects very early on, with potential predictive power in determining a future deterioration of a populations fitness [167, 45, 346, 275]. A few of these markers have been proven useful, such as vitellogenin, which in male fish is a biomarker for endocrine disruption indicating decline in reproductive potential [9, 138, 159]. The advent of functional genomics technologies has provided the community with powerful tools which can be used to identify more complex molecular signatures predictive of toxicity [326, 329, 277, 140, 18, 30, 41, 258]

Mussels represent an excellent indicator species due to their wide range of habitat, including both salt and freshwater, geographical distribution and their ability to filter vast amounts of water (*e.g.* *Mytilus edulis* can filter up to 15ml water in a minute) [264, 128, 363, 118, 237]. Mussels have been used in the Mussel Watch program long before the advent of functional genomics technologies [118]. Physiology-based biomarkers of toxic effects in mussels, such as scope for growth and survival stress tests (*i.e.* time to death outside water) [135] have been recommended by the International Council for the Exploration of the Sea (ICES) [115]. Mussel biomarkers have been shown to have a broad range of applications ranging from assessing the effects of UV filters [15], urban wastewater [73], oil pollution [23, 236, 291, 320], offshore gas platforms [126] and a wide range of environmental pollution [96, 114, 363].

More recently, functional genomics has been applied to study the mussel stress response, such as salinity or response to the tidal cycle, in a number of environmentally relevant species [61, 19, 76, 60, 201, 256]. While most of the ′omic′ studies have relied on gene expression profiling, other omics technologies, such as metabolomics [91, 319, 290, 178, 140] and proteomics [8, 168, 292] have also been used and may provide a strongly

phenotype-oriented view of the animal that aligns well with the objective of environmental effects monitoring.

Because of the vast amount of data generated by these studies, untangling the effects of multiple environmental stress factors and identifying molecular mechanisms controlling organism physiology requires advanced computational methods. The work in this chapter tests the potential of systems biology to develop metabolism-based biomarkers of environmental relevance. I approach this challenge by modelling the complexity of the global metabolite changes during the annual cycle of blue mussel *Mytilus edulis* and their relationship with environmental location, physical parameters such as temperature and salinity, and physiological parameters such as gonadal stage and adipogranular tissue index (ADG rate). These models identify a hierarchy of connected molecular and physiological events that are linked to gonadal development. I show that these metabolic profiles can predict sex in the reference site. Ultimately, I prove that the utility of this approach by showing that in more polluted waters, the metabolic state predicts the presence of intersex organisms, generating a hypothesis that would need to be tested in future studies.

3.4 Methods

3.4.1 Overview of the analysis strategy

Figure 3.1 summarizes the different components of the study, including the data analysis strategy. The first step (data production) involves the generation of a dataset representing the metabolic state of mussels sampled from Exmouth (reference site) and Southampton (more polluted site) over the period of one year (Figure 3.1A). During the data acquisition the mantle tissue is sampled to acquire both histological and ^{1}H-NMR spectroscopy-based metabolomics data. Meanwhile, the environmental variables are collected. The second step (data analysis) involves the identification of metabolites linked to sex, site and sea-

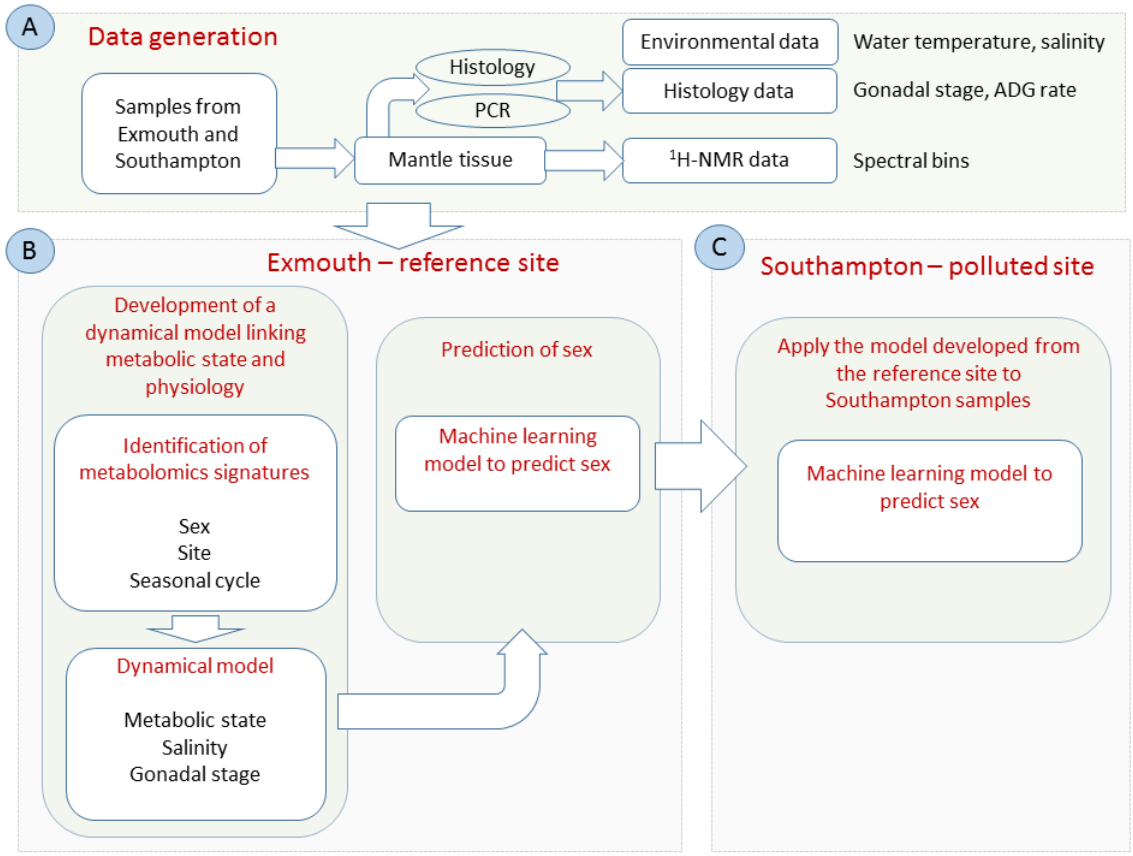


Figure 3.1: Overview of the data generation and analysis strategy

sonal cycle in the Exmouth reference site, the development of a dynamical model linking molecular changes to environmental and physiological parameters, and the development of a statistical model that can predict sex from the metabolote levels (Figure 3.1B). In the third step, the application of the sex prediction model (developed for the reference site and applied to the Southampton site) led to the hypothesis that the molecular state characterizing sex in mussels sampled from the Exmouth reference site is altered in organisms sampled from the more polluted site (Figure 3.1C).

3.4.2 Sample collection, determination of sex and physiological variables (CEFAS)

A total of 50 mussels were collected from a rural reference site (Exmouth) and an industrial harbor site (Southampton), every month over a period of one year [31]. Species, sex and

physiological parameters (gonadal status and the adipogranular tissue scoring index – ADG rate) were determined as previously described [31]. Gonadal status was defined as the degree of gonadal maturity scored on a scale between 0 to 4. The ADG rate was defined as the adipogranular tissue scoring index, where 0 represents absence of ADG cells in vesicular connective tissue and 4 shows that majority of connective tissue volume are ADG cells. ADG cells are characterized by intracellular granules storing protein and small amount of lipids and glycogen as energy reserves. Only individuals for which sex could be determined by histology were used for ^{1}H-NMR metabolomics. In the analysis steps up to model creation, I concentrated on *M. edulis* only, as this species was present in both sites. However, due to low numbers of male *M. edulis* in Southampton, *M. galloprovincialis* and the hybrids of both species were included in the prediction of sex. ADG rate, gonadal stage, parasite load and temperature, as described here have been previously published by Bignell *et al.* [31] and are used in this thesis as variables in the model.

3.4.3 ^{1}H-NMR metabolomics analysis (Mark Viant's group)

Metabolite extraction: Following dissection and subsequent snap freezing in liquid nitrogen, polar metabolites were extracted from mussel tissues using a methanol:water:chloroform solvent system and a Precellys 24 homogeniser (Stretton Scientific, UK), as described previously [351]. Each dried polar metabolite extract was resuspended in 650 µl of sodium phosphate buffer solution (0.1M, 90% H2O/10% D_2O, pH 7.0) containing an internal chemical shift standard of 1 mM sodium 3-trimethylsilyl-2,2,3,3-d4-propionate (TMSP). Data acquisition: Samples were analysed using a Bruker Avance III 500 MHz NMR spectrometer operating at 500.18 MHz ^{1}H resonance frequency, equipped with a 5 mm cryoprobe and BACS-60 automatic sample changer (Bruker Biospin, Coventry, UK). For each sample a two-dimensional ^{1}H, ^{1}H J-resolved (JRES) NMR spectrum was acquired using 16 transients per increment for 16 increments, collected into 16k data points, and spectral widths of 6009 Hz (12 ppm) in F2 (chemical shift axis) and 50 Hz in F1 (spin-spin coupling constant axis), with a 4.0-s relaxation delay. Datasets were zero-filled in

F1 and both dimensions multiplied by sine-bell window functions prior to Fourier transformation. JRES spectra were tilted by 45°, symmetrized about F1, and then calibrated (TMSP, 0 ppm), all using TopSpin (Bruker). Data were exported as the 1-D skyline projections (along F2) of the JRES spectra (termed pJRES) [205]. Pre-processing: Each spectrum was binned between 10 and 0.2 with a bin width of 0.005ppm. Two regions were excluded (4.46 to 5.15 ppm, water; 7.6 to 7.76 ppm, residual chloroform from the extraction method). Data were normalized to total spectral area (TSA). Next, due to slight variation in the chemical shifts of some peaks, bins were compressed by calculating their mean. Six regions were compressed (7.11 to 7.16 ppm, 7.96 to 7.99 ppm, 7.99 to 8.02 ppm; 8.18 to 8.20 ppm, 8.26 to 8.29 ppm, and 8.58 to 8.61 ppm). A generalized log (Glog) transformation was performed with lambda = 3.75e-9 [241].

3.4.4 Statistical analysis: ANOVA and clustering

Metabolite bins were used for statistical analysis (ANOVA) and clustering (HOPACH [327, 251]). Using the larger dataset representing the full annual cycle developed with the reference site (Exmouth) samples, we first performed a two-factor ANOVA (as implemented in TMev) and analyzed the effects of time (12 months of field sampling) and sex for *Mytilus edulis*. Since a smaller number of time points was available for the Southampton site a three-factor ANOVA (implemented in the statistical environment R [260]) analyzing the effects of time, sex and site (Exmouth and Southampton sites) was performed. In this ANOVA, time was factorized in three groups (April-May, June-July and December-January). Spectral peaks with FDR-adjusted p-value $< 1\%$ were considered significant. Data visualization of site-effects in clusters was performed on a more stringent statistical threshold to focus on the most significant results (FDR adjusted p-value $<$ 1e-8). To identify sex-specific markers across all species included a two-factor ANOVA was generated. Due to the reduced number of samples within certain species, sampling time was binned into 3 groups (July-Sept, Oct-April, May-June). For visualization purposes, the statistically significant metabolite bins identified by the first ANOVA were standardised

($\mu = 0$, $\sigma = 1$). In order to reduce the complexity of the dataset we clustered significant metabolite spectral bins by using HOPACH [327, 251]. The first round of clustering gave 7 clusters and was performed using the *abscos* similarity measure. Since this groups together negatively as well as positively correlated profile we performed an additional round of HOPACH clustering on cluster 3 that contained such heterogeneous profiles (Figure 3.5, Figure 3.4).

3.4.5 Principal Component Analysis (PCA)

In order to visualize changes in the metabolic state of mussels across the annual cycle a principal component analysis, as implemented by the *prcomp* command in the Stats package within the statistical environment R [260], was used. In short, 3 different PCAs were performed, 1) utilizing all the metabolite bins, 2) using averages computed for each cluster, and 3) for specific subsets of clusters or metabolites (3.14). To visualise the extend of the differences between the two sites a PCA was generated based on the Exmouth samples only. Using the rotation matrix, which represent the calculated PCA parameters used to transform a dataset into principal components, the Southampton samples were projected into the same principal component space.

3.4.6 Metabolite annotations (Jonathan Byrne, Jaanika Kronberg-Guzman

The annotations of the metabolites contributing most to each cluster were then determined using the web-based automated identification tool developed at the University of Birmingham (http://www.bml-nmr.org/). This tool makes use of a library of ^{1}H-NMR spectra of ca. 200 pure metabolite standards [204]. NMR chemical shift data from the literature [343, 151] were also used to check the metabolite identities.

3.4.7 Modelling the seasonal dynamics of mussel's metabolic state

Dynamical models linking metabolite clusters, physiological measurements (gonadal stage, ADG rate, parasite load) and environmental parameters (salinity, water temperature) were built separately for male and female mussels in Exmouth using the algorithm TimeDelay ARACNE as implemented in the TDARACNE R package [362]. As our original dataset consisted of 12 time-points (12 months), we have used polynomial interpolation for each cluster of metabolites to get 120 time points (5th degree polynomial was used). In the TDARACNE models we choose a DPI tolerance of 0.15 and a time delay $N = 20$. The models were represented in a graphical format by graphs where each node represented a cluster of highly correlated metabolites or one of the physiological or environmental measurements. Edges represented the strength of time-dependent relationship between variables. Nodes representing the metabolite clusters were color-coded to reflect the percentage of metabolite bins differentially expressed between Exmouth and Southampton.

3.4.8 Classification

To develop a predictive model able to classify sex a Support Vector Machine (SVM) from R package Kernlab [157] (linear kernel, $C = 70$, $cross = 4$) was used. To train the model the Exmouth dataset was randomly split 5000 times into 75% training and 25% testing datasets. For each split an SVM model was generated based on the significant metabolites. Additional internal training/test splits were used to combat overtraining. Final model accuracy was determined by predicting sex on the independent test dataset. All 5000 models were then used to predict sex for the Southampton samples (including additional samples from *Mytilus galloprovincialis* and hybrid of *M. galloprovincialis* and *M. edulis*). The resulting 5000 predictions for each sample were then averaged to define a representative sex prediction. To confirm the visual inspection that GABA and ATP can discriminate sex in the winter in Exmouth but not Southampton, another KSVM model

was built based on 2 metabolites (ATP and GABA) with linear kernel as before ($C = 70$, $cross = 4$), but only with cross-validation.

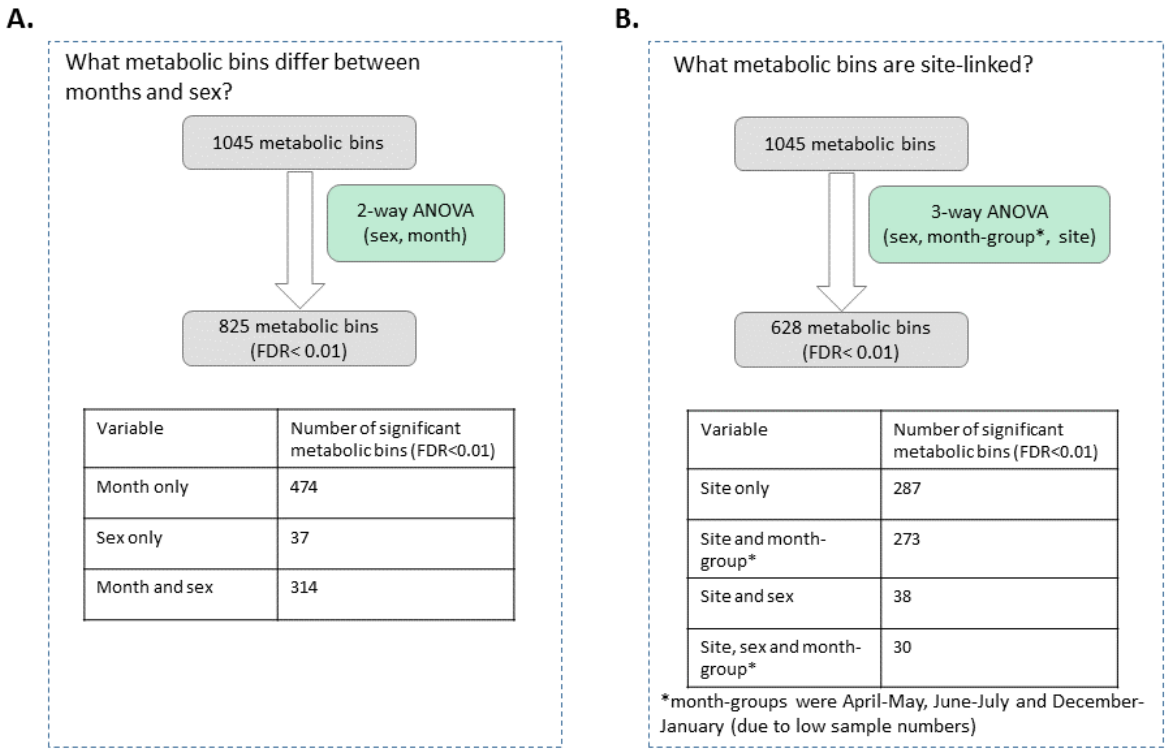
3.4.9 Re-analysis of parasite load from previously published data

Previously published data [31] representing the same mussels were re-analysed to test whether mussels in the Exmouth and Southampton sites were showing alterations in parasite load that could be consistent with the predictions of our model. We tested whether the proportions of counts of parasite load for *Steinhausia mytilovum* and bucephalids in the two sites are the same by using the prop.test function within the statistical environment R [260].

3.5 Results

3.5.1 *Mytilus edulis* mantle metabolic state changes in relation to seasonal cycle, sex and environmental location

I first tested the hypothesis that the metabolic state of the mussel mantle changes during the annual cycle, and that such variation reflects sex and time. By using a two-factor ANOVA in the Exmouth site, I found that 79% of the spectral bins are changing significantly in at least one of the factors tested (825 out of 1045, FDR-adjusted p-value<1%). Of these, 45% (474) were linked to seasonal variation, 30% (314) were changing both in time and between sex, and 37 were sex-specific but time-invariant (Figure 3.2 A). In the comparison between samples from the two geographical sites I discovered 60% (628) site-linked spectral bins. From these site-linked spectral bins, 273 were changing in time only, 38 were sex-specific, and 30 were changing in both time and sex (Figure 3.2 B).



Variable	Number of significant metabolic bins (FDR<0.01)
Month only	474
Sex only	37
Month and sex	314

Variable	Number of significant metabolic bins (FDR<0.01)
Site only	287
Site and month-group*	273
Site and sex	38
Site, sex and month-group*	30

Figure 3.2: ANOVA for metabolic bins. A. ANOVA for metabolic bins in relation to month and sex. Tests were performed using 11 months. B. ANOVA in relation to site, month-group and sex to find site-linked metabolic bins. ANOVA was performed for 3 month-groups (Apr-May, Jun-Jul, Dec-Jan) due to lack of male samples in some months.

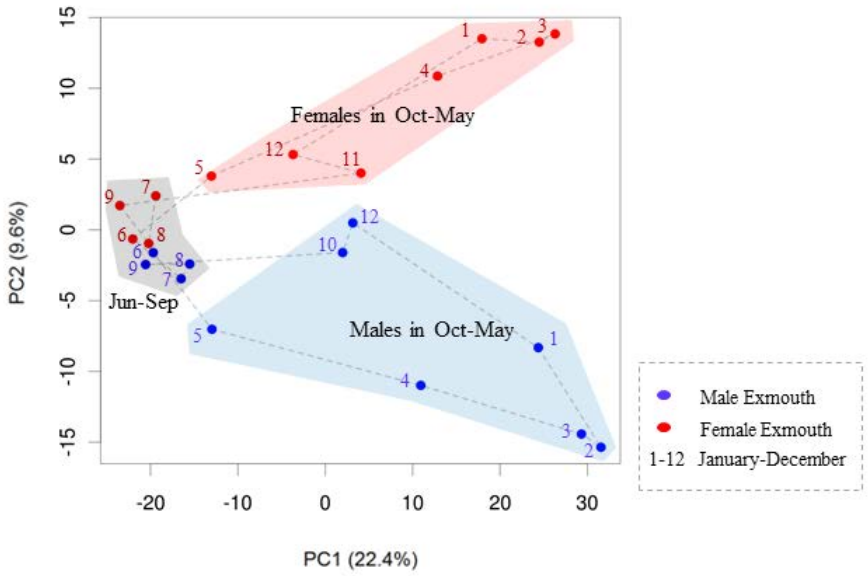
3.5.2 Sex-specific metabolites show different dynamics of seasonal variation in the two sampling sites

Having shown that sex, sampling site and time all affect the metabolic state of mussels, I asked whether the dynamics of metabolite changes are different in the two sampling sites. I first addressed this question by analysing the dynamics of change in the metabolic state of male and female mussels, in the two sampling sites, by using Principal Component Analysis (PCA). Figure 3.3 shows that in Exmouth, male and female mussels are similar in the summer months, start diverging in the autumn, and reach maximum differentiation in late winter (February and March). Interestingly, while samples from the reference site show considerable separation between males and females in the period of peak gonadal development in the winter and early spring (Figure 3.3), Southampton mussel samples collected in December and January are mixed with female samples, suggesting the existence of location-specific alterations during the period of gonadal development (3.3).

3.5.3 Dynamical models, representing seasonal variation identify metabolite profiles linked to temperature, ADG rate and gonadal stage

The PCA revealed that the metabolic profile of environmentally-sampled mussels follows a sex-specific trajectory across a year. It also revealed that the seasonal dynamics are affected in mussels sampled from the Southampton site to the extent that male mussels in winter show a very similar metabolic profile to the samples of female mussels from the reference site. In order to model the relationship between changes in the metabolome and organism physiology across the seasonal cycle, I applied a computational method designed to learn the structure of a dynamic network from observational data. In order to reduce the complexity of the modelling task, I first set to reduce the number of variables to model. Clustering of metabolites changing over the seasonal cycle revealed that the dynamics of the annual cycle can be described by 9 clusters (Figure 3.4, Figure 3.5). Clusters 1, 2, 3.1, 3.2 and 4 have positive correlation with gonadal stage, with maxima

A. Exmouth



B. Exmouth and Southampton

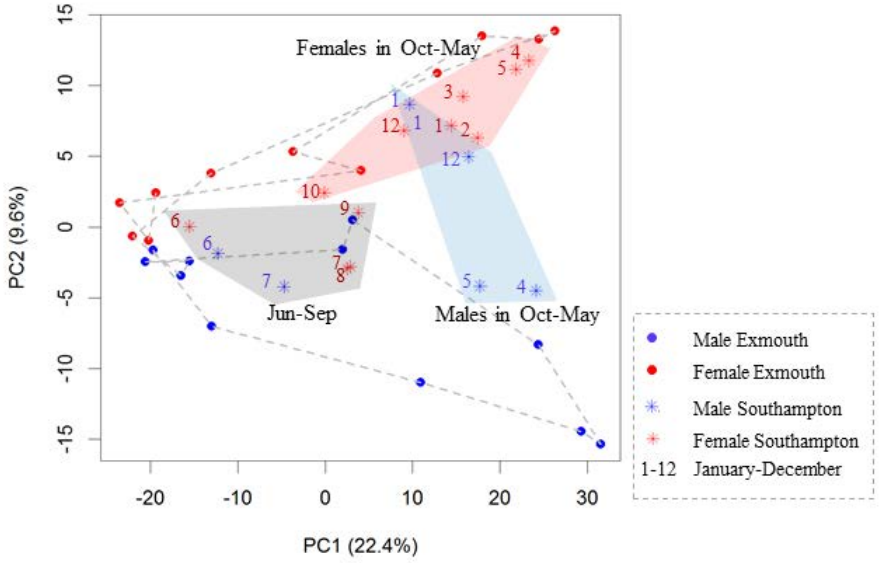


Figure 3.3: Principal component analysis (PCA) of cluster medians in both sex and sites during the annual cycle. A: Cluster medians of male and female mussels in the Exmouth (reference) site. Red represents female mussels and blue male mussels. Summer months are highlighted with a gray box, months from October to May are highlighted with red and blue boxes for female and male mussels, respectively. B. Cluster medians of male and female mussels in Exmouth and Southampton sites. Red represents female mussels and blue male mussels. Summer months of Southapton mussels are highlighted with gray box, months from October to May for Southampton mussels are highlighted with red and blue boxes for female and male mussels, respectively. Numbers 1-12 represent months from January to December. Exmouth mussels are also shown as reference (same as A).

in the winter and negative correlation with temperature, with maxima in the summer (Spearman correlation, as shown in Table 3.1, and Table 3.2). Six clusters (3.1, 3.2, 3.3, 5, 6 and 7) are similar between male and female mussels, as shown in Figure 3.4. Clusters correlated with gonadal stage (1, 2 and 4) show a stronger sex specific response. PCA shows that cluster profiles are sufficient to capture the dynamics of the annual cycle (Figure 3.14).

Then a computational approach was applied to our dataset representing the median cluster profiles as well as relevant physiological variables (gonadal stage, count of ADG cells, shown in Figure 3.5) and environmental parameters (salinity and water temperature). I sought to use sex-specific models developed with this approach to identify molecular signatures that may be linked to gonadal development. The model developed to represent seasonal dynamics in females (Figure 3.6) places temperature as the most upstream node, directly connected to ADG rate and salinity. ADG rate further connects to a downstream layer of metabolite clusters. Interestingly, three metabolite nodes directly connect to gonadal stage which is the most downstream node. The model developed to represent male mussels (Figure 3.7) shows a different structure, from an initial visual inspection. In the male model, temperature is directly upstream of 2 metabolite clusters, 5 and 6, and through them, also cluster 3.2. In the female model, temperature is upstream of the same clusters, although through ADG rate. In both models, cluster 3.3 is central in the network and downstream of ADG rate. Interestingly, in both male and female dynamical models, metabolomics clusters 3.2 and 3.3 are directly upstream of gonadal stage and include the highest percentage of metabolites that are at different concentrations in mussels derived from the two sampling sites (Figure 3.6, Figure 3.7).

3.5.4 Development of sex-specific biomarkers

The analysis of the model of the female mussels described in Figure 3.6 has shown that metabolites that are present at different concentrations in mussels sampled from different geographical locations are upstream of gonadal development. Moreover, as the principal

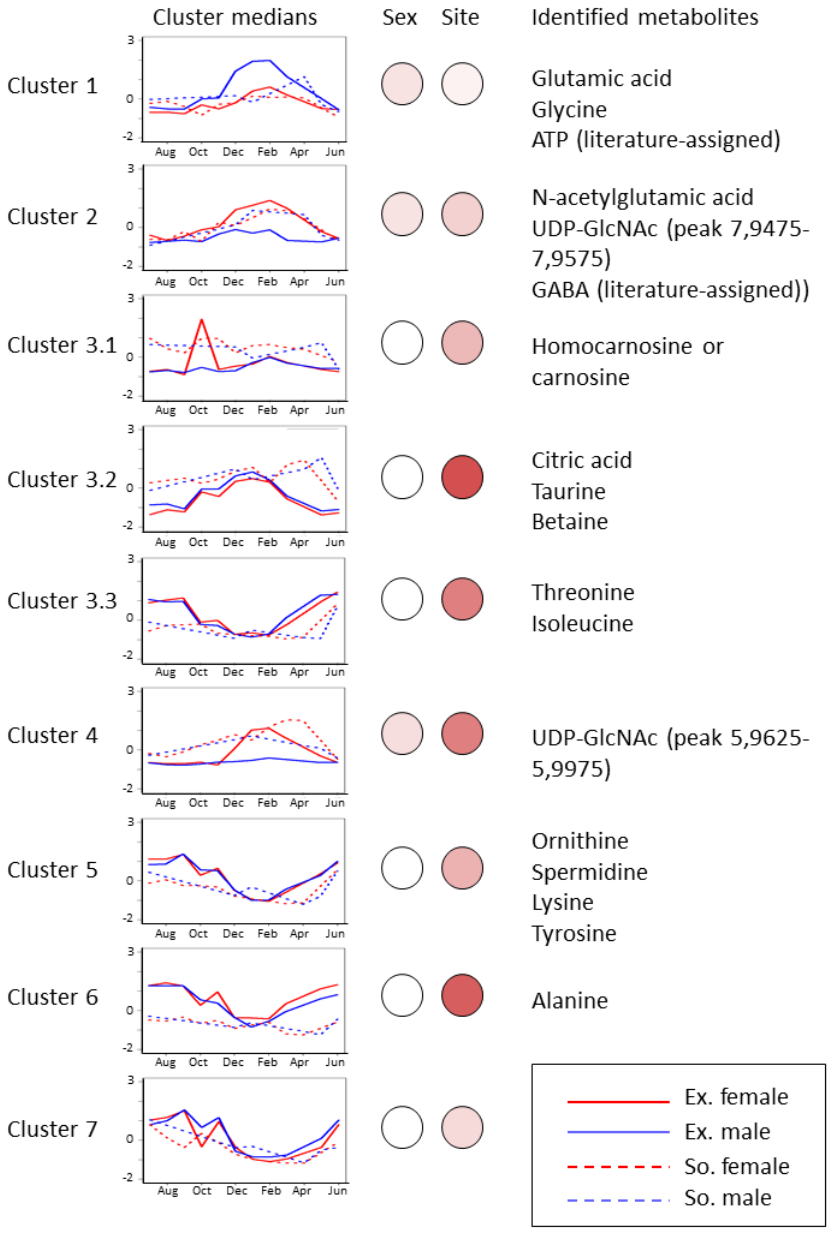


Figure 3.4: Clustering of metabolic profiles in *Mytilus edulis*. Metabolite bins of Exmouth mussels have been clustered with HOPACH, a clustering tool that determines the number of clusters automatically. Cluster medians are used for visualization. The same Exmouth clusters are used to show cluster medians of Southampton samples. Female mussels are represented with red and males with blue, Exmouth with solid line and Southampton with dashed line. Percentage of significantly different (FDR adjusted p-value 1e-8) bins in each cluster is shown by colour intensity. Putative metabolite identities are also shown.

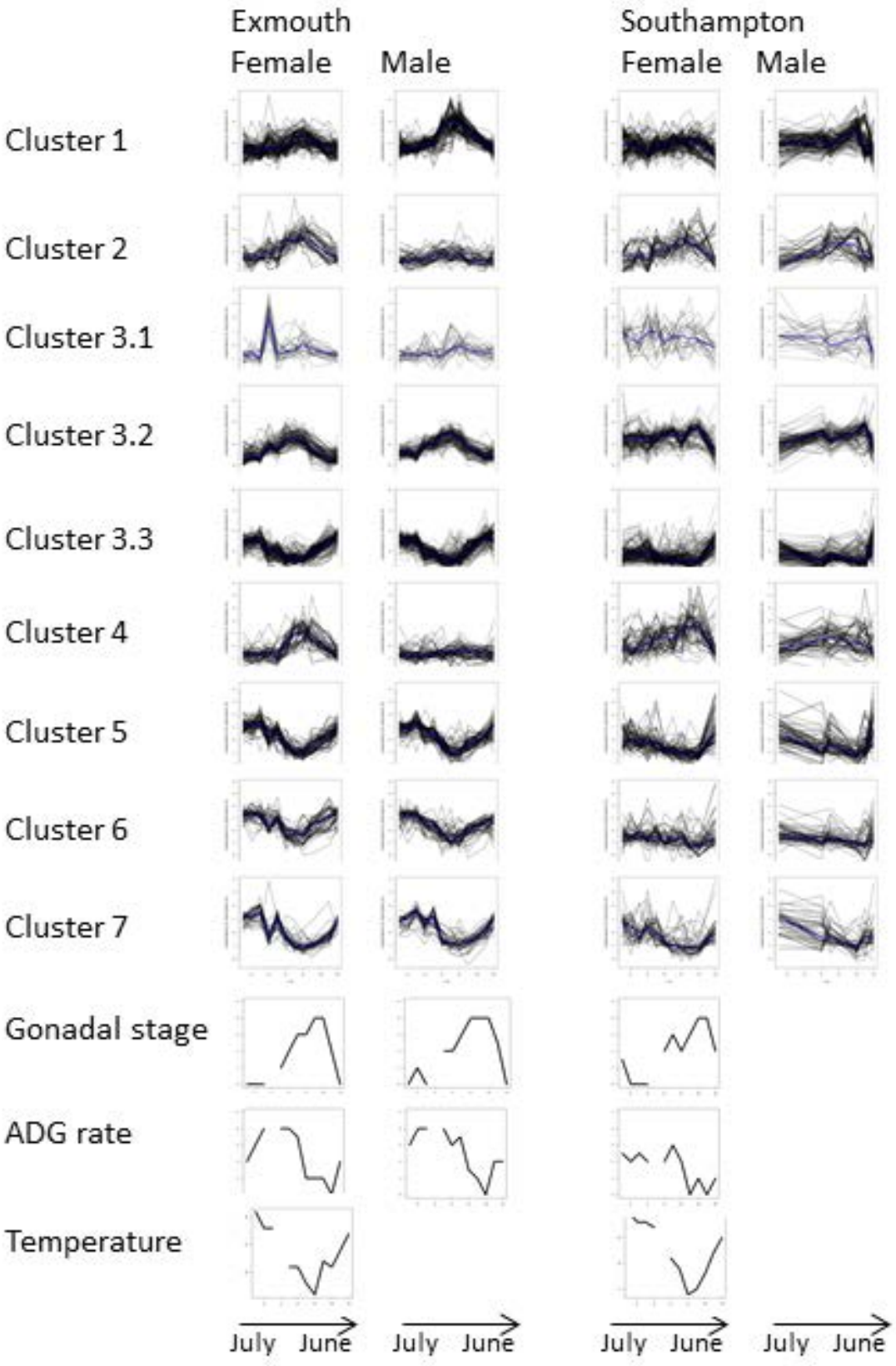


Figure 3.5: Alternative clustering visualisation for temporal profiles of female and male mussels (*Mytilus edulis* in Exmouth and Southampton

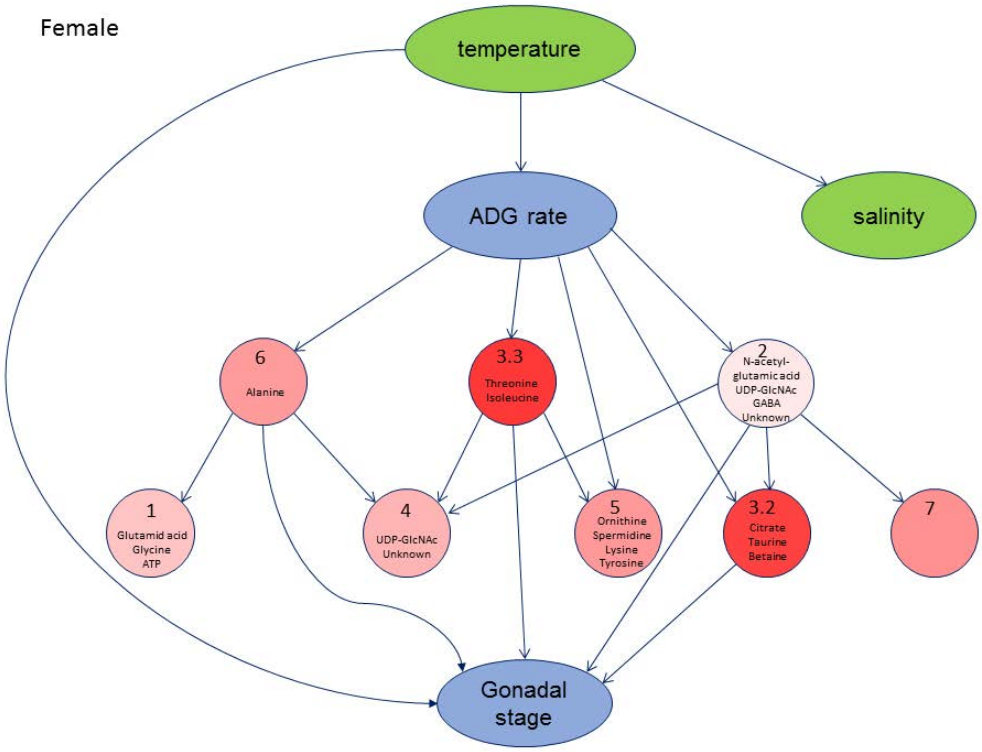


Figure 3.6: Dynamical TDARACNE model of the female mussels (*Mytilus edulis*) sampled from Exmouth (reference site). Red nodes represent metabolite clusters, green nodes environmental variables and blue nodes physiological variables. Intensity of red shows the percentage of metabolites significantly different (FDR adjusted p-value 1e-8) between sites.

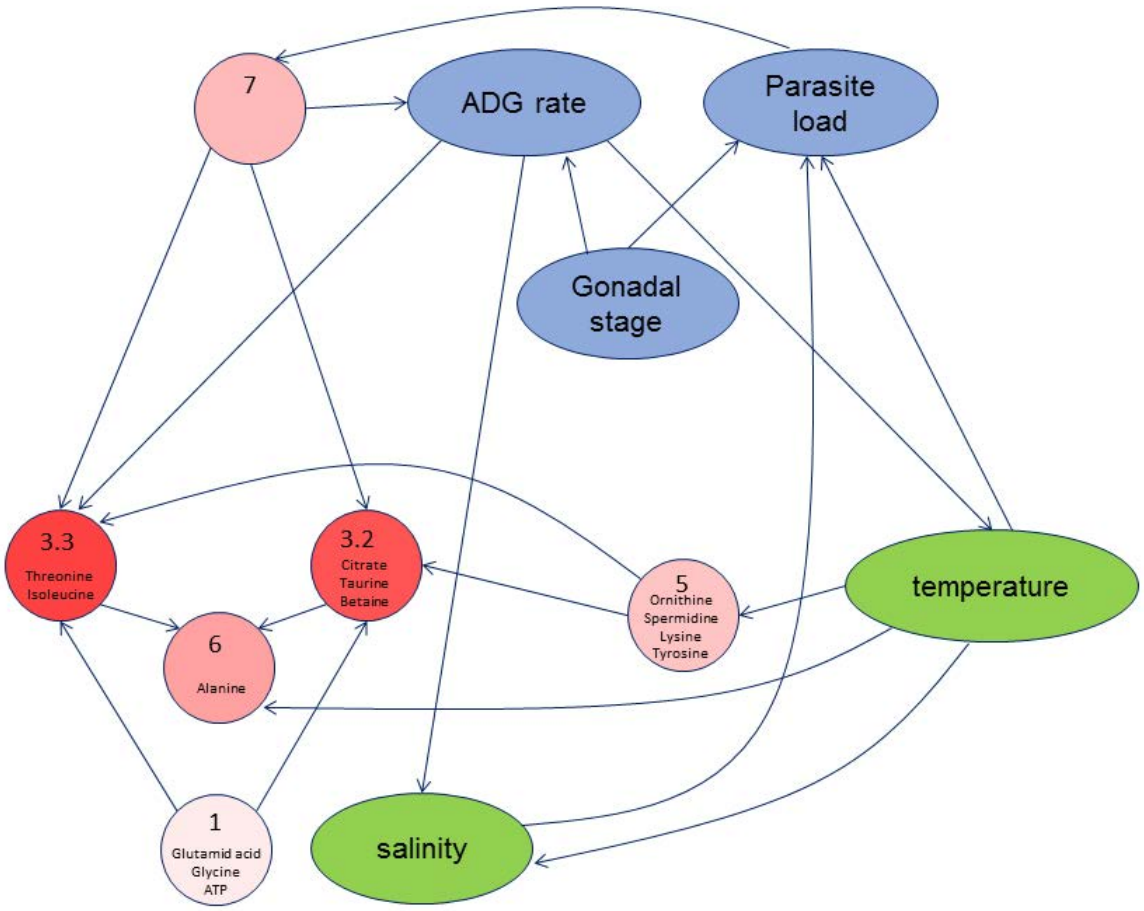


Figure 3.7: Male TDARACNE model in *Mytilus edulis*). Blue: physiological measurements, green: environmental measurements, red: metabolite levels. Intensity of red for metabolite levels indicate the percentage of metabolite bins significantly different between Exmouth and Southampton

component analysis (Figure 3.10) showed that especially metabolites in clusters upstream of gonadal stage might reflect the effects of site, we reasoned that such metabolites may be effective biomarkers to monitor the detrimental effects of environmental stress on the development of the reproductive system. I tested this by developing a statistical model that can predict sex from the metabolic state of mussels sampled from the reference Exmouth site (Figure 3.11, Figure 3.8, Figure 3.9). The model was very accurate in identifying male and female mussels on the basis of their metabolic profile (97.8% accuracy) in the months between October and April. The model is still effective but less accurate at other times of the year (Figure 3.11). This is consistent with the timing of the development of the gonads (greatly reduced after spawning in late spring and early summer). Metabolites contributing to the prediction were mostly mapped to clusters 1, 2 and 4 (Table 3.4), from which cluster 1 is mostly changing in males and not in females, and clusters 2 and 4 are female-specific, with no change in male mussels in the reference site. The top 20 metabolite bins predictive of sex are shown in Table 3.3 together with their putative identities.

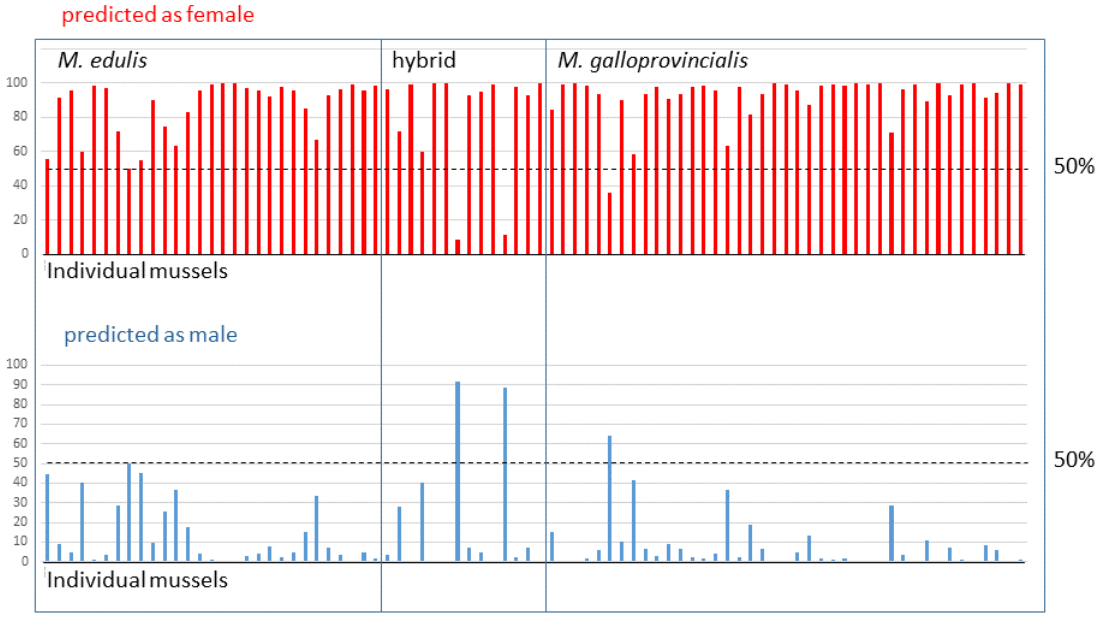


Figure 3.8: Sex predictions for female Southampton mussels for all species (*Mytilus edulis*, *Mytilus galloprovincialis* and their hybrid) from October to April. Bars represent individual mussels.

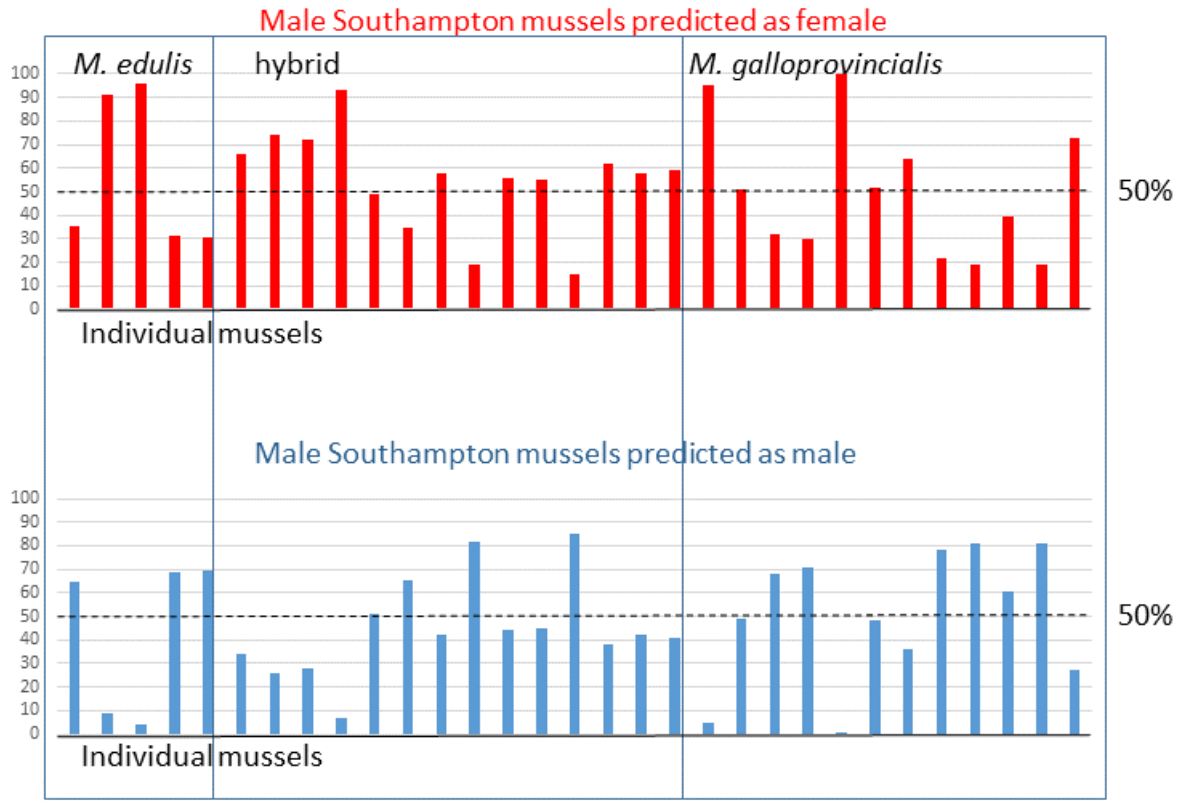


Figure 3.9: Sex predictions for male Southampton mussels in December and January: all species (*Mytilus edulis*, *Mytilus galloprovincialis* and their hybrid). Bars represent individual mussels.

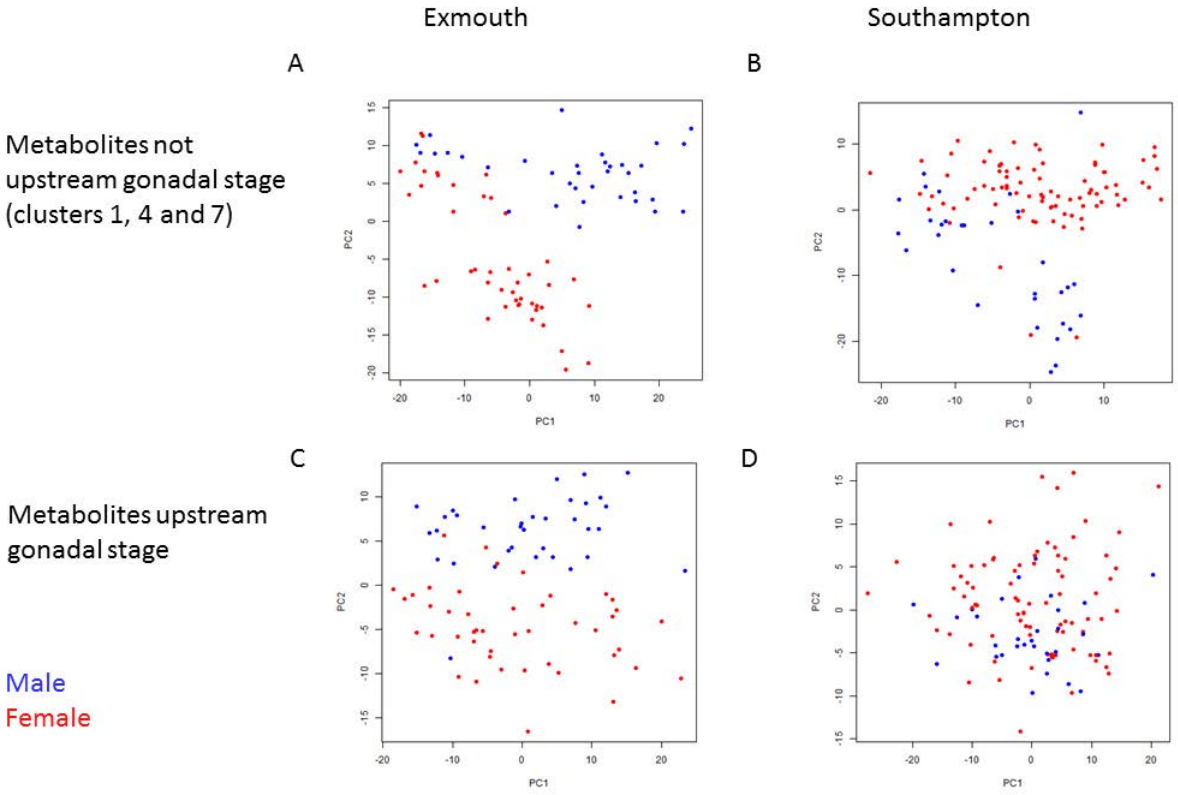


Figure 3.10: Principal component analysis of metabolites not upstream of gonadal stage (A, B) and upstream of gonadal stage (C, D) in the winter months (October to April). Red indicates individual female mussels and blue individual male mussels. The analysis only includes *Mytilus edulis*

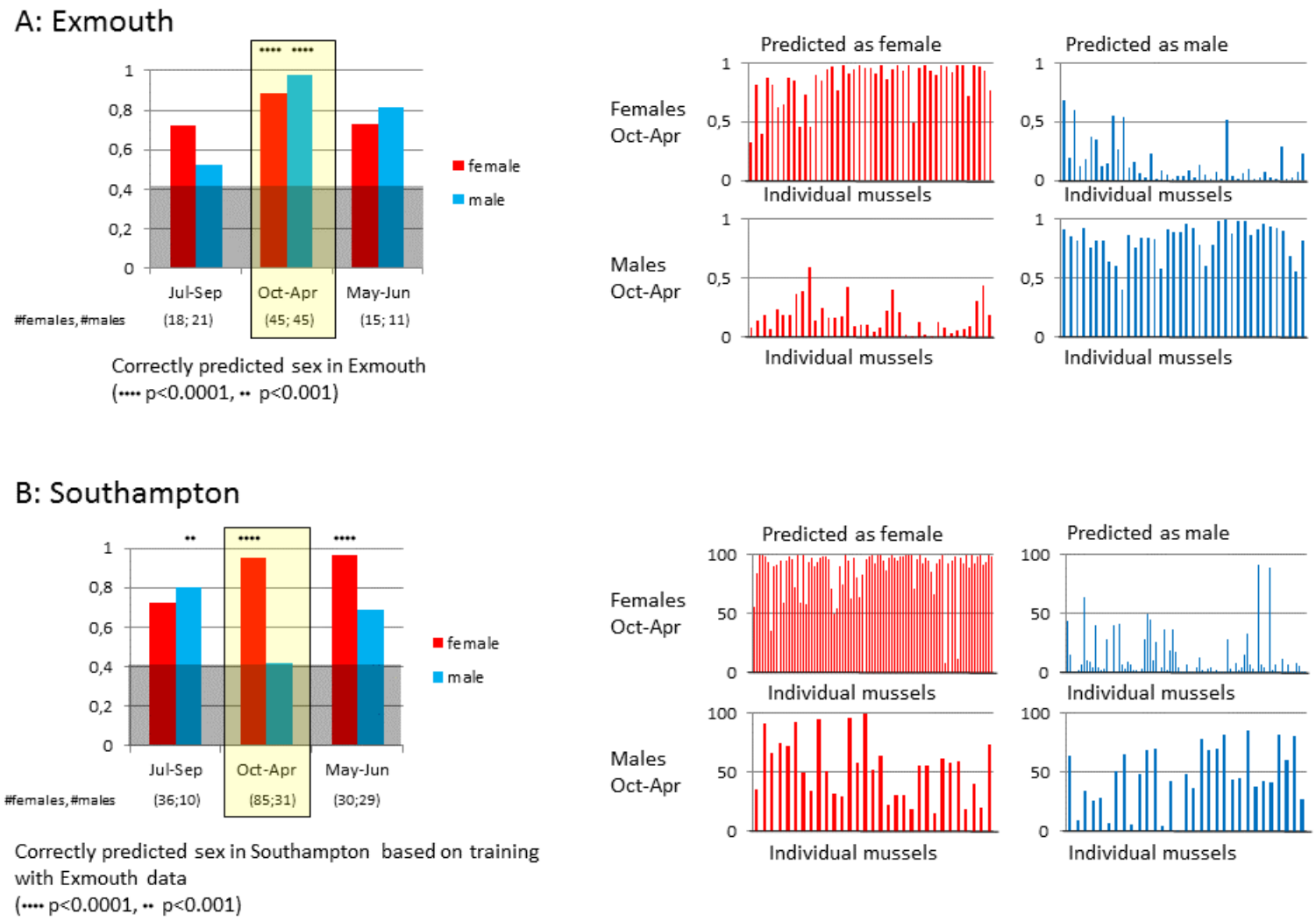


Figure 3.11: A: Prediction accuracy of sex in Exmouth using support vector machine with 25% and 75% cross validation (left), and prediction of sex for individual mussels (right). B: Prediction accuracies of Southampton mussels (left) and sex prediction for individual mussels in Southampton (right). For panel A, Exmouth, only *Mytilus edulis* was used, for panel B, *Mytilus edulis*, *Mytilus galloprovincialis* and their hybrid were used

I then tested whether the markers developed from the Exmouth dataset could be used to predict sex in the Southampton site. Only 41.9% of male mussels were predicted as "male" during the winter months (Figure 3.11B left). Conversely many of the male mussels were predicted as "female" with high probability (Figure 3.11B right, Figure 3.9). Some examples of individual metabolites that contribute to the sex-prediction model are shown in Figure 3.12. As expected, all of them show a differential concentration in males and females in the Exmouth site but not all of them in the Southampton site. We clustered these metabolites in 4 distinct groups on the basis of their sex and site profiles: 1) same sex-specific pattern between the two different sites (Figure 3.12 A: glycine, glutamate, unknown metabolites from cluster 4); 2) Different level of metabolite in the two sites but sex specific differences are in the same direction (Figure 3.12 B); 3) Different level of metabolite in the two sites and metabolites in Southampton males are similar to females in the Exmouth site (Figure 3.12 C: gamma-aminobutyric acid (GABA), ATP, UDP-GlcNAc, ornithine); 4) Different level of metabolite in the two sites and metabolites in Southampton females are similar to males in the Exmouth site (Figure 3.12 D: citrate). Notably, two metabolites that most contribute to sex prediction (GABA and ATP) are sufficient to separate males and females sampled from Exmouth in the winter (overall accuracy 90% with ksvm model) but fail to do so in the Southampton site (model accuracy 70% overall, but 82.4% for females and 25.8% for males) (Figure 3.13).

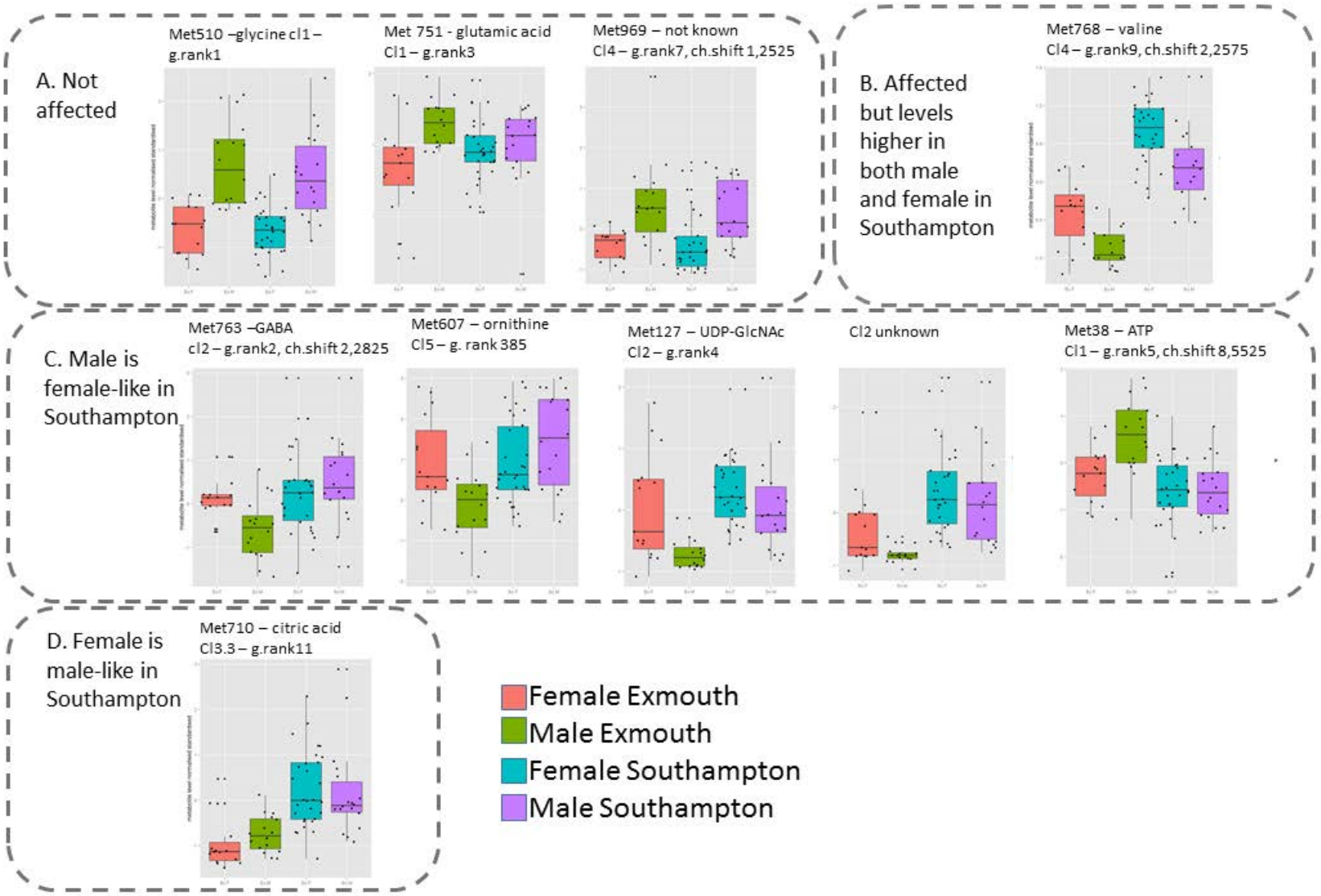


Figure 3.12: Representative examples of top sex-predicting metabolites. A: male and female mussels are not affected by site. B: Both sexes are affected, with increased metabolite levels. C: Metabolite levels in male mussels in Southampton are more similar to levels of female mussels. D. Metabolite levels in female mussels in Southampton are more similar to levels of male mussels. In Exmouth, *Mytilus edulis* was used, for Southampton, *Mytilus edulis*, *Mytilus galloprovincialis* and their hybrid were used.

Table 3.1: Correlation of various environmental and physiological parameters with metabolite cluster medians in female mussels (*Mytilus edulis*)

Parameter	Cl. 1	Cl. 2	Cl. 3.1	Cl. 3.2	Cl. 3.3	Cl. 4	Cl. 5	Cl. 6	Cl. 7
Temperature	-0.79	-0.76	-0.72	-0.81	0.72	-0.55	0.79	0.8	0.67
ADG rate	-0.48	-0.34	-0.17	0.14	0.22	-0.66	0.48	0.19	0.69
Gonadal stage	0.92	0.94	0.77	0.63	-0.87	0.79	-0.92	-0.85	-0.88
Parasite load	0.22	-0.03	0.06	-0.31	0.07	0.28	-0.22	0.11	-0.35

Table 3.2: Correlation of various environmental and physiological parameters with metabolite cluster medians in male mussels (*Mytilus edulis*)

Parameter	Cl. 1	Cl. 2	Cl. 3.1	Cl. 3.2	Cl. 3.3	Cl. 4	Cl. 5	Cl. 6	Cl. 7
Temperature	-0.74	-0.6	-0.51	-0.77	0.8	-0.64	0.77	0.88	0.56
ADG rate	-0.56	-0.25	-0.66	-0.16	0.13	-0.85	0.62	0.59	0.65
Gonadal stage	0.93	0.47	0.73	0.66	-0.71	0.82	-0.94	-0.91	-0.88
Parasite load	-0.08	-0.34	0.34	-0.31	0.38	0.13	-0.01	0.09	-0.26

Table 3.3: Identities for top 20 of sex-differentiating metabolite bins in *Mytilus edulis*. Metabolite bins are grouped based on their assigned metabolite identities form the BML database and literature. Where known, pollution-effects are indicated, as are site-effects form our study

Met.code	Chemical shift	sex-rank	Putative metabolite identity from BML	Identification based on chemical shift from literature	effect in pollution (literature)	Effect in our study
Met506, met509, met510, met514, met503	3.5475, 3.5525, 3.5675, 3.5275, 3.5825	1, 2 ,3, 5, 17	Glycine	3.563 [343]	up in pollution [343]	Not affected
Met751	2.3425	4	Glutamate	2.34 [343]	up in pollution [343]	Not affected
Met38, met39	8.5525, 8.5475	6, 19	-	ATP 8.548 [343], ATP 8.53 [151]	down in pollution [343]	Males in Southampton similar to females (down)
Met763, met764, met765, met766	2.2825, 2.2775, 2.2725, 2.2675	7, 8, 18	-	GABA 2.28 [151]		Males in Southampton similar to females (up)
Met127, met288	7.9475, 6.0075	9, 10	UDP-GlcNAc	7.953 [343]	up in pollution [343]	Males higher in Southampton, but not higher than females
Met969	1.2525	11	-	1.258 [343]	down in pollution	Not affected
Met6	8.8475	12	-			Not affected
Met768	2.2575	13	-	Valine 2.25 [151]		Higher in both males and female
Met 549	3.3525	14	-			
Met710	2.5475	15	Citric acid			Higher in both male and female
Met771	2.2425	20	N-acetylglutamic acid			

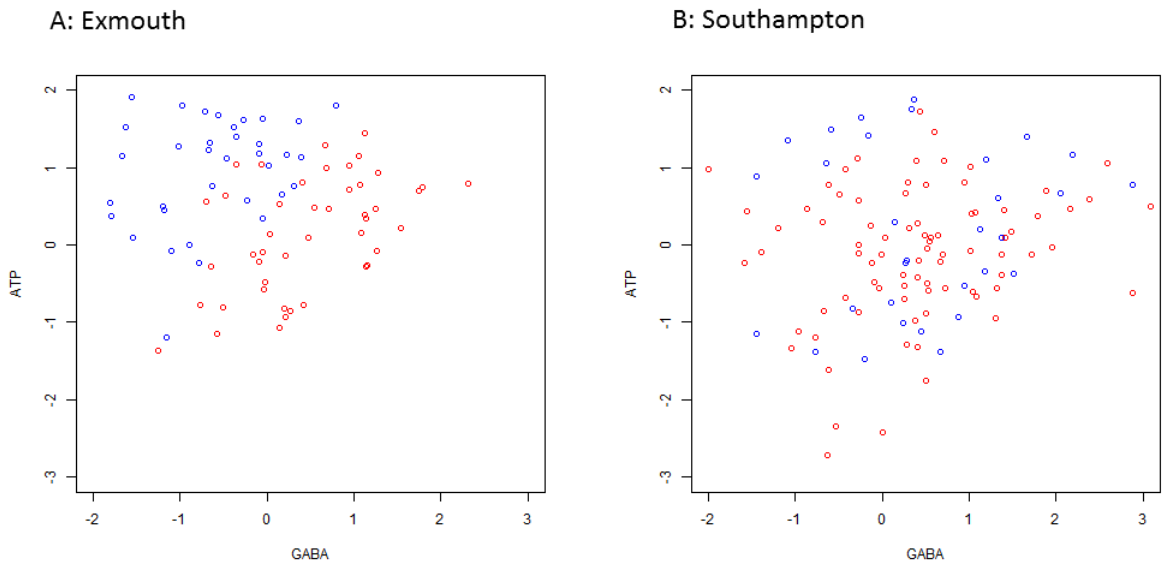


Figure 3.13: ATP and GABA levels separate male and female mussels in Exmouth (A), but not in Southampton (B). Red dots represent female and blue dots male mussels. In Exmouth, *Mytilus edulis* was used, for Southampton, *Mytilus edulis*, *Mytilus galloprovincialis* and their hybrid were used.

Table 3.4: GSEA enrichment for sex-predicting metabolite bins in *Mytilus edulis* for each cluster. Column 3 indicates how many nodes upstream of the gonadal stage was that cluster.

Cluster	FDR	nodes upstream gonadal stage
Cluster 2	0.000	2
Cluster 1	0.108	Not upstream
Cluster 4	0.126	Not upstream
Cluster 3.2	0.457	1
Cluster 5	1.000	Not upstream
Cluster 6	1.000	1
Cluster 7	1.000	Not upstream
Cluster 3.1	1.000	Not in network
Cluster 3.3	1.000	1

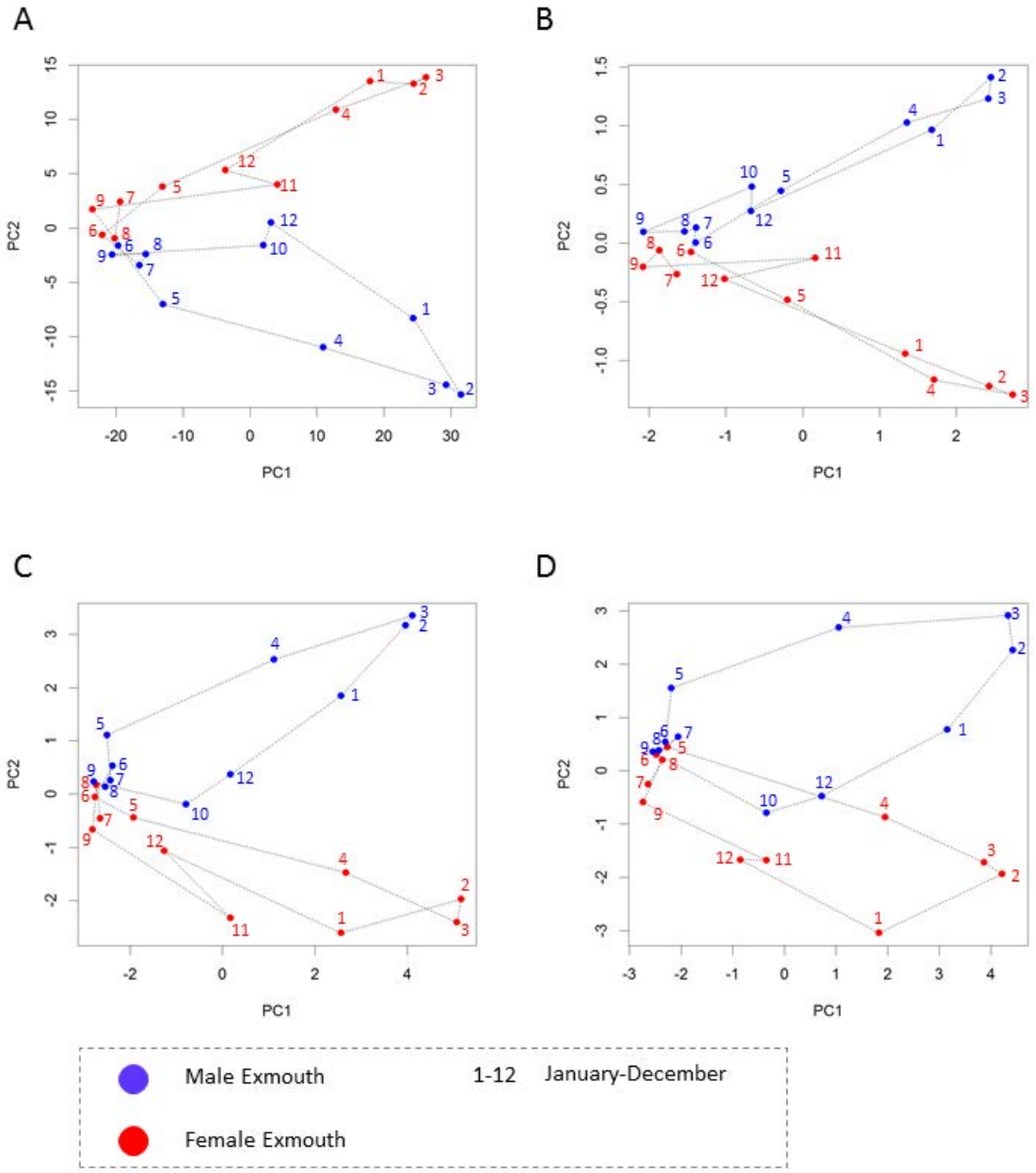


Figure 3.14: Principal component analysis (PCA) describing the annual cycle in the Exmouth site for male (blue) and female (red) *Mytilus edulis*. A: all significant metabolites; B: 9 cluster medians; C: 20 metabolite bins for the clusters (including all identified metabolites plus 5 significant peaks); D: only 15 identified metabolites. Numbers 1-12 represent months from January to December. October is missing for females and November for males due to no samples of relevant sex.

3.6 Discussion

The most important finding of our study is the demonstration that the metabolic state of mussels sampled from the environment correlates with multiple biologically relevant parameters such as sex, life cycle, and gonadal development, and that it is potentially informative of site-specific environmental stressors. Our work provides useful knowledge to formulate hypothesis on the molecular basis of important biological processes such as gonadal development and provide markers to improve environmental monitoring, opening the possibility of revealing the effect of pollution before adverse phenotypic effects are detected.

3.6.1 Are sex-specific metabolites important in sex determination?

The five metabolites that are most predictive of sex were glycine, glutamate, gamma-aminobutyric acid (GABA), ATP and UDP-GlcNAc. Glycine has been previously shown to have higher levels in male mussels [141]. ATP is an energy currency of the cell and as it showed higher dynamics in male mussels and was correlated with gonadal development, it is possible that the sex-specificity is related to the content of ATP in sperm. This is supported by previous studies that relate ATP content with sperm quality [34]. In addition to having different dynamics in male and female mussels, ATP was also affected by site. Interestingly, a previous ^{1}H-NMR metabolomics study in which models predictive of the scope for growth are reported, showed that pentachlorophenol (PCP, used as pesticide, wood preservative), a chemical that uncouples oxidative phosphorylation, increases respiration rate and reduces scope for growth in lab-exposed mussels [140]. Moreover, using ^{1}H-NMR data for adductor muscle for the same mussels as in the current study, the scope for growth was predicted to be lower in Southampton than in Exmouth [140]. ATP concentrations have been shown to be inversely correlated with respiration rate in sea urchin [101], meaning that site-effects might be due to specific chemicals which affect

the oxidative phosphorylation, increase respiration rate and lower ATP concentrations. Unfortunately, we have no data available for the two sites about chemical concentrations. Interestingly, gamma-aminobutyric acid (GABA) was linked to gonadal development in the model represented in Figure 4 and ATP and UDP-GlcNAc were differentially abundant in the Southampton site. Another interesting property of these metabolites is that three of them (glycine, GABA and glutamate) are known neurotransmitters (reviewed in [36],[99], [361]).

In addition to ATP being the energy molecule in the cell and controlling a very large number of biological mechanisms, it also operates a wide range of channels involved in modulating the neuronal synapse and changes in its levels can have significant effects in nerve conduction [255], reviewed in [48]). Since glycine, glutamate, GABA and ATP are all operating various ionotrophic receptors, it is reasonable to hypothesize that they may be involved in the control of sex determination or at least in the development and differentiation of the gonads. There are several lines of evidence in support of this hypothesis. A recent gene expression study in the scallop has revealed that sodium- and chloride depending GABA transporters and sodium- and chloride dependent glycine transporters were over-expressed in the ovary [196]. Importantly, GABA has been shown to be present and functional in bivalves [286, 193]. In oyster, the homolog of glutamic acid decarboxylase (cgCAD), the rate-limiting enzyme for conversion of glutamate into GABA, has been identified and shown to be functional [194]. The GABA transporter (GAT2) is also been found to be functional in oyster [286]. These observations raise the question of what the role of GABA in the reproductive system of molluscs may be. Although there are no data in mussels, we know that in vertebrates, GABA regulates GnRH neurons [342]. Importantly, GnRH plays role in gonadal development in mollusc as well. In the scallop, GnRH-like peptide has been found in the central nervous system where it stimulates spermatogenesis. In vivo studies in scallop have showed that GnRH accelerates spermatogenesis in males while inhibiting oocyte development in females whereby shifting sex balance towards males [229]. GnRH is also involved in ovarian cell proliferation

in abalone [233]. In the absence of a real understanding of sex specification in mussels the possibility that GABA may also regulate GnRH neurons in molluscs represents an exciting hypothesis.

3.6.2 Sex prediction models across geographical sites

We have shown that the metabolic state of male mussels sampled from the Southampton site is similar to females. It is possible that chemicals in the water of the Southampton sampling site or other environmental and health factors, may be responsible for changes in the development of the mussel gonads that our model is detecting. This raises the potential that male individuals with the molecular state of a female in the more polluted site might show the detectable changes underlying complex pathological indicators of health , offering a new tool for environmental biomonitoring. In fact samples derived from the Southampton site show a significantly higher parasite load for *Steinhausia mytilovum* ($p - value = 2.2e - 16$) and bucephalids ($p - value = 0.002976$) (p-values calculated based on data from [31]) which are known to affect gonadal maturation often leading to a reduced ADG rate. Moreover, it has been demonstrated that polluted water downstream of municipal effluents or on shipping routes can alter the male-female ratio or the percentage of intersex mussels [102, 287] or up-regulate vitellogenin [73]. This observation raise the possibility that samples from the Southampton site may be affected in the intersex ratio, although this was not detected by the original histological analysis.

3.6.3 Conclusions

Results of this chapter show the potential of data-driven systems biology approaches using metabolomics for describing normal seasonal cycles of mussels. Instead of concentrating on specific pathways or mechanisms, I show that groups of metabolites with distinct seasonal dynamics are associated with gonadal development. This lead to the development of sex-prediction models, which reveal markers responsible for molecular intersex in the more

polluted site. It is possible that molecular intersex in the polluted site might represent a molecular state just below the exposure threshold where organism phenotypic changes become visible. If this was true, the biomarkers I have identified would be a very useful monitoring tool. I also show the potential of this approach for generating hypothesis – both for physiology and for potential effects of pollution in terms of intersex. The hypotheses are experimentally testable. Firstly, the role of GABA in gonadal development in mussels could be studied, most importantly its relation to GnRH, by injection of GABA, or by GAT interference (similarly as in [286] to study copper accumulation) and levels of GnRH measured. The role of GABA could be also studied in relation to other possible genes or proteins involved in sex determination, for example by transcriptomics or proteomics. Indeed, if this neurotransmitter induces changes related to gonadal development (either GnRH or other genes/proteins), it would also be possible to test which chemicals affect the levels of GABA. This could change the way we view mussels as sentinel species.

CHAPTER 4

DATA-DRIVEN SYSTEMS BIOLOGY APPROACH GIVES INSIGHT INTO A COMPLEX PROCESS OF WATER REMEDIATION

4.1 Contributions

- Jaanika Kronberg-Guzman[1,2,3] analysed the pre-processed microarray data (which was provided by Tim Williams), found differentially expressed genes, performed the exploratory analysis of chemical and gene expression data, integrated the gene expression data with other data in the form of similarity networks, did network modularisation, functional annotation and interpretation. Analysed and interpreted the model created by Alberto Cassese and Marina Vannucci. Wrote the chapter.

- Timothy D. Williams[1] was involved in the microarray experimental design, experimental work and microarray data pre-processing. Provoded the description of the methods that he did (indicated) and mediated the sharing of different types of data of collaborators.

- Albertinka Murk[4] was involved in the mesocosm experiments.

- Erwin Roex[4] was involved in the passive sampler measurements and has written a draft about the chemical data. However, the analysis is non-overlapping and is different than in this thesis.

- Laine Wallace[1] was involved in the microarray experimental work.
- Alberto Cassese[6] developed the Bayesian model on a small number of genes and chemicals and provided the results to be analysed and interpreted in the context of the full system.
- Philipp Antzcak[2] selected the genes and chemicals to be used in the Bayesian model for Alberto Cassese.
- Edwin Foekema[4] was the leader of the passive sampler measurements.
- Marina Vannucci[6] was involved in the development of the Bayesian model, being the supervisor of Alberto Cassese at the time.
- Ron van der Oost[5] was involved in the design of the whole study (mesocosms, passive samplers and microarrays).
- Kevin Chipman[1] was involved in the design of the whole study.
- Francesco Falciani[1,2] guided the systems biology side of the study, with discussions about methodology. He also mediated the collaboration with Alberto Cassese and Marina Vannucci.

Affiliations

1. School of Biosciences, University of Birmingham, UK
2. Institute of Integrative Biology, University of Liverpool, UK
3. Institute of Genomics, University of Tartu, Estonia
4. Wageningen University, Netherlands
5. Waternet, Netherlands
6. Rice University, USA

4.2 Introduction

Environmental pollution, linked to industry, intense farming and to a growing urban populations pose a challenge to secure clean water resources [337, 336]. Pesticides from agriculture [110, 169], personal care products [202], medicines from both domestic effluent and hospitals [223, 55, 24] and other chemicals, such as flame retardants, plasticisers [218, 163] and industrial chemicals enter our ecosystem despite extensive use of waste water treatment plants (WWTP). The European Water Framework directive [77] aimed to oblige European countries to achieve cleaner surface waters, however, the improvement of water has been modest: In 2015, 47% of all waters still did not meet the required good environmental status (reviewed in [338]).

Current wastewater treatment plants do not remove all chemicals efficiently [221]. The REACH programme estimates that there are at least 30000 different chemicals currently in use in domestic and industrial settings in the EU [94], many of these will enter WWTP. It has been shown that level of removal depends on the type of technology used [221, 334] and also the diversity of chemicals [83].

One possibility for achieving the goal set by the Water Framework Directive (WFD) is to improve the quality of water that is discharged into surface waters from waste water treatment plants. Some countries, such as Switzerland, have already taken action by legislation (Swiss Water Protection Act, 2016), requiring selected WWTPs to be upgraded to use either activated carbon or ozonation (reviewed in [35]). Membrane filtration has also been shown as effective for reducing concentrations of some chemicals of emerging concern [175]. However, such advanced purification methods can be applied to selected WWTPs in a limited number of countries.

4.2.1 Alternative additional remediation

"Effluent polishing" is a biological remediation method that has been recommended for improving water quality before releasing it from the WWTP [166]. Effluent polishing

techniques can use constructed wetlands before water enters the surface waters. This type of remediation has been used for decades since it is a low-cost natural option [52, 166]. One effluent polishing method is the Waterharmonica concept (www.waterharmonica.nl) that has been used in the Netherlands. Constructed wetlands have been studied in respect to various pollutants, including phosphorus, nitrogen and ammonia, as well as more recent pollutants, such as emerging contaminants, personal care products and nanoparticles [52]. However, most studies so far have focussed on one chemical or type of chemicals. For example, constructed wetlands have been shown to be effective in reducing phenols and endocrine disrupting compounds [70, 316, 240, 68].

The use of biological markers and especially omics measurements have the potential to show effects of chemicals that are not (yet) measured and also make it possible to use advanced machine learning methods to study the effects or predict components of chemical mixtures [104, 80]. The stickleback (*Gasterosteus aculeatis*) has been described as a particularly suitable species for environmental studies due to being present in all Europe, Northern Asia and America [158] and stickleback biomarkers, especially the male-specific glue-protein spiggin [131], have been demonstrated to be suitable for biomonitoring [270]. A cDNA microarray has also been developed for stickleback [109] and has been used to show the effects of various chemicals [109, 159, 274].

In a laboratory-based model of wastewater remediation, gene expression changes in *Daphnia magna* have been modelled taking into account the changing chemical concentrations [53]. In real wastewater treatment plants, trancriptomics has been used to study the effects of the WWTP upgrades [213]. In another study, gene expression signatures of fathead minnow exposed to different effluent gradients were studied [28]. However, to the best of my knowledge, gene expression has not been used to analyse the consecutive stages of a whole system of remediation together with measured chemical concentrations and traditional toxicity tests.

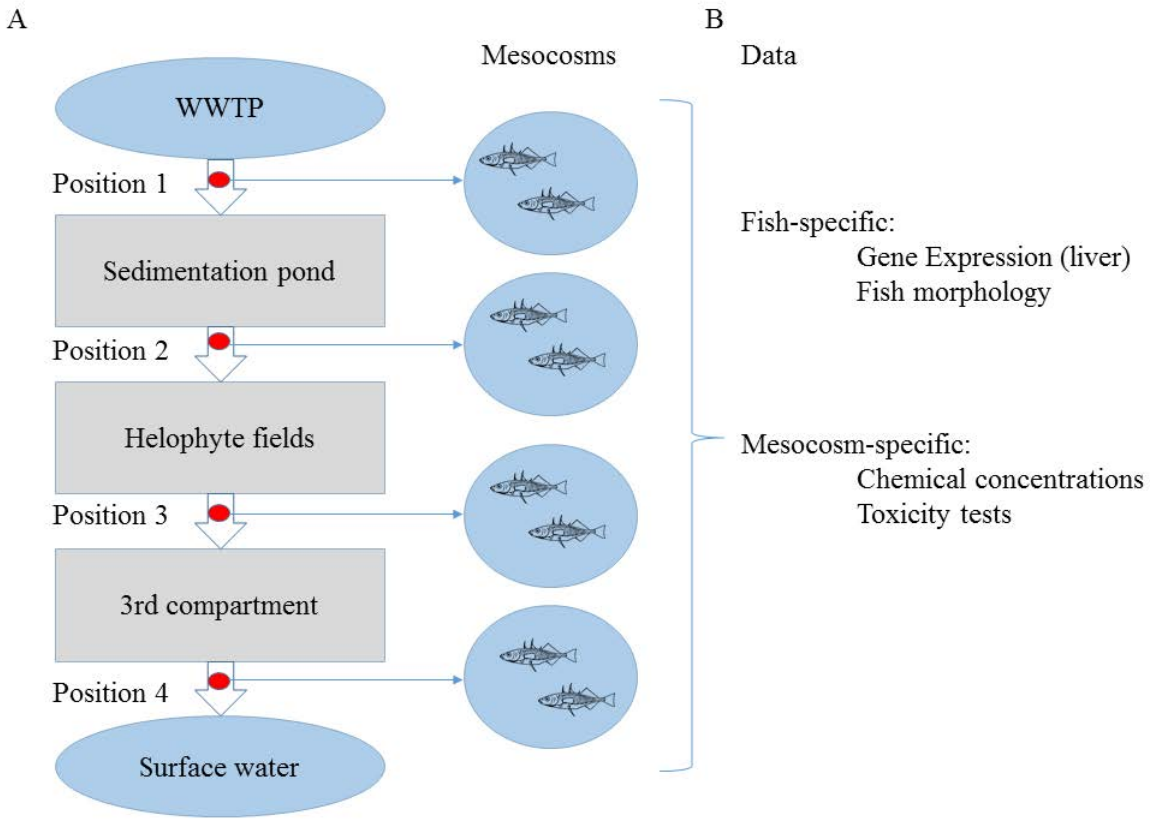


Figure 4.1: Overiew of the study. Panel A shows the general overview of the Waterharmonica remediation after the WWTP and water flow from 4 positions into mesocosms. Panel B shows the types of data that were generated for each fish and mesocosm.

4.2.2 Aims of this chapter

In this chapter, I use advanced systems biology integrating chemical concentrations from passive samplers, traditional toxicity tests, stickleback morphology and stickleback liver gene expression to understand the effects of the Waterharmonica effluent polishing system on sticklebacks.

The Waterharmonica wastewater remediation system was used in 3 different sites in the Netherlands, following the conventional wastewater treatment. The sites were Grou (Figure 4.2 A), Hapert (Figure 4.3 A) and Land van Cuijk (Figure 4.4 A) and the wastewater treatment type and capacity are described in Table 4.1. In addition to the conventional treatment, water was remediated using a type of constructed wetland, Waterharmonica (www.waterharmonica.nl). Waterharmonica consists of three consecutive stages of remediation: a sedimentation pond, helophyte fields and a 3rd compartment which can vary in different sites (Table 4.1, Figure 4.2 B, Figure 4.3 B and Figure 4.4 B).

In our study, treated wastewater entered the Waterharmonica and was pumped to mesocosms from 4 sampling positions as indicated in Figure 4.1 A. In each mesocosm, sticklebacks were exposed to the continuous water flow from the sampling position during a period of 1 year. For each mesocosm, chemical concentrations were measured and traditional toxicity tests performed. For every fish, morphology and liver gene expression were measured (Figure 4.1 B).

First, the Waterharmonica remediation system is evaluated on water quality by chemical concentrations and risks. Secondly, stickleback gene expression across the sequential stages of remediation is analysed. The ultimate aim is to understand how gene expression, traditional toxicity tests and morphological measurements of sticklebacks can help inferring the improvement in water quality within the different stages of purification.

The analysis has revealed that the approach of using systems biology for the integration of gene expression with chemical concentrations and morphological endpoints can provide biologically meaningful hypotheses about how chemicals might affect stickleback. More specifically, the demonstration that through gene expression, chemical concentrations are

linked to stickleback growth provide an exciting hypothesis. Additionally, I demonstrate that many chemicals with concentrations below the predicted no effect concentration (PNEC) are correlated with gene expression and decrease during remediation, suggesting that these chemicals have effect during chronic exposure. The analysis also shows that many of the chemicals correlated with gene expression decrease during remediation.

Table 4.1: Characterisation of different sites by their wastewater treatment plant and Waterharmonica processes

	Grou	**Hapert**	**Land van Cuijk**
Type of purification process	Carrousel	Oxidation ditch	Activated sludge
Capacity (inhabitants equivalent)	25000	71000	175000
Additional remediation	Waterharmonica	Waterharmonica	Waterharmonica
WH stage 1	Sedimentation pond	Sedimentation pond	Sedimentation pond
WH stage 2	Helophyte fields	Helophyte fields	Helophyte fields
WH stage 3	Ecological lagoon	Wetland forest	Discharge ditch

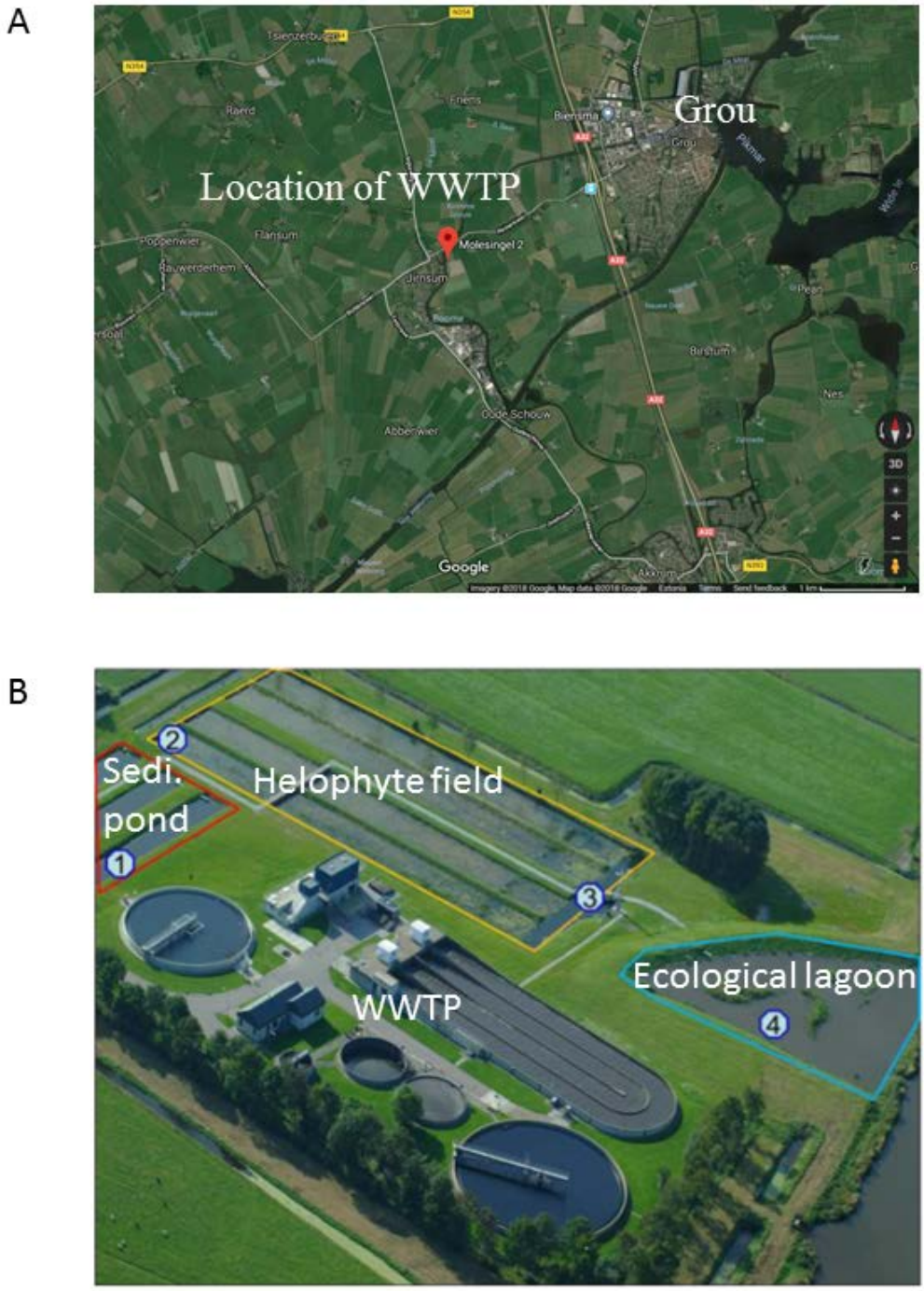


Figure 4.2: Overview of the Grou site. Panel A shows the location of the WWTP and panel B shows the layout of the Waterharmonica remediation system at this site. Numbers 1-4 on different remediation stages indicate positions from where water was flowing to a mesocosm where fish of this study were living. Panel A was generated with Google maps. The image on Panel B was adapted from the report of Waternet

A

B

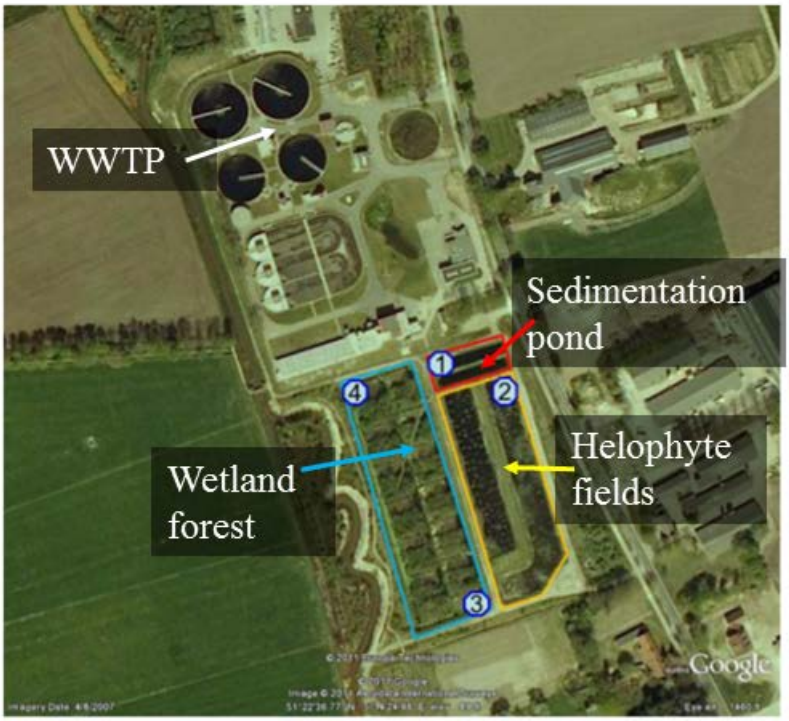


Figure 4.3: Overview of the Hapert site. Panel A shows the location of the WWTP and panel B shows the layout of the Waterharmonica remediation system at this site. Numbers 1-4 on different remediation stages indicate positions from where water was flowing to a mesocosm where fish of this study were living. Panel A was generated with Google maps. The image on Panel B was adapted from the report of Waternet

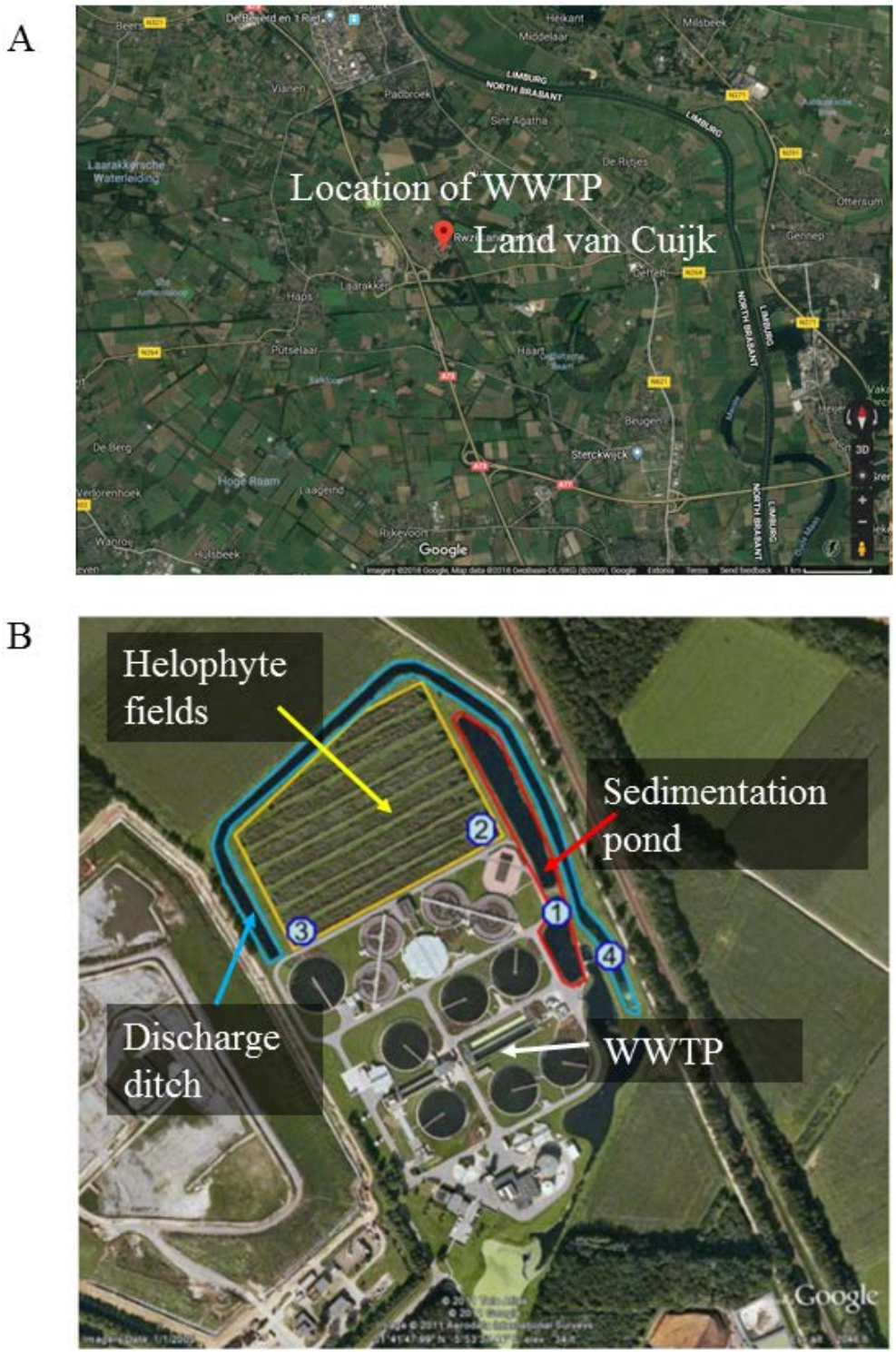


Figure 4.4: Overview of the Land van Cuijk site. Panel A shows the location of the WWTP and panel B shows the layout of the Waterharmonica remediation system at this site. Numbers 1-4 on different remediation stages indicate positions from where water was flowing to a mesocosm where fish of this study were living. Panel A was generated with Google maps. The image on Panel B was adapted from the report of Waternet

4.3 Methods

4.3.1 General overview of all methodology

The overview of the methodology is shown on Figure 4.5. Figure 4.5 A describes the mesocosm set-up and data generation. Briefly, data was generated from 4 mesocosms connected to 4 water sources along the Waterharmonica remediation system in 3 different sites. In each of the mesocosms in 3 sites, organic chemicals were measured by passive samplers. Traditional toxicity tests were also done using water from the mesocosms. For each individual stickleback, morphology parameters were determined and liver gene expression measured.

4.3.2 Mesocosm set-up (Ron van der Oost)

In each of the sites, Waterharmonica additional remediation was used after the Waste Water Treatment Plant and before water was released into the surface water. The Waterharmonica consists of three consecutive stages, first and second were same between all sites and a third one differed. The first stage was a sedimentation pond containing algae and Daphnia. The second stages were helophyte fields, where reeds are the main organisms contributing to the remediation. The third stage was ecological lagoon in Grou, wetland forest in Hapert and discharge ditch in Land van Cuijk. The sampling positions to be investigated were chosen to be 1)after the WWTP, 2)after sedimentation pond, 3)after helophyte fields and 4)after the 3rd compartment. From each sampling position, water was continuously pumped (60 l / h) into 4 mesocosms, made of polyester, with the following measurements: 2m diameter, 1m depth, volume 2.5 m^3). The mesocosms were sheltered with cages or ropes, to avoid public from interfering with the experiment. In each of the mesocosms, there were aquatic plants of the species *Myriophylum spicatum* for providing shelter and nesting material for sticklebacks.

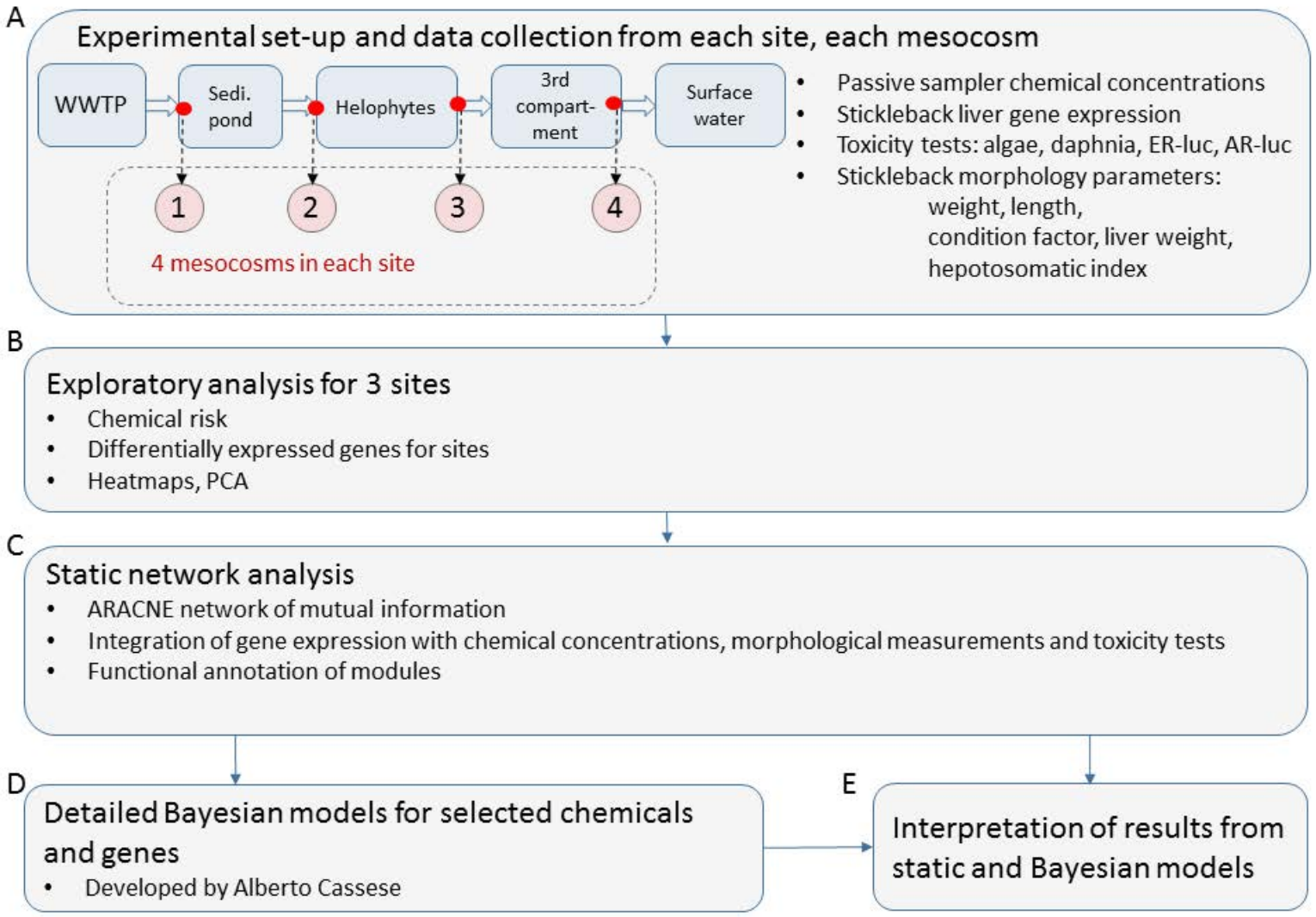


Figure 4.5: Overview of the methodology. Part A shows the additional remediation stages after the Wastewater Treatment Plant and the direction of water flow. Locations of water sources for mesocosms are also indicated. All types of data generated are outlined. The data was first used for exploratory analysis, as described in part B. After exploratory analysis, a full systems biology integration of all data was performed (part C). From a subset of chemicals and genes, Bayesian model was developed by Alberto Cassese (D) and interpreted in the context of the static model (E).

4.3.3 Exposures (Ron van der Oost, methods description adapted from of Tim Williams)

The F0 generation consisted of adult sticklebacks migrating from the sea to fresh water were caught in spring 2010 from Den Helder, Netherlands. These sticklebacks, after being acclimated to freshwater at IMARES in Dan Helder, were placed in mesocosms (5 male and 10 female in each mesocosm). In July 2010, after removing the parents, 80 offspring from the F1 generation were kept in each mesocosm. The sticklebacks were counted in October 2010 and mesocosms were cleaned. In April 2011, sticklebacks were sorted into "male", "female" and "unknown" based on their appearance. 10 male and 10 female sticklebacks were killed and morphological parameters (weight, length, liver weight) determined. Livers were placed in in *RNA later* (Ambion) for transcriptomics.

4.3.4 RNA preparation (Tim Williams, methods description adapted from Tim Williams)

Stickleback livers (< 10mg wet weight) that had been preserved in *RNAlater* were homogenised and extracted with Trizol (Sigma-Aldrich, Dorset, UK) and chloroform, followed by isopropanol precipitation. RNA was resuspended and purified using a Qiagen RNEasy Mini Kit (Qiagen, Venlo, The Netherlands). On-column DNAase I ligation was used to remove genomic DNA. NanoDrop (NanoDrop Products, Wilmington, DE, USA) procedure was used for assessing the RNA quality: after resuspensing the RNA in nuclease-free water, absorbances were measured for 230nm, 260nm and 280nm using the NanoDrop 1ul spectrophotometer. The thresholds for RNA quality were A260/A280 > 1.7 and A260/A230 > 1.5 and samples which did not pass these thresholds were re-purified using ethanol/acetate precipitation. RNA samples were shipped to the University of Birmingham on dry ice. In Birmingham, the quality was further evaluated by the Agilent Bioanalyser 2100 with Eukaryote Total RNA Nano chip (Agilent, Wokingham, Berkshire, UK). A quality threshold of RIN>7 was used and samples not passing this threshold were excluded from further processing.

4.3.5 Stickleback sequences and annotation (Tim Williams, adapted from Tim Williams)

Predicted stickleback cDNA sequences were downloaded from Ensembl [150] and stickleback Unigene [64] sequences from Genbank. To identify the overlap of Unigenes with Ensembl predicted cDNAs, these sets of sequences were compared using MEGABLAST [217]. When there was a match between a pair of Unigene and Ensembl sequences, the Ensembl sequence was kept due to a higher probability of being a full-length sequence. When the Unigene sequence did not match any Ensembl sequence, the Unigene was kept. The Blast threshold for a match was $P < 1E-06$. There were 29118 predicted Ensembl sequences and 6425 Unigene sequences did not match any of them, giving a total of 35538 sequences. The annotation of the sequences was done by several resources. In addition to the Stickleback genome annotations, annotations were predicted based on orthologs by using Biomart [164]. Secondly, BLASTx was used, searching against Genbank and Swissprot databases. Finally, BLAST2GO [59] was used to match orthologs from other species and for GO categories [62] and KEGG pathways [62].

4.3.6 Microarrays (Tim Williams, methods description adapted from Tim Williams)

The 35538 sequences resulting from the combination of Unigene [64] and Ensembl [150] stickleback sequences were used for designing the Agilent array, using the Agilent's EArray design algorithm. Where possible, 3x60-mer oligonucleotides were designed for each sequence. The resulting array design, Agilent 027680 consisted of 2x105000 probe arrays on each slide. First, an optimization experiments was conducted using the F0 sticklebacks for selecting 15000 probed to be used in the microarray in the main experiment. The optimization array was done using pooled RNA from mixed F0 males and females from a reference site and Grou, Hapert and Land van Cuijk. The microarray experiment was done using the standard Agilent Protocol (Agilent Technologies, Santa Clara, CA) and hybridized microarrays scanned with Agilent G2565BA microarray scanner (Agilent Tech-

nologies, Santa Clara, CA). From the 3 probes of every sequence, the best one with higher fluorescence spot was selected. For these, additional criteria were applied: their intensity had to be 2 standard deviations higher than the background score ($(F > B + 2SD)$). As the number of resulting probes was lower than 15000, additional probes were added which are known be protein-coding which were marginally below the initial threshold. Using these 15000 probes, the 8x15000 microarray was designed using Agilent's EArray (Design ID: 029767). This is the design which was used for generating the data in this thesis.

4.3.7 Microarray pre-processing (Tim Williams, methods description adapted from Tim Williams)

Data was Quantile normalized using the Genespring GX software (v 11.5 or 7.3) (Agilent). Spots with fluorescence close to the background level (base/proportional score intensity<19) were removed, as were spots flagged as 'bad' or 'marginal' by the scanner software.

4.3.8 Passive sampler measurements: (Ron van der Oost, Edwin Foekema)

Passive sampler measurements and used methods are described in an unpublished report by Roex *et al.* [269]. Briefly, large number of micropollutants were measured: lipophilic compounds (PAHs, PCBs, organochloride compounds, phthalates and synthetic musks), alkylated phosphates, plant protection products and Pharmaceuticals and Personal Care Products. The samplers, after being in the mesocosms of various stages of the Waterharmonica for a year, for the same duration as sticklebacks, were rinsed with local surface water and a scouring pad. The samples were stored at -20 until further processing. The chemicals were measured with GC-MS, LC-MS and LC-MSMS. The concentrations in the passive samplers were back-calculated to aqueous concentrations and certainty of measurements calculated [269], based on being detected at least 3 times. Detailed statistical analysis of the chemical concentrations has been described in an unpublished report [269]

by the project partners of Waternet (www.waternet.nl).

4.3.9 Chemical analysis (Jaanika Kronberg-Guzman)

The aim of chemical analysis in this thesis is to provide context for gene expression data interpretation. Full statistical analysis of data from passive samplers is written in an unpublished report [269] by collaborators.

For each chemical detected by passive sampler, annual means were calculated per mesocosm per site. Chemical risks were calculated by the following formula: concentration/PNEC, where PNEC is the Predicted No-Effect Concentration. PNEC values were provided for each chemical by Ron van der Oost. For the heatmap and clustering of chemicals, distances between chemicals were calculated as $dist = (1 - (cor(t(x)))$ and for clustering, complete linkage was used. Heatmap was made in gplots [341] package in R [260].

4.3.10 Gene expression exploratory analysis (Jaanika Kronberg-Guzman)

Differentially expressed genes between mesocosms were found in each site and sex separately using SAMR [312] with FDR 0.01. Although blank control mesocosms existed, they were not used in the gene expression analysis as the fish were fed different diet. Mesocosm 4 was not used in this analysis because there were missing samples in this mesocosm in some sites and also because of a possibility of flood into mesocosm 4 from mesocosm 1. However, this mesocosm is shown in the heatmap and PCA visualisations. Significant genes for every site and sex were visualised with heatmaps. These heatmaps were also used for outlier removal (Cuijk.exposed.4.F.4, Hapert.exposed.3.F.9 and Hapert.exposed.4.F.3, Grou.exposed.3.F.3 removed). Principal component analysis was used to visualise the overall dynamics of the process. For this, R command prcomp from the Stats package [260] was used.

4.3.11 Network analysis (Jaanika Kronberg-Guzman)

ARACNE [212] was used for network reconstruction, using all differentially expressed genes from all sites, separately for male and female sticklebacks. In addition to differentially expressed genes, physiological parameters were also used. Mutual information thresholds were 0.171 and 0.186 for the networks of female and male sticklebacks, corresponding to p-value of $10e-7$. DPI 0.15 was applied. Chemical concentrations and toxicity tests were added by calculating Spearman correlations between gene expression (medians for every pond) and these mesocosm-specific measurements. For finding out which correlation threshold to use, p-values were calculated to correspond to FDR 0.05, with the following formula: FDR/(number of genes x number of chemicals). For female sticklebacks, the p-value corresponding to FDR 0.05 was found to be $7.42e-08$ and for male sticklebacks, $3.08e-08$. To find the correlation corresponding to each p-value, the correlations were calculated for reshuffled data: for every chemical and gene pair, the gene expression and chemical concentration vectors were sampled without replacement and correlations calculated between these random vectors. For every p-value, ks.test was performed to calculate whether it is different from the distribution of p-values from the random data. For male, for p-value $3.08e-08$, the correlation cut-off were 0.89 and -0.89. For female, for p-value $7.42e-08$ the correlation cut-off were 0.9 and -0.9. Chemical-gene correlations that were significant according to these thresholds were saved as .sif file (node - node correlation). Both network of gene-gene and gene-chemical were read into Cytoscape [285]. Instead of using defined pathways (with ortholog mapping to human of zebrafish pathways) for the functional annotation of the network, we have decided to use community clustering with the aim of dividing the network into distinct gene sets that have more edges between the set than with other nodes outside the set. For modularisation, GLay algorithm [304] from the ClusterMaker package [224] was used. The modularisation was done again for a second time for larger modules. All modules were saved as node lists ready for further analysis.

4.3.12 Network module enrichment (Jaanika Kronberg-Guzman)

Using a Gene Ontology [62] and KEGG [235] term file provided by Tim Williams (using Blast2GO [59] program to assign GO and KEGG terms to stickleback genes by mapping the sequences to paralogues in other organisms), custom list enrichment script was written for counting the presence of each KEGG and GO term in a list and calculating the EASE score [147] using 4 variables: count of term x in list A, count of term x in the whole gene set, size of list A, size of whole gene set. EASE score is a modified Fischer test that is fore example used in the DAVID database and software [149].

4.3.13 Bayesian model (Philipp Antzcak, Marina Vannucci, Alberto Cassese, Jaanika Kronberg-Guzman)

This methods section describes work done by Philipp Antzcak, Alberto Cassese and Marina Vannucci. As part of this thesis, I have interpreted their results biologically and in the context of the static networks shown in previous paragraphs of this thesis.

The Bayesian model was ran by Alberto Cassese using a previously published method [53]. In this framework, expression level of a gene in each pond is a sum of mean gene expression of the same gene in the previous pond plus the gene-pond specific difference. Variable selection for differentially expressed genes also uses prior probability of change that takes into account the chemical concentrations. Chemicals to use were selected as representatives of each cluster of chemicals with additional condition that in CTD [72, 71], there is information available about this chemical. The genes were selected as targets of selected chemicals, which have been shown to be modulated in response to environmental chemicals. The selection of chemicals and genes was done by Philipp Antzcak. Alberto Cassese developed the model and made Figure 4.23 and Figure 4.24. I have generated Table 4.11 based on his data and interpreted the results in the context of the static network.

4.4 Results

4.4.1 Changes in chemical concentrations in the Waterharmonica effluent polishing system follow complex dynamics

The Waterharmonica effluent polishing system is composed of three sequential stages: a sedimentation pond, helophyte fields and a third compartment that can vary between sites (ecological lagoon in Grou, wetland forest in Hapert and discharge ditch in Land van Cuijk). In order to understand the efficacy of the remediation process, the changes in concentration of various organic micropollutants measured by using passive samplers were analysed.

In order to facilitate the interpretation of the dynamics of chemical concentrations over the different purification stages, cluster analysis was used to identify groups of chemicals with similar concentration profiles across the different purification stages. Visual inspection of the resulting dendrogram (Figure 4.6) shows that three sites have different chemical profiles. More precisely, three clusters could be identified. Interestingly, while many chemicals show a desired monotonic decrease in concentration moving from effluent to the final purification stage, some of the chemicals show a transient increase in concentration.

In order to study this phenomenon more in depth, an in-depth analysis of the chemical profiles was performed. First, the number of chemicals showing a desired monotonic decrease along the purification stages were determined. The dynamics of chemicals in each large cluster were visualised separately for each site (Figure 4.7, Figure 4.8, Figure 4.9).

These separate clusters of each site were used to categorise the chemicals into two groups: with "expected" (where chemical concentration decreases) and "unexpected" (chemical concentrations does not decrease) dynamics and are summarised in Table 4.2. Cluster 3, which represents chemicals with the highest concentrations in the Hapert site, has the largest number of chemicals with decreasing concentrations during the purifica-

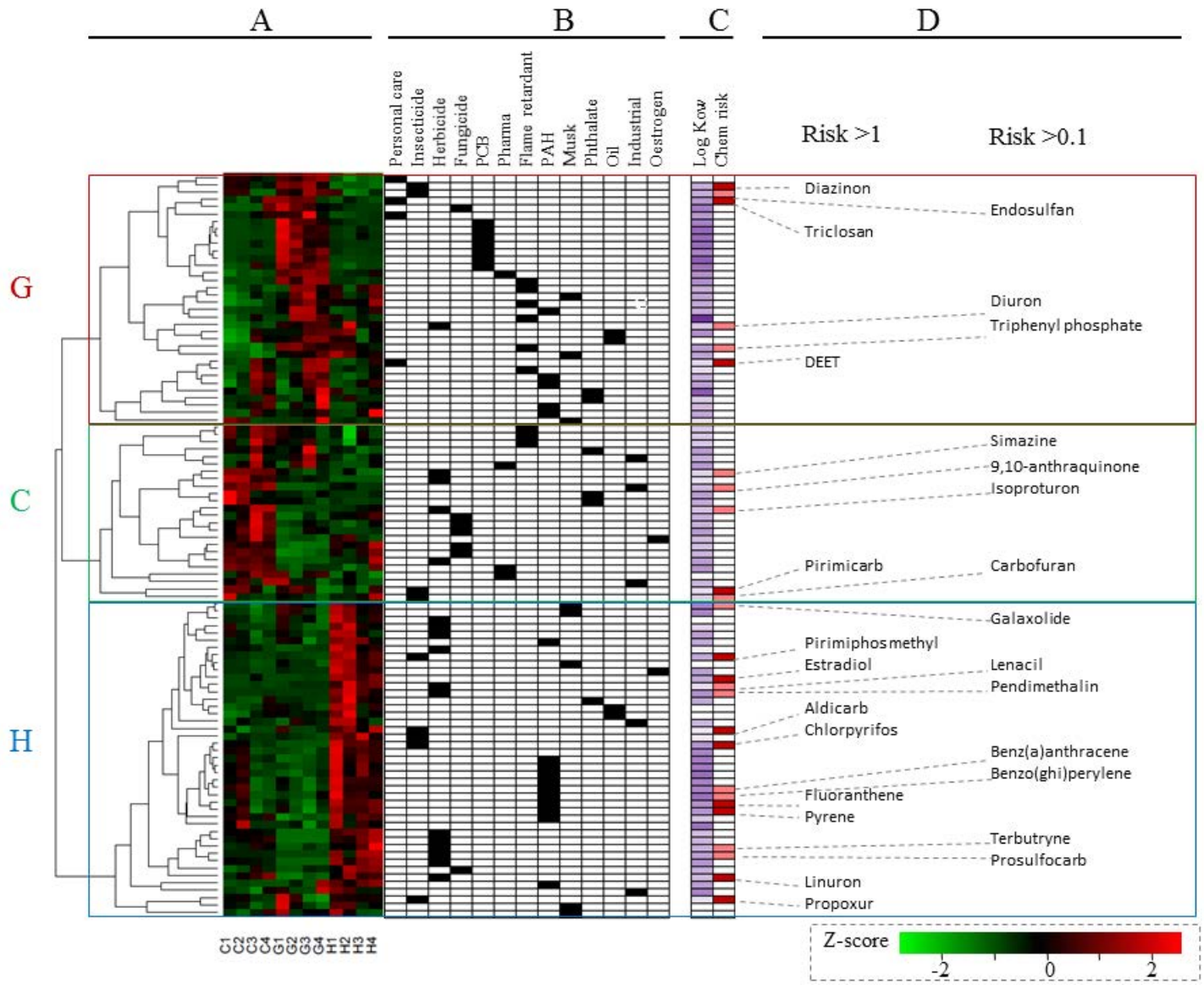


Figure 4.6: A. Clustering and heatmap visualisation of all chemicals measured with certainty in all sites (C1-C4 represent the average chemical concentration in each of the Land van Cuijk mesocosms, G1-G4 in Grou and H1-H4 in Hapert). Green represents low chemical concentration and red high chemical concentration. B. Chemical classes for each chemical where black represent belonging to a certain chemical class (zoomable heatmap with single chemical names is available in the electronic appendix). C. Chemical risk shown as red for risk > 1 ("high risk") and pink for risk > 0.1 ("medium risk"). D. Chemicals with high and medium risk in each cluster.

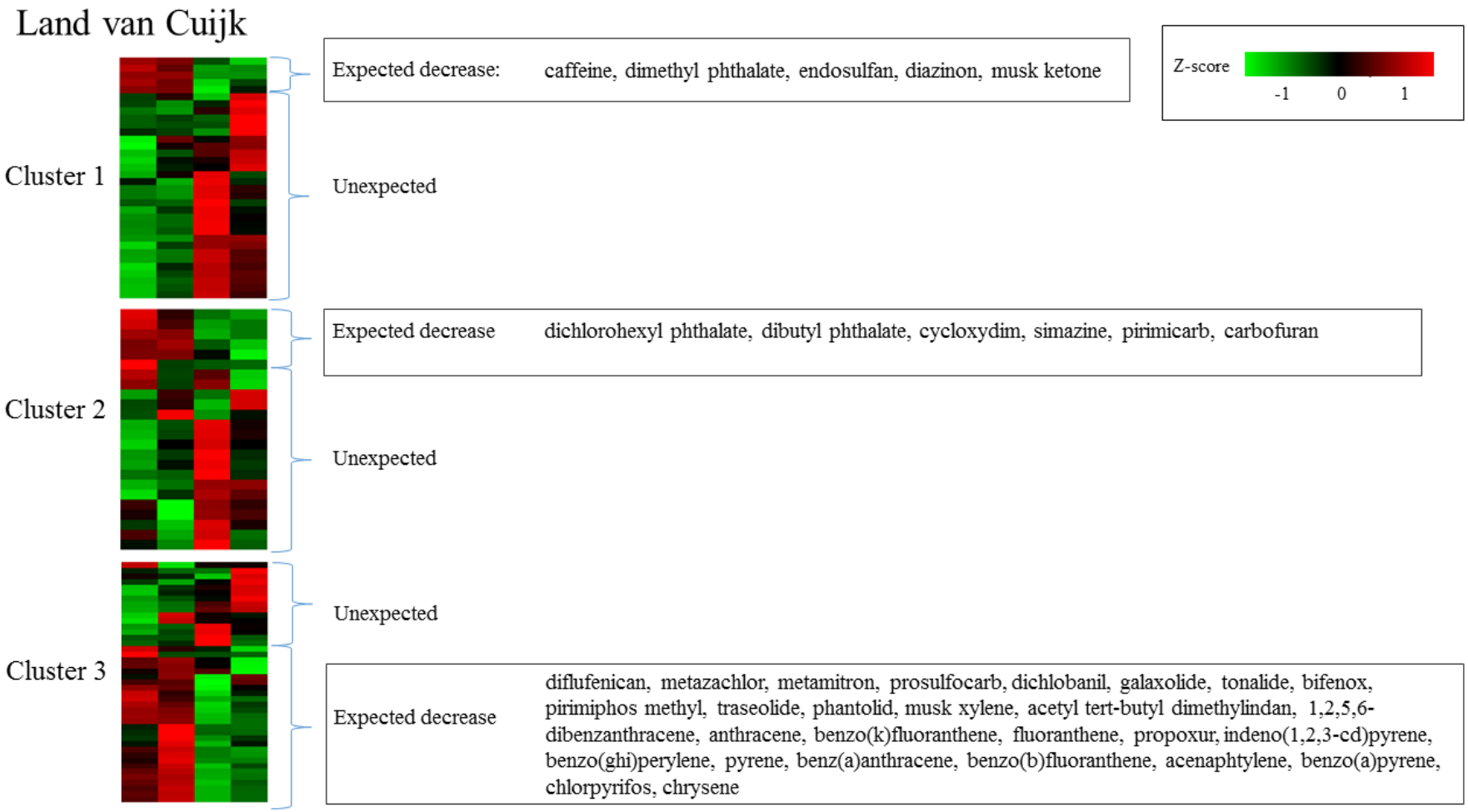


Figure 4.7: Heatmap visualisation of every cluster in Land van Cuijk as found in overall cluster analysis (Figure 4.6). Red represents high chemical concentration and green low concentration. Chemicals in clusters where concentrations are decreasing during remediation stages are outlined

Figure 4.8: Heatmap visualisation of every cluster in Grou as found in overall cluster analysis (Figure 4.6). Red represents high chemical concentration and green low concentration. Chemicals in clusters where concentrations are decreasing during remediation stages are outlined

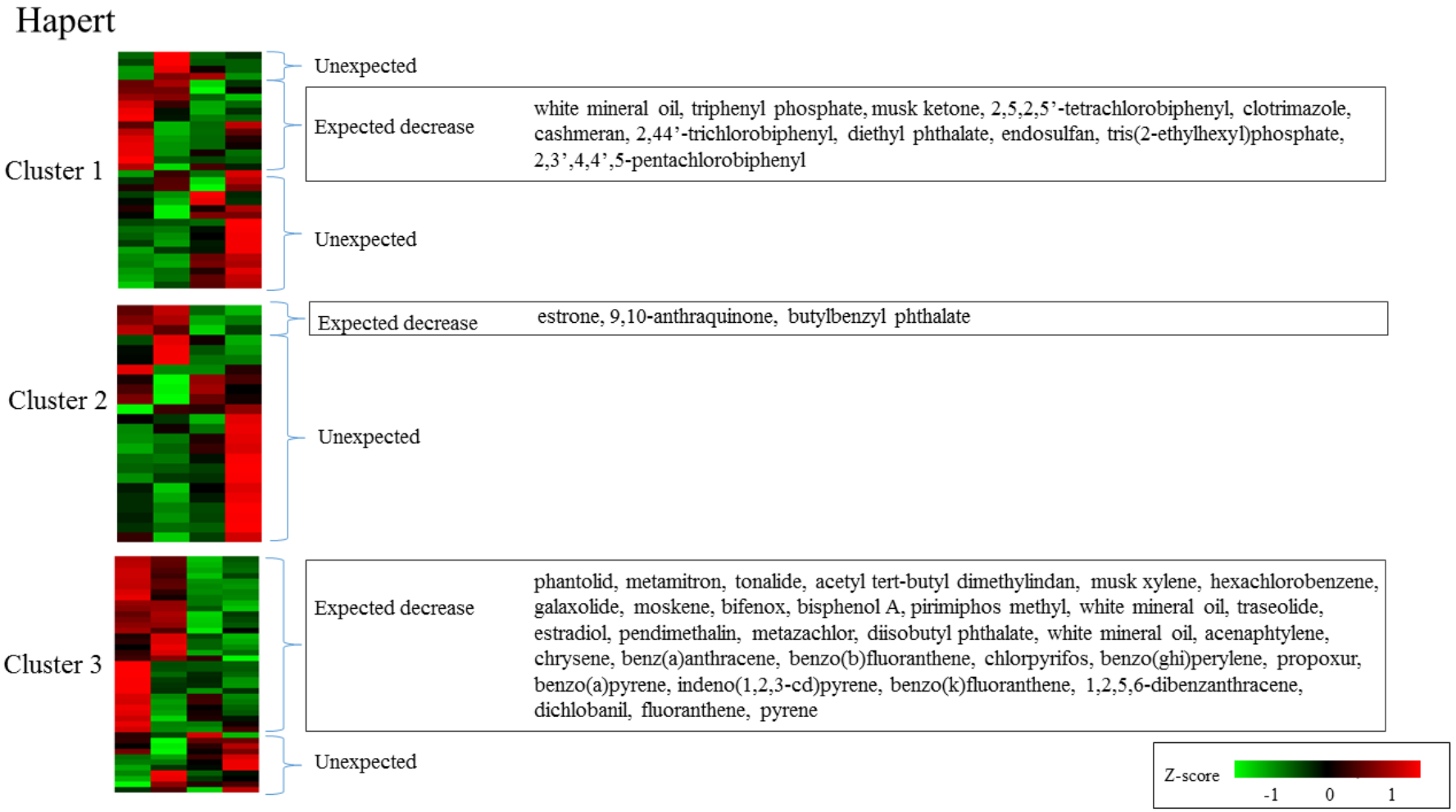


Figure 4.9: Heatmap visualisation of every cluster in Hapert as found in overall cluster analysis (Figure 4.6). Red represents high chemical concentration and green low concentration. Chemicals in clusters where concentrations are decreasing during remediation stages are outlined

tion process (65.1% in Cuijk, 44.2% in Grou and 74.4% in Hapert decrease in cluster 3, Table 4.2). Chemicals that decrease in cluster 3 include polycyclic aromatic hydrocarbons (PAHs), musks, herbicides, industrial chemicals and insecticides (Figure 4.7, Figure 4.8, Figure 4.9).

Table 4.2: Numbers and percentages of chemicals decreasing as expected in all sites, separated into three main clusters

	Cluster size	**Cuijk chemicals decreasing**	**Grou chemicals decreasing**	**Hapert chemicals decreasing**
Cluster 1 (highest in G)	34	5 (14.7%)	9 (26.5%)	12 (35.5%)
Cluster 2 (highest in C)	24	6 (25.5%)	5 (20.8%)	3 (12.5 %)
Cluster 3 (highest in H)	43	28 (65.1%)	19 (44.2%)	32 (74.4%)
All chemicals	101	39 (38.6%)	33 (32.7%)	47 (46.5%)

Cluster 1, which represents chemicals with the highest concentrations in Grou, only a small percentage of chemicals show the expected monotone decrease (14.7% of chemicals in Cuijk 26.5% in Grou and 35.5% in Hapert) (Table 4.2. Cluster 1 contains all PCBs and personal care products measured (Figure 4.6). Interestingly, PCB concentrations decrease in both Grou and Hapert (Figure 4.8 and Figure 4.9).

Cluster 2 represents chemicals with the highest concentrations detected in the Cuijk site and contains chemicals which are not removed very efficiently (25% of chemicals decrease in Cuijk, 20.8% in Grou and 12.5% in Hapert). Interestingly, in all sites, chemicals that are not decreasing have higher concentrations in mesocosms 3 or 4 compared to mesocosms 1 and 2. Another interesting dynamic is the increase in mesocosm 2, followed by decrease in mesocosm 3 with return to similar concentrations as in mesocosm 1 (Figure 4.7, Figure 4.8, Figure 4.9).

4.4.2 Chemical risk analysis confirms complex dynamics in the remediation process

Although analysing concentrations and their dynamics is interesting, it is difficult to separate the dynamics at concentration ranges where they might have biological effects from low concentration fluctuations that may be unlikely to affect organisms. In order to select only chemicals at concentrations which might have biological effects, biological risk was calculated (computed as the ratio between the concentration of a chemical and the predicted no-effect concentration (PNEC) [58]) over the remediation process (Table 4.3). The chemicals with high and medium risk are mapped to clusters in Figure 4.6.

Table 4.3: Number of chemicals with different chemical risk ranges. Chemical risk was calculated as concentration/PNEC (predicted no-effect concentration)

	Number of chemicals
All chemicals detected	120
Chemical risk >1	17
Chemical risk >0.5	21
Chemical risk >0.1	35

In order to find out whether chemical removal is linked to site or same type of chemicals are removed in all sites, chemical sets that decrease in risk from mesocosm 1 to mesocosm 3 were compared between the sites. The results indicate that from 72 chemicals decreasing in any site, 28 are decreasing in all sites. There are further 21 chemicals that are decreasing in 2 sites and only 23 chemicals are decreasing in only 1 site (Figure 4.10, Table 4.4). From high-risk chemicals, aldicarb, chlorpyrifos, fluoranthene, pirimiphos methyl, propoxur and pyrene are all decreasing in all sites. From 11 chemicals of medium risk (risk > 0.1), 6 decrease in all sites. Moreover, most PAHs that decrease during remediation do so in all sites (9 from 11) Table 4.4. Finally, we asked whether the efficiency of the purification process is affected by the lipophilicity of the chemical. We addressed this question by comparing the distribution of a measure of lipophilicity (log K_{ow}) between chemicals that decrease along the purification process (group 1) and chemicals that are not eliminated

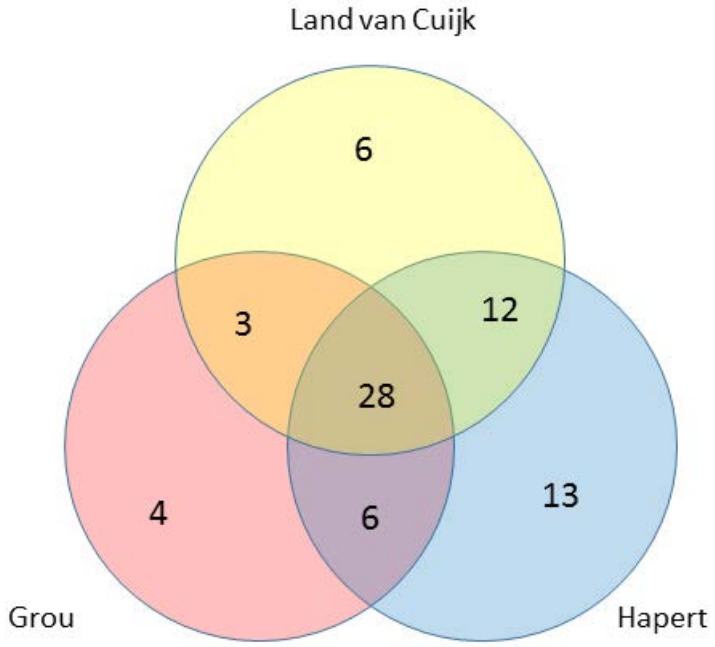


Figure 4.10: Number of chemicals decreasing in each of the sites and overlaps between sites.

efficiently (group 2) by the remediation process (Figure 4.11). These distributions are indeed different ($p - value = 0.0123$), indicating that chemicals with higher lipophilicity are eliminated more efficiently.

4.4.3 The transcriptional state of stickleback livers correlates with changes in chemical concentrations

Having characterised changes in chemical concentration and chemical risk over the remediation process, we then set to assess whether the molecular state of the fish livers is correlated with the concentration of chemicals in the water. We first set to assess whether the remediation process affects stickleback liver gene expression. We compared gene expression in livers between different remediation stages and discovered considerable differences. Such differences were detected in both male and female fish (Figure 4.12).

More specifically, in female sticklebacks there are 3024 differentially expressed genes

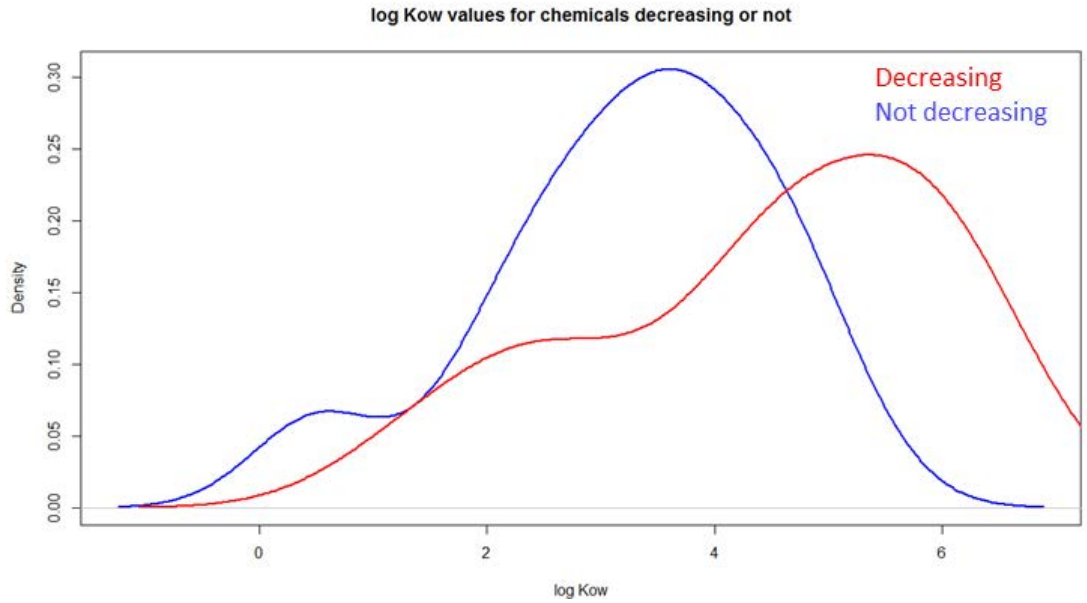


Figure 4.11: log K_{ow} values of chemicals decreasing and not decreasing during remediation. The group of chemicals decreasing was defined as decreasing in at least 1 site and the group of chemicals not decreasing was defined as not decreasing in any site.

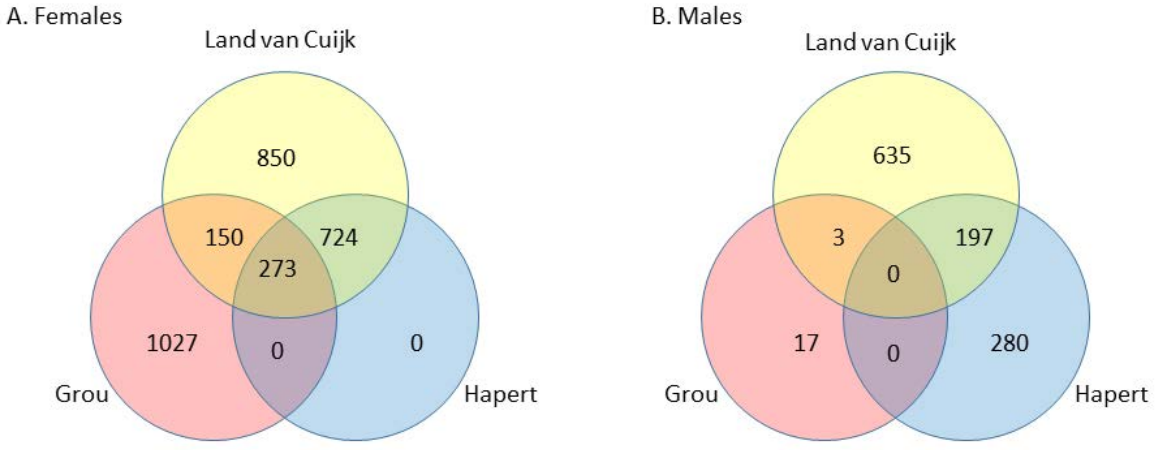


Figure 4.12: Venn diagram of overlap between differentially expressed genes (FDR 0.01) in stickleback livers across sampling points 1-3 in three sites: Land van Cuijk, Grou and Hapert. Sampling point 4 was not used due to possible flooding from sampling point 1 into 4. Two separate Venn diagrams show differentially expressed genes in female (A) and male (B) sticklebacks.

Table 4.4: Chemical risk decreasing between mesocosms 1 and 3 (after sedimentation pond and after helophyte fields – remediation stages in common between all sites). Red background colour indicates chemicals that have high risk in at least one site (risk > 1), yellow shows chemicals with medium risk (> 0.1) and white chemicals with risk < 0.1. Sites are shown as G for Grou, H for Hapert and C for Land van Cuijk. Chemicals type is shown for each chemical: insecticides (ins.), herbicides (herb.), fungicides (fung.), polycyclic aromatic hydrocarbons (PAH), flame retardants (flame r.), industrial (ind.), phthalates (phth.), pharmaceuticals (pharm.), personal care (pers.).

G, H and C	H and C	H and G	C and G	H	C	G
Aldicarb (ins.)		Oxamyl (pest.)	Thiacloprid (ins.)	Estradiol (estro.)	Diazinon (ins.)	Triclosan (per.)
Chlorpyrifos (ins.)					Metalachlor (herb.)	
Fluoranthene (PAH)					Pirimicarb (ins.)	
Pirimiphos methyl (ins.)						
Propoxur (ins.)						
Pyrene (PAH)						
1,12-benzoperylene (PAH)	Endosulfan (ins.)	9,10-anthraquinone (ind.)	Carbofuran (ins.)	Pendimethalin (herb.)		
Tebufenpyrad (ins.)	Triazamite (ins.)					
Benz(a)-anthracene (PAH)						
Galaxolide (musk)						
Simazine (herb.)						
Triphenyl phosphate (flame r.)						
1,2,5,6-dibenzanthracene (PAH)	Benzo(k)-fluoranthene (PAH)	Clotrimazole (fung.)	Anthracene (PAH)	Bisphenol A (ind.)	Cycloxydim (herb.)	Acetaminophen (pharm.)
4-dichlorobenzene (ind.)	Caffeine (pers.)	Difenoconazole (fung.)		butylbenzyl phthalate (phth.)	Cyprodinil (fung.)	Chlorpropham (herb.)
Acetanaphtylene (PAH)	Dichlobanil (herb.)	Iprodione (fung.)		Diethyl phthalate (phth.)	Triallate (herb.)	Pyrazon (ins.)
tAcetyl methyl tetramethyl tetralin (musk)	Diethyl-stilbestrol (pharm.)	tri-(2-chloroisopropyl)-phosphate (flame r.)		Estrone (estro.)		
Benzo(b)-fluoranthene (PAH)	Dimethyl phthalate (phth.)			Gemfibrozil (pharm.)		
Benzo(a)pyrene (PAH)	Enilconazole (fung.)			Nonylphenol (xenoestro, ind.)		
Bifenox (herb.)	Indeno(1,2,3-cd)pyrene			Prochloraz (fung.)		
Chrysene (PAH)	Metamitron (herb.)			Propioconazole (fung.)		
Dibutyl phthalate (phth.)	Musk ketone (musk)			Tributyl phosphate (flame r.)		
Dichlorohexyl phthalate (phth.)	Pentachloro-benzene (flame r.)			Tris(2-butoxyethyl)-phosphate (flame r.)		
Diflufenican (herb.)				Tris(2-ethylhexyl)-phosphate (flame r.)		
Hexachloro-benzene (ind.)						
Metazachlor (herb.)						
Musk xylene (musk)						
Thiabendazole (fung.)						
Triadimefon (fung.)						

between mesocosms 1-3 (FDR 0.01, Figure 4.12 A). In female sticklebacks (Figure 4.12 A), Land van Cuijk has the largest number of differentially expressed genes across the 3 stages of purification (1997), followed by Grou (1450). Hapert has 997 differentially expressed genes. There are 273 genes that are differentially expressed in all three sites. While Grou and Land van Cuijk both have differentially expressed genes specific only for these sites, all differentially expressed genes in Hapert females are same as in Land van Cuijk. Grou and Land van Cuijk share 150 differentially expressed genes specific for these two sites but not Hapert (Figure 4.12 A).

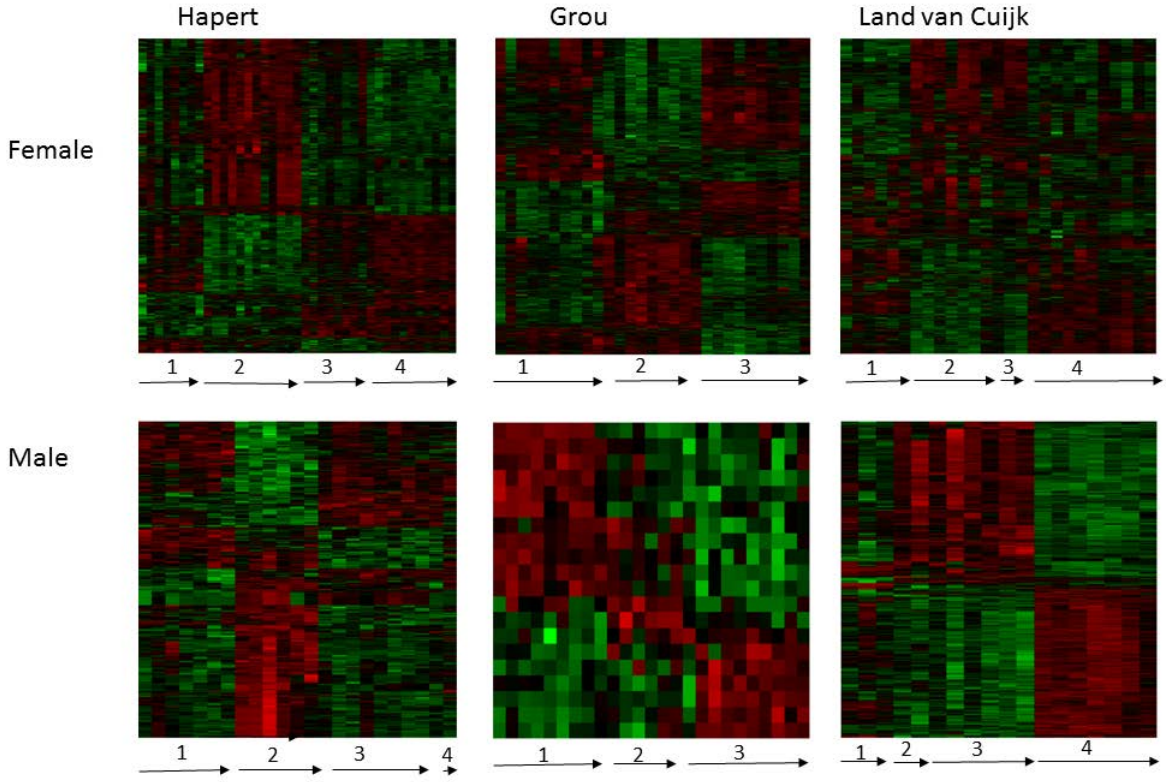


Figure 4.13: Heatmaps showing differentially expressed genes (FDR 0.01) for every site for male and female stickleback. Position 1: after WWTP, position 2: after sedimentation, position 3: after helophyte fields bed, position 4: after ecological lagoon/wetland forest/discharge ditch.

In male sticklebacks, there are 1132 differentially expressed genes between mesocosms 1-3 (FDR 0.01, Figure 4.12 B). The largest number of differentially expressed genes is in Land van Cuijk (835), followed by Hapert (477) and Grou (20). There are no differentially expressed genes in common between all sites in male sticklebacks. Hapert and Land van

Cuijk share 197 differentially expressed genes. Grou and Land van Cuijk share 3 genes while Hapert and Grou share none. This is similar to females, where the largest number of differentially expressed genes in common is between Hapert and Land van Cuijk, while Grou and Hapert share none. However, in male sticklebacks, there are 280 Hapert-specific differentially expressed genes. In order to visualise the overall dynamics of gene expression during the different stages of remediation we performed cluster analysis (Figure 4.13). Visual inspection of the heatmaps of clustered genes revealed the existence of "cyclic" dynamics. Majority of genes increase or decrease in expression between mesocosms 1 (after the WWTP) and 2 (after the sedimentation pond) and then they return to the initial expression levels in mesocosm 3 (after helophyte fields). Heatmaps also show that there are some genes for which expression either monotonally increases or decreases. For example, in Hapert and Grou females, there are genes that show lower expression in mesocosms 1 and 2 and higher expression in mesocosm 3. In Grou females, some genes have higher expression in mesocosms 1 and 2 and lower in mesocosm 3. In Hapert males, there are genes that show higher expression in mesocosm 1 and lower in mesocosms 2 and 3. In Grou males, the 20 differentially expressed genes all follow unidirectional pattern – either increasing or decreasing from mesocosm 1 to mesocosm 3.

In order to visualise the changing molecular state of fish livers we performed Principal Component Analysis (PCA) (Figure 4.14). The analysis revealed that in Hapert males and females and in Grou females, largest changes in gene expression as described by principal component 1 are "cyclic" (Figure 4.14). Principal component 2 also describes dynamics during remediation steps – from mesocosm 1 to mesocosm 3, all changes are in the same direction in male and female stickleback in Hapert and female stickleback in Grou. Males in the Grou site are different from Hapert and females in Grou, with less significantly expressed genes (20 in total), showing unidirectional changes on principal component 1 and cyclic changes in principal component 2. Stickleback gene expression in Land van Cuijk differs from Grou and Hapert – remediation steps 2 and 3 are not as distinct as in other sites and also the characteristic cyclic changes as seen in Hapert and

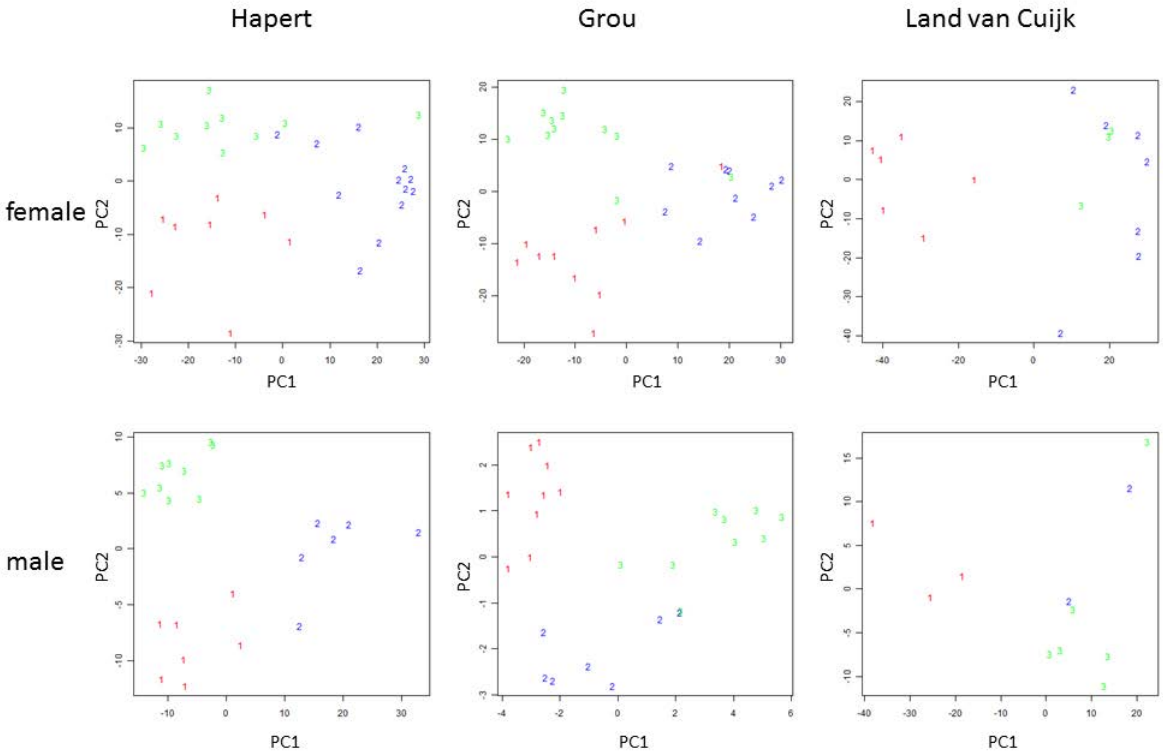


Figure 4.14: Principal components 1 and 2 in Hapert, Grou and Land van Cuijk positions 1-3, for males and females. PCA was performed in the space of significant genes (FDR 0.01)

Grou are not seen in Land van Cuijk.

The analysis of chemicals showed that chemical concentrations follow complex patterns, some unidirectional as expected during remediation, but some more unintuitive, with chemical increase in the later steps of remediation. Exploratory analysis of gene expression also showed that there are various patterns of dynamics. Most changes described by principal component one follow "cyclic" changes and 2nd component describes changes that are more intuitive – gene expression either increases or decreases during the remediation steps. To understand the complex associations between chemical concentrations and gene expression, an integrated analysis of all available data should be performed.

4.4.4 A systems biology approach linking gene expression to chemical concentrations of high-risk chemicals and algae toxicity

Network integration of all data

In order to gain more understanding of the whole system of remediation we integrated chemical concentrations and their relationship with stickleback morphology (as described by traditional endpoints), gene expression in the stickleback liver, and toxicity tests. A convenient way of integrating and analysing such data is to use network representation, where chemicals, genes, physiological parameters and toxicity are represented by nodes and the connecting edges represent the similarity between the concentration and expression profiles. Because of the differences in physiology we made separate networks to represent both male and female stickleback. Resulting networks included 3306 nodes and 2233 nodes in females and males, respectively).

The network representing female stickleback has 3306 nodes and 86871 edges (Figure 4.15). For ease of interpretation, this large network could be divided into different but interconnected network modules, based on gene function. Modularisation of the large

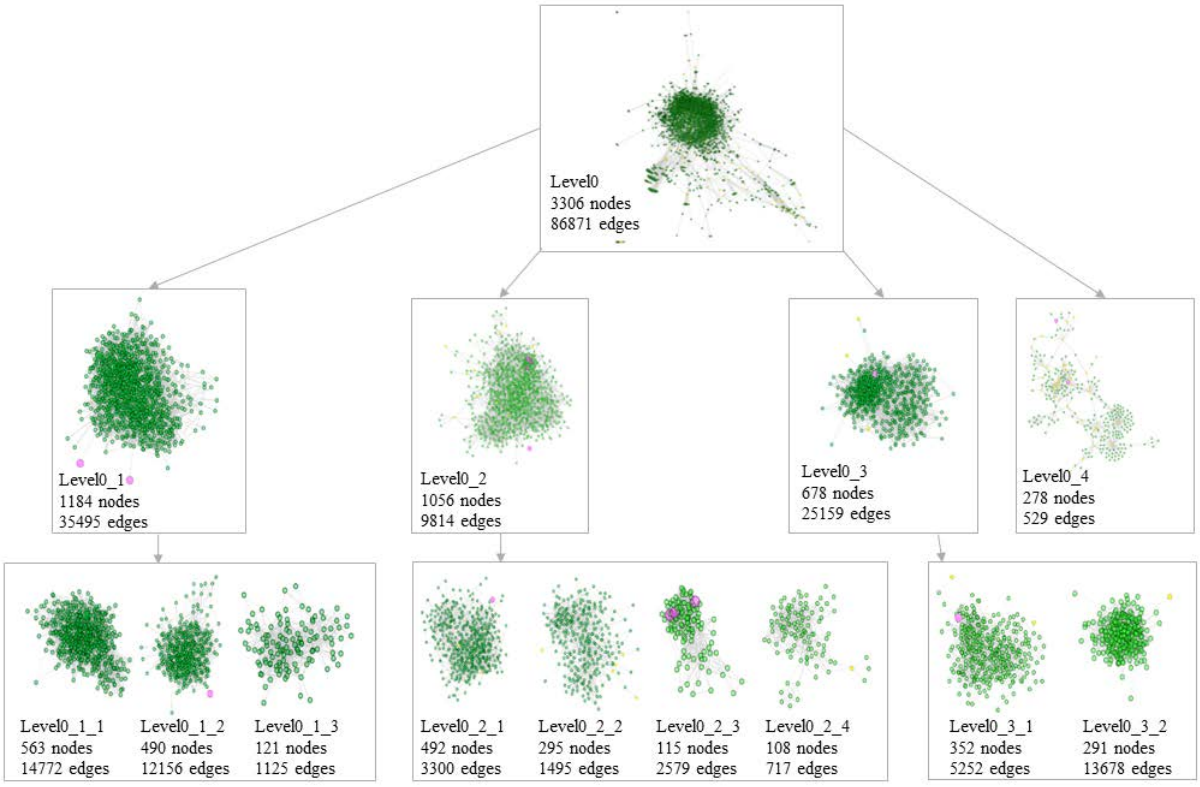


Figure 4.15: GLay modularisation of ARACNE mutual information network of differentially expressed genes (from all sites) in the female stickleback. Modularisation was performed twice (recursively). Green nodes represent genes, pink nodes physiological measurements or toxicity tests and yellow nodes are concentrations of chemicals captured by the passive sampler

Table 4.5: Individual modules of the female network — two rounds of modularisation are shown.

Module	Nodes	Health	Toxicity tests	Chemicals
0_1	1184		MTT, liver weight, hepatosomatic index, ER-luc	
0_1_1	536		liver weight, hepatosomatic index	
0_1_2	490		ER-luc	
0_1_3	121			
0_2	1056	weight, length, condition factor	microtox TU EC50	11
0_2_1	492		microtox TU EC50	3
0_2_2	295			5
0_2_3	115	weight, length, condition factor		
0_2_4	108			1
0_3	678	% survival Oct 2010		3
0_3_1	352	% survival Oct 2010		2
0_3_2	291			1
0_4	278		AR-luc, algae TU EC50	47

network of female stickleback resulted in 4 modules (sizes 1184, 1056, 678 and 278 nodes) (Figure 4.15). Three larger modules could be modularised further (module sizes were 563, 490 and 121 nodes for the network of 1184; 492, 295, 115 and 108 nodes for the network of 1056; 352 and 291 nodes for the network of 678). Modules are different in respect to the number of chemicals correlated with gene expression and various physiological parameters and toxicity tests are also in different modules (Table 4.5). For example, MTT, liver weight, hepatosomatic index and ER-luc assay are in module 0_1 in the female stickleback network. Module 0_2 contains weight, length and condition factor, and also MTT TU EC50 and 11 chemicals. These chemicals include phthalates (dicyclohexyl phthalate, diethyl phthalate, butylbenzyl phthalate, dibutyl phthalate), white mineral oil, metazachlor, gemfibrozil, carbofuran, diethylstilbestrol, pyrazon and aldicarb in module 0_2 (Table 4.6). Module with most chemicals (47) was 0_4, which is correlated with the AR-luc assay and algae TU EC50 toxicity test (Table 4.5).

Table 4.6: Chemicals correlated with gene expression (FDR < 0.05) in different modules of the female stickleback network

Module	**Chemicals**
0_1	-
0_2	white mineral oil, metazachlor, dicyclohexyl phthalate, diethyl phthalate, gemfibrozil, carbofuran, diethylstilbestrol, butylbenzyl phthalate, pyrazon, aldicarb, dibutyl phthalate
0_3	1,2,5,6-dibenzanthracene, iprodione, acetaminophen
0_4	43 chemicals shown in Figure 4.16

Detailed overview of module 0_4 (Figure 4.16) shows that from chemicals correlated with gene expression, several have high or medium risk. Many of the chemicals are also decreasing in at least one of the sites. Especially in the central part of the network, linked more to Algae TU EC50, there are many polycyclic aromatic hydrocarbons (PAHs).

The network of male stickleback is smaller due to fewer differentially expressed genes (Figure 4.17, Table 4.7) – however, the overall network is larger than the sum of differentially expressed genes and chemicals, as there were many genes which are not significantly differentially expressed between 3 stages of remediation, but are correlated with chemical concentrations. The integrated similarity network of all differentially expressed genes, physiological parameters, toxicity tests and chemical concentrations has 2233 nodes. This large network could be modularised further, resulting in 5 subnetworks after first round of modularisation. The sizes were 441, 413, 337, 305, 202 and 140 nodes. Two larger networks were modularised further (194, 162 and 91 nodes for the network of 441; 243 and 140 nodes for the network of 413). Similar to the network of female stickleback, weight, length and condition factor are in one module in the network of male stickleback (Table 4.7). Also, most chemicals are correlated with genes in one module (0_58), which is correlated with algae TU EC50 toxicity assay (Figure 4.18, Table 4.7). Another module contains genes correlated with liver weight and hepatosomatic index, which is again similar to the network of female stickleback. However, in the network of male stickleback,

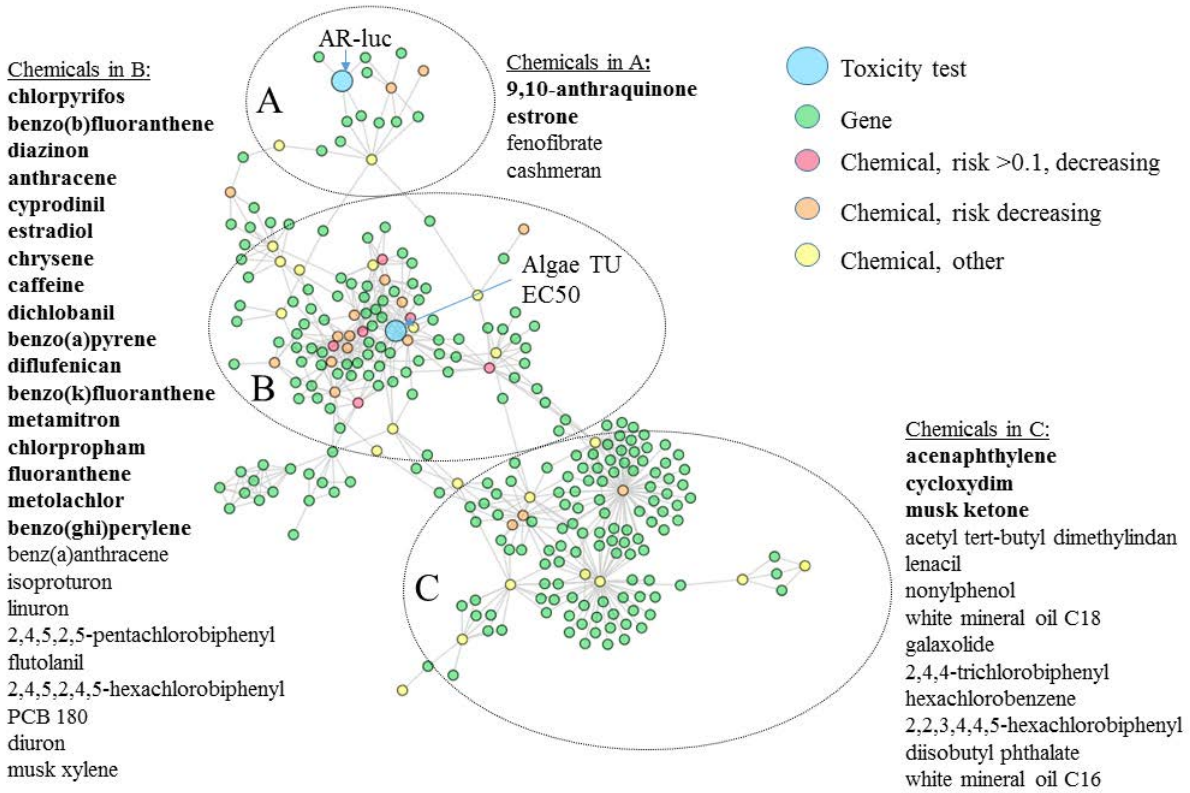


Figure 4.16: Module 0_4 in the network of female stickleback. Chemicals in different areas of the network are shown – bold font indicates chemicals that have decreasing chemical risk in at least one site

% survival is also in the same module. In the network of male stickleback, ER-luc and AR-luc assays are in a module that also contains 9 chemicals (Table 4.7).

The network approach shows that differentially expressed genes in both male and female sticklebacks can be modularised into distinct clusters (Figure 4.15, Figure 4.17, Table 4.5, Table 4.7). Interestingly, some of the physiological measurements, chemical concentrations and toxicity tests are also related to specific modules. In networks of both female and male stickleback, there are modules that contain most of the chemicals and interestingly, genes in these modules are also correlated with algae TU EC50 toxicity test.

Further analysis of chemicals and genes in the two modules containing most chemicals in both male and female sticklebacks reveals that 26 chemicals from a total of 50 are correlated with both male and female gene expression in these modules (Figure 4.19), including 7 with high risk. Among the 26, there are 1 PAH with chemical risk > 1 and 2 PAHs with chemical risk > 0.1. However, many other PAHs with lower chemical risk are correlated with gene expression in both male and female stickleback and PAHs are also decreasing during the remediation stages in most sites. In addition to PAHs, there are some PCBs and musks correlated with gene expression – for these, PNEC is not known, so chemical risk could not be calculated.

There are 5 chemicals correlated with gene expression in the network module of male stickleback (M_0_58) that are not correlated with genes in the network module of female stickleback (F_0_4) (Figure 4.19 A, Table 4.9). All of these have chemical risk < 0.1 and are phthalates, pesticides and a pharmacological (Paracetamol). There are more chemicals correlated with only module F_0_4 (Figure 4.19 A, Table 4.10): from 19 chemicals (musks, industrial chemicals, pesticides, pharmacological agents, PCBs and oils), 3 had chemical risk > 0.1 and others < 0.1 or not known.

Despite many chemicals correlated with gene expression in both male and female sticklebacks, genes associated with chemical concentrations in these modules are different. This could be because of sex-specific effects. However, from especially from the analysis of most chemical-correlated modules, it is clear that high-risk chemicals affect both male

and female stickleback. Moreover, even chemicals with risk < 1 such as some PAHs which decrease in all sites are correlated with gene expression in both sexes.

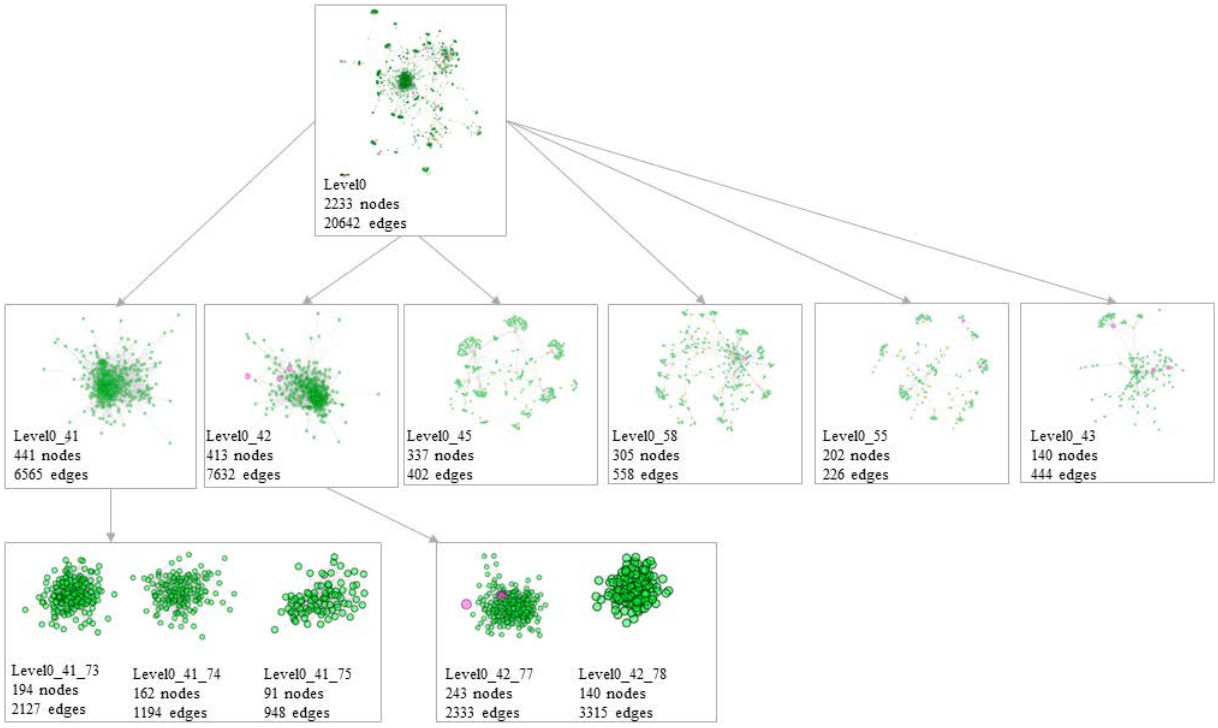


Figure 4.17: GLay modularisation of ARACNE mutual information network of differentially expressed genes (from all sites) in the male stickleback. Modularisation was performed twice (recursively). Green nodes represent genes, pink nodes physiological measurements or toxicity tests and yellow nodes are concentrations of chemicals captured by the passive sampler

Table 4.7: Individual modules of the male network – two rounds of modularisation are shown

Module	Nodes	Health	Toxicity tests	Chemicals
0_41	441			
0_41_73	194			
0_41_74	162			
0_41_75	91			
0_42	413	weight, length, condition factor		
0_42_77	243	weight, length, condition factor		
0_42_78	140			
0_45	337			14
0_58	305		algae TU EC50	31
0_55	202		ER-luc, AR-luc	9
0_43	140	% survival April 2010, liver weight, hepatosomatic index		1

Table 4.8: Chemicals correlated with gene expression (FDR < 0.05) in different modules of the male stickleback network

Module	**Chemicals**
0_41	-
0_42	-
0_43	clotrimazole
0_45	4-dicylochlorobenzene, aldicarb, dibutyl phthalate, hexachlorobenzene, cycloxydim, dicyxlohexyl phthalate, diisobutyl phthalate, galaxolide, lenacil, musk ketone, white mineral oil, white mineral oil c16, white mineral oil c18
0_55	2,3,4,4,5-pentachlorobiphenyl, 9,10-anthraquinone, estrone (POS), fenofibrate, cashmeran, cyprodinil, iprodione, isoproturon, white mineral oil c17
0_58	29 chemicals as shown in Figure 4.18

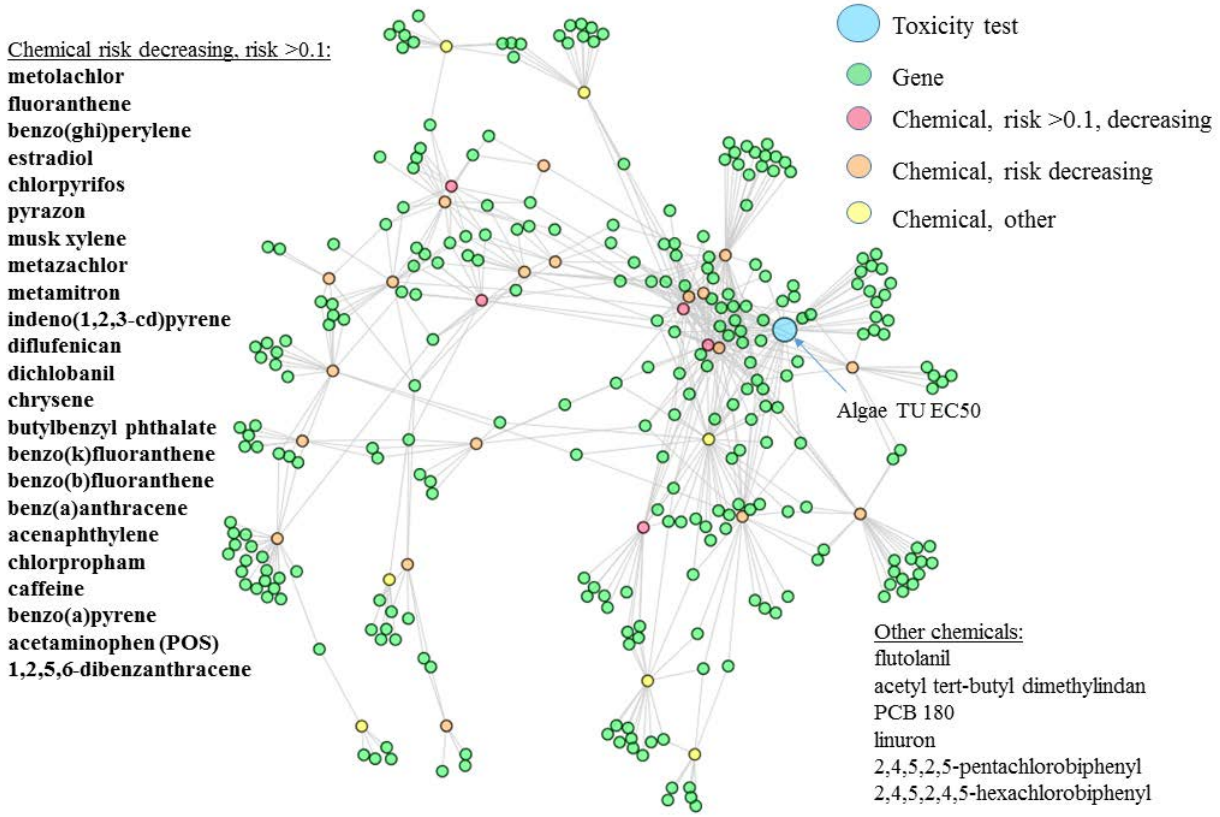
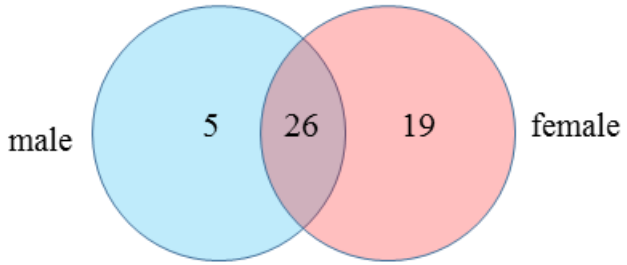


Figure 4.18: Module 0_58 in the network of male stickleback.

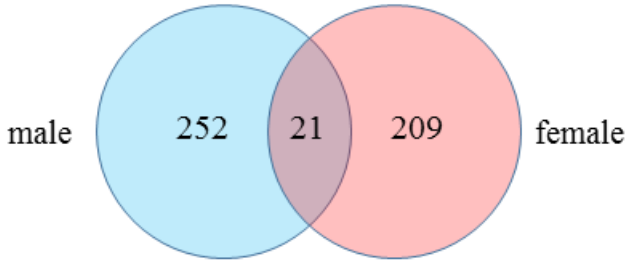
Table 4.9: Chemicals correlated with gene expression in only male stickleback network module

Chemicals in male module 0_58	**Type**	**Ch. risk**	**Decreasing?**
1,2,5,6-dibenzanthracene	PAH	< 0.1	all
Metazachlor	herbicide	< 0.1	all
Acetaminophen	pharmacological	< 0.1	G
Pyrazon	insecticide	< 0.1	G
Butylbenzylphthalate	phthalate	< 0.1	H

A



B



C

Chemicals associated with both modules	Type	Ch. risk	Decreasing?
fluoranthene	PAH	>1	all
Chlorpyrifos	insecticide	>1	all
metolachlor	herbicide	>1	C
Diazinon	Insecticide	>1	C
Estradiol	estrogen	>1	H
flutolanil	fungicide	>1	-
Linuron	herbicide	>1	-
1,12-benzoperylene	PAH	>0.1	all
benz(a)anthracene	PAH	>0.1	all
acenaphthylene	PAH	<0.1	all
anthracene	PAH	<0.1	all
benzo(b)fluoranthene	PAH	<0.1	all
Benzo(a)pyrene	PAH	<0.1	all
chrysene	PAH	<0.1	all
diflufenican	herbicide	<0.1	all
musk xylene	musk	<0.1	all
acetyl tert-butyl dimethylindan	musk	No PNEC	all
benzo(k)fluoranthene	PAH	<0.1	H and C
indeno(1,2,3-cd)pyrene	PAH	<0.1	H and C
dichlobanil	herbicide	<0.1	H and C
Caffeine	personal care	<0.1	H and C
metamitron	herbicide	<0.1	H and C
2,4,5,2,5-pentachlorobiphenol	PCB	No PNEC	C and G
Chlorpropham	herbicide	<0.1	G
PCB 180	PCB	No PNEC	G
2,4,5,2,4,5-hexachlorobiphenol	PCB	No PNEC	-

Figure 4.19: A: Chemicals associated with modules 0_4 in females and 0_58 in males. B: Genes in modules 0_4 of female and 0_58 of male stickleback. C. annotation of chemicals overlapping between male and female modules 0_4 and 0_58

Table 4.10: Chemicals correlated with gene expression in only female stickleback network module

Chemicals in female module 0_4	**Type**	**Ch. risk**	**Decreasing?**
Galaxolide	musk	> 0.1	all
9,10-anthraquinone	industrial	> 0.1	H and G
Diuron	herbicide	> 0.1	-
Hexachlorobenzene	industrial	< 0.1	all
Cycloxydim	herbicide	< 0.1	C
Cyprodinil	fungicide	< 0.1	C
Estrone	estrogen	< 0.1	H
Musk ketone	musk	< 0.1	H and C
Nonylphenol	industrial	< 0.1	H
Fenofibrate	pharmacological	< 0.1	-
Isoproturon	herbicide	< 0.1	-
Lenacil	herbicide	< 0.1	-
Diisobutylphthalate	phthalate	no PNEC	all
Moskene	musk	no PNEC	H and G
2,2,3,4,4,5-hexachlorobiphenol	PCB	no PNEC	G
2,4,4-trchlorobiphenol	PCB	no PNEC	H and G
White mineral oil c16	oil	no PNEC	H and G
White mineral oil c18	oil	no PNEC	H and G
6,7-dihydro-1,1,2,3,4-pentamethyl-4-(5H)indanone (Cashmeran)	fragrance	no PNEC	H

Functional annotation of network modules

Functional annotation of both male and female networks show that modules are enriched in different biological functions (Figure 4.20). Significant KEGG pathways ($FDR < 0.05$) are related to metabolism (purine metabolism, pentose phosphate pathway, galactose metabolism, steroid biosynthesis), signalling (PPAR signalling pathway, p53 signalling pathway, mTOR signalling pathway), cell cycle (DNA replication, cell cycle), immunity and infection (antigen processing and presentation, shigellosis, pathogenic Escherichia coli infection) and diabetes (insulin signalling pathway, type I diabetes mellitus, maturity onset diabetes of the young). Module of the female stickleback that contains most chemicals is enriched in mTOR signalling. Network module of male stickleback with most chemicals is not significantly enriched in any KEGG pathways. Modules correlated with length, weight and condition factor (F_0_2 and M_0_42) are enriched in antigen processing and presentation, pyruvate metabolism, histidine metabolism in the female stickleback network and cell cycle, PPAR signalling pathway, biosynthesis of unsaturated fatty acids, aminoacyl-tRNA biosynthesis, alpha-linolenic acid metabolism, steroid biosynthesis, synthesis and degradation of ketone bodies, terpenoid backbone biosynthesis, pyrimidine metabolism, DNA replication and ECM-receptor interaction in the network module of males. Gene Ontology Biological Process enrichment analysis (Figure 4.21, Figure Elec.Supp2 in the Electronic Supplementary) shows that in networks of male and female stickleback, xenobiotic metabolic process is enriched in three modules: M_0_41_75 and F_0_2 and F_0_2_3. M_0_41_75 is a small module of 91 nodes and is not correlated with any physiological measurements, toxicity tests or chemicals. Module F_0_2_3 is also small module, but is correlated with weight, length and condition factor. Module F_0_2 is a large module (parent module of F_0_2_3) and is also correlated with microtox TU EC50, weight, length, condition factor and 11 chemicals. Biological processes that were enriched in modules correlated with most chemicals, M_0_58 and F_0_4, were response to heat, eye morphogenesis and response to cadmium ion. In addition, the male module was enriched in antigen processing and presentation, steroid biosynthetic process

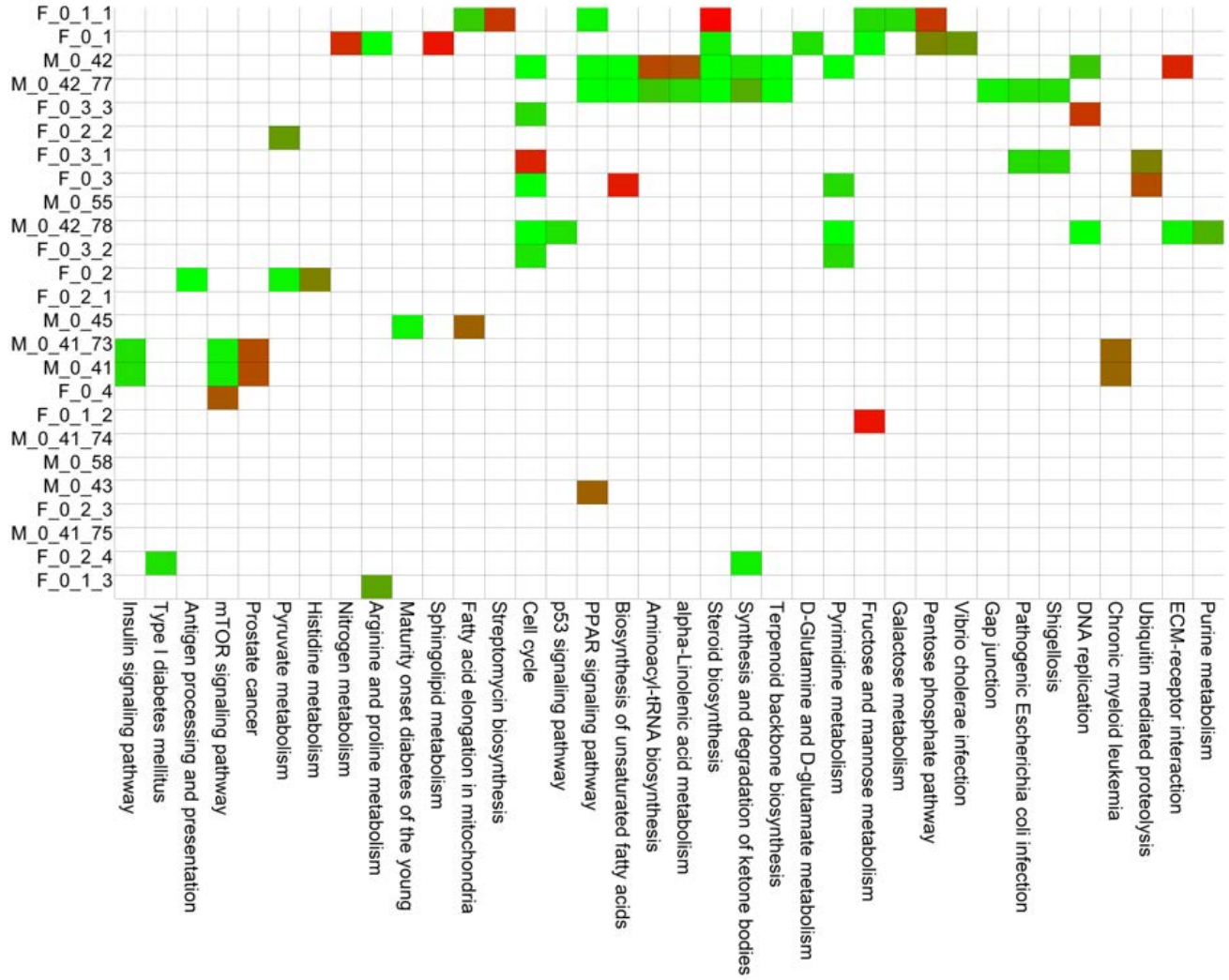


Figure 4.20: Enriched KEGG functions in all network modules. Only significant ($FDR < 0.05$) enrichment is shown (green to red), cells where $FDR > 0.05$ are shown as white.

and ubiquinone biosynthetic process. The network module of female stickleback was additionally enriched in rRNA processing. Network analysis of all differentially expressed

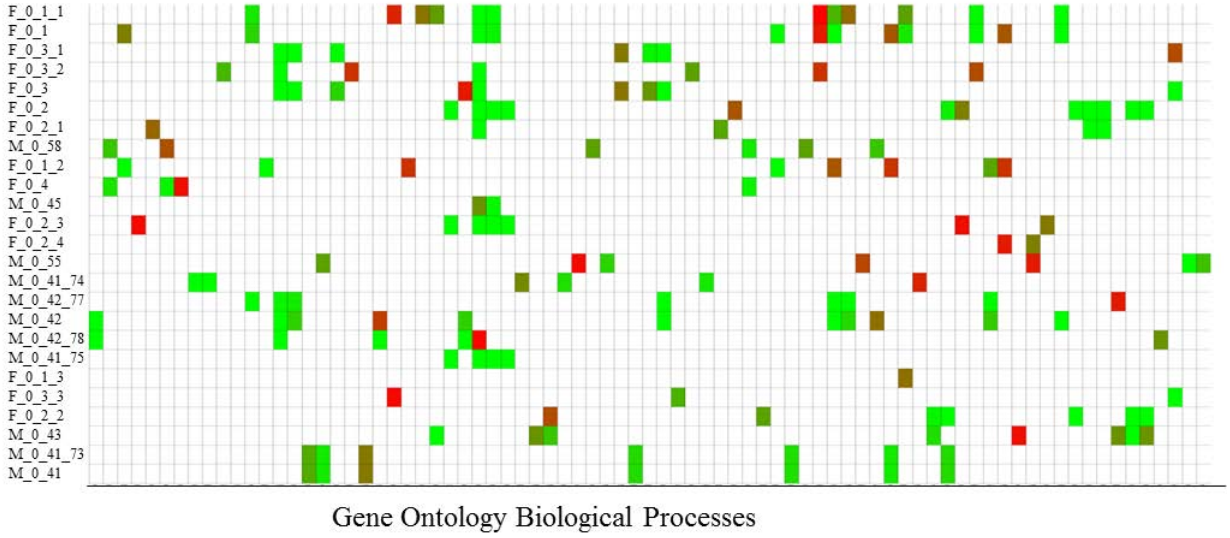


Figure 4.21: Heatmap showing significant Gene Ontology terms in each of the modules. Only significant ($FDR < 0.05$) values are shown in colour (green to red) for terms for which at least 3 genes for this term were present in the module. The aim of this figure is to show that different modules are enriched in different Gene Ontology terms and a large zoomable figure with all Gene Ontology terms is in the electronic supplementary (Electronic Supplementary Elec.Supp2).

genes integrated with chemical concentrations, physiological measurements and toxicity tests has showed that there are distinct parts of the network associated with certain chemicals, physiological parameters and toxicity tests and these modules are also enriched in several KEGG pathways and Gene Ontology Biological Processes. However, this static network, although having advantages of integrating large amounts of data, cannot capture the dynamical characteristics of the real system. Therefore, a more detailed dynamical model, on a smaller scale might be able to give further insights into the remediation system.

4.4.5 Building a Bayesian model that integrates chemicals and biological response and that explicitly models the wastewater remediation process

Work in this section is the result of a collaboration with Dr. Alberto Cassese (Rice University) and Prof. Marina Vannucci (Rice University) who have developed the Bayesian model for wastewater remediation [53].

Static networks as shown in previous paragraph have the advantage of allowing integration of a very large number of variables. However, the static network does not include the information that the state of a given remediation stage depend on the previous one and influence the subsequent. In order to address this issue and to identify the chemicals that are more likely to affect gene expression, the mathematical framework developed by Cassese *et al.* [53] was used. Model-based data integration is data-intensive and therefore can only integrate a smaller number of genes with the available chemical concentration data. Therefore, a subset of chemicals and genes to perform this analysis was selected. Chemicals to use were selected so that they represent different types of dynamics during the remediation process (Figure 4.22). The selected chemicals from each cluster can therefore be treated as representatives of the cluster, making it possible to interpret the model for chemicals with similar dynamics. As different from the static network, Bayesian model was done for male and female sticklebacks together to have more data for each remediation step.

For the current study, only the Hapert site was used as an example as in this site, there was data from all remediation steps (4 mesocosms). In this site, there were more chemicals of high risk than in other sites and PCA of male and female stickleback gene expression showed similar dynamics. Chemicals selected as representatives of clusters were triclosan, alpha-HCH, permethrin-trans, diisobutyl phthalate, chrysene, permethrin-trans, butylbenzyl phthalate, triphenyl phosphate, PCB-28, diethyl phthalate, di-n-octyl phthalate, tetramethrin and estrone. Genes selected for the model were selected to be associated with the selected chemicals according to the CTD database.

As the static models and Bayesian model were run concurrently, the chemicals used in these are not overlapping. This is due to the fact that in the static network, the aim was to integrate data for all 3 sites, separately for male and female sticklebacks, and only chemicals which were detected with certainty in all sites were used (*i.e.* no missing values) to maximize the number of data points. However, as the Bayesian model was ran separately for all sites, but male and female together, it was possible to include certain chemicals in the model of one site which had missing values in other sites. Due to this, some for the chemicals used in the Bayesian model are not in the static networks because they were filtered out due to missing data or uncertainty in some of the sites. Figure 4.22 shows the overlap of selected chemicals with the chemicals that were used in the static network analysis: triclosan, triphenyl phosphate and diethyl phosphate as representatives from the cluster that has highest concentrations in Grou, butylbenzyl phthalate and estrone from the cluster that has highest concentrations in Land van Cuijk and diisobutyl phthalate and chrysene from the cluster that has highest concentrations in Hapert.

4.4.6 Identifying chemical drivers of transcriptional response

The model we developed allow us to identify the most relevant transcriptional changes and in particular the ones that correlate to changes in the concentrations of chemicals. In our model these genes are identified by having a large posterior probability of inclusion (PPI) (Figure 4.23). The largest number of changes occur between the 1st and the 2nd mesocosms while the smaller number of genes are identified between the 3rd and the 4th mesocosms. A PPI threshold of 0.99 applied to the three transitions show that 29 genes from 73 are most affected, with PPI $>$ 0.99 in at least 1 transition (Table 4.11). The set of most affected genes includes vitellogenin and estrogen receptor 1. Other genes, encoding proteins of cytochrome P450 family, glutathione peroxidase, progesterone receptor, peroxisome proliferator activated receptors are present in both sets of high PPI and lower PPI (Electronic Supplementary Table 1). The expression of some genes changes during all transition steps. One such example is vitellogenin. VTG1 has PPI $>$ 0.99 in all

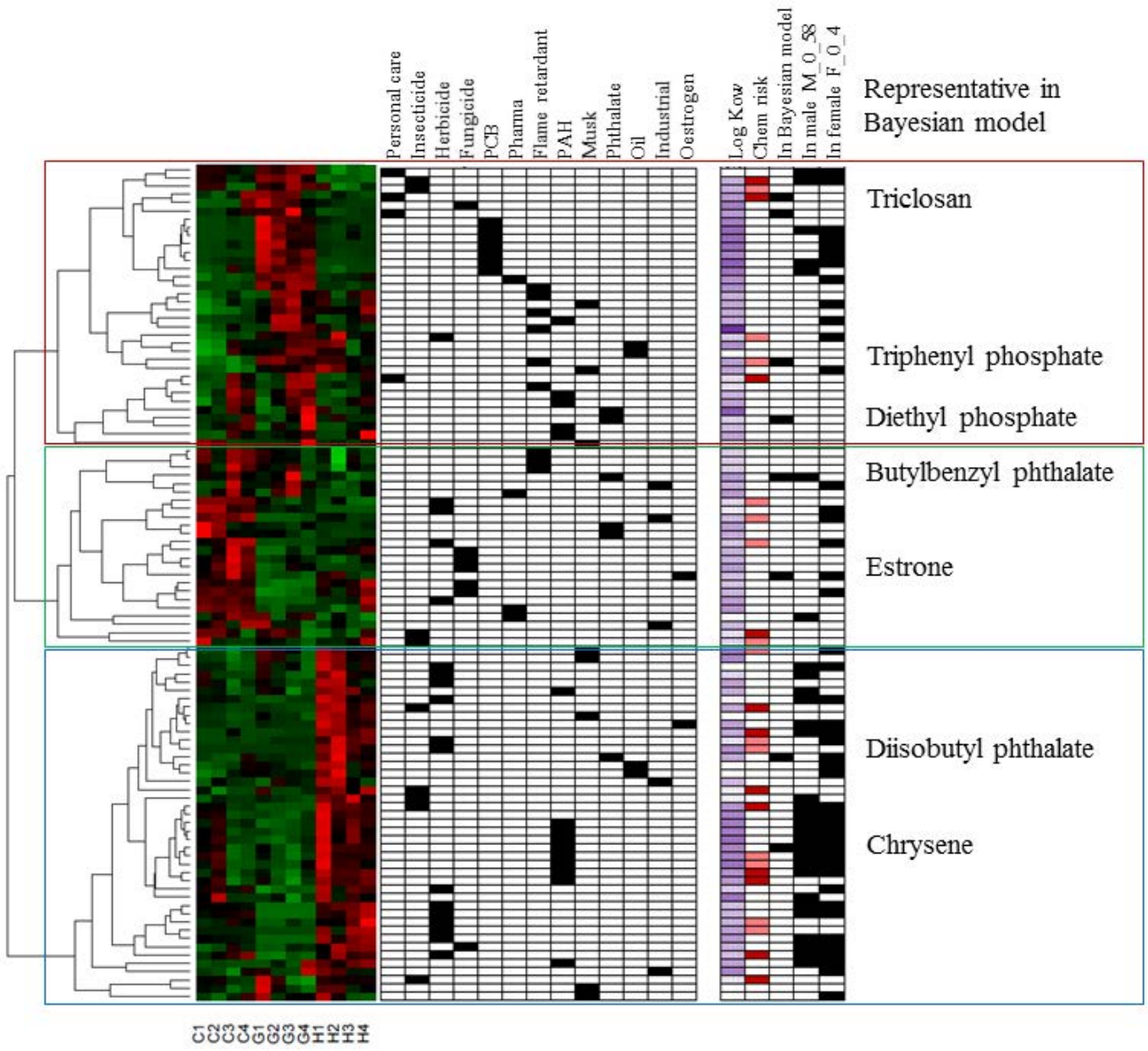


Figure 4.22: Heatmap of chemical concentrations, the log K_{ow} of each chemical, and also whether the chemical was used in the Bayesian model, whether it was correlated with gene expression in male or female stickleback network.

transitions and VTG2 and VTG3 in 1st and 3rd transitions. ESR1 and CYP1C1 also have PPI > 0.99 in all transitions. Some genes have PPI > 0.99 in two transitions: GSTR1, STAR, STARD3NL, STARD9, SULT1ST4, SULT3A1, ZP3, ZP4, VTG2 and VTG3.

4.4.7 Important chemicals

The model we have developed allow us to identify the chemicals that are most likely to be responsible for the changes in transcription. The relative importance of each of the chemicals in driving gene transcriptional changes in the 3 transitions is shown in Figure 4.24. Triclosan and alpha-HCH are in top 3 chemicals in all transitions. Permethrin-trans has larger effect only in transitions 1 and 3 and has more modest effect in the second transition.

Interestingly, chrysene, diisobutyl phthalate, estrone and butylbenzyl phthalate, which are also correlated with gene expression in the static networks, are also within the first 5 chemicals most likely driving the transcriptional changes. In all three transitions, these top chemicals contributing to gene expression changes were also correlated with gene expression in the static networks (Table 4.12). Additionally, tri-phenyl phosphate was placed 4th in the final remediation stage. When considering each of these chemicals in more details, triclosan has effect on all remediation stages, whereas diisobutyl phthalate and chrysene have effect on 2 stages of remediation. Butylbenzyl phthalate and tri-phenyl phthalate have most effect on a single stage of remediation as summarised in Table 4.12. These chemicals of effect are correlated with several other chemicals which are also correlated with gene expression in the static networks (Table 4.12).

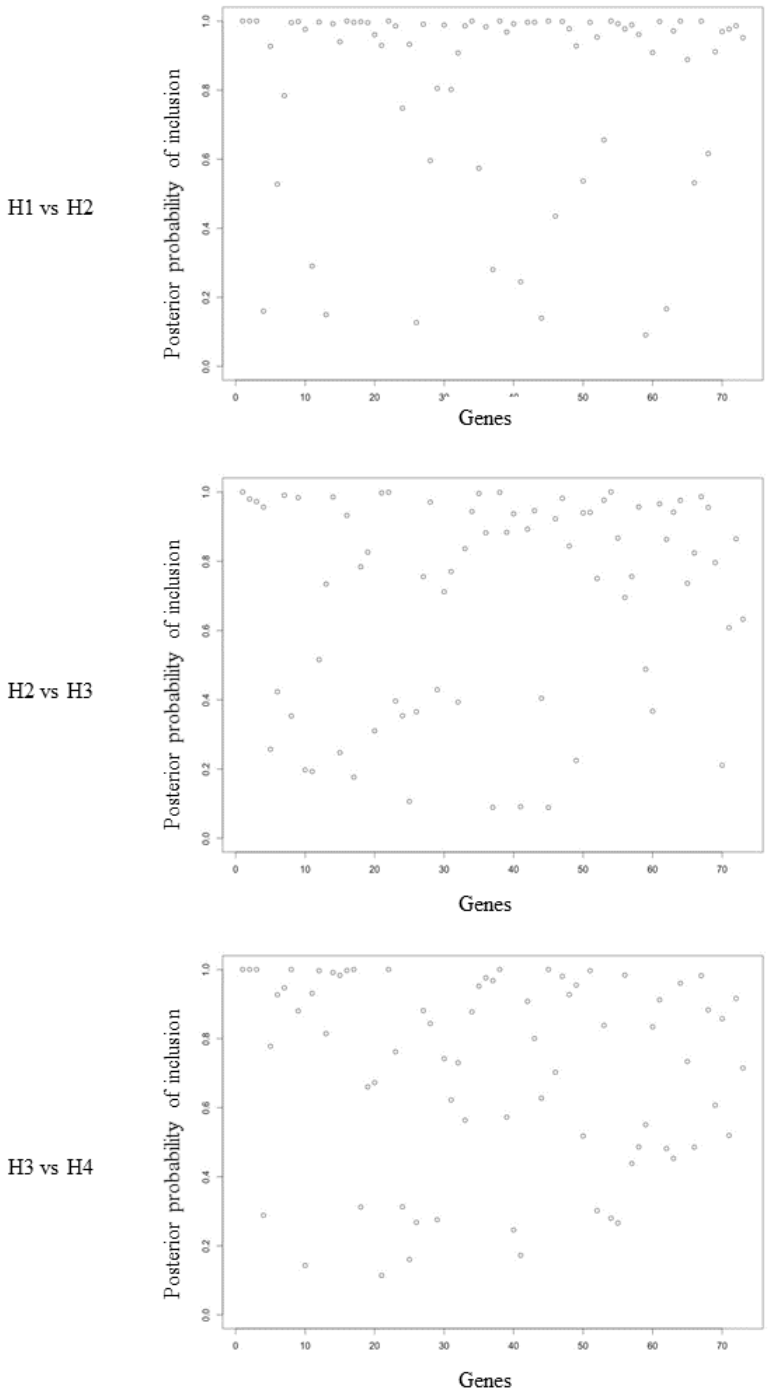


Figure 4.23: Posterior probability of inclusion for 73 genes in three transitions: H1 vs H2 (pre-sedimentation pond to pre-helophyte fields), H2 vs H3 (pre-helophyte fields to after helophyte fields) and H3 vs H4 (after helophyte fields to after 3rd remediation compartment).

Table 4.11: Genes with posterior probability of inclusion > 0.99 in at least one transition. H2 vs H1 is the transition from sedimentation pond to pre-helophyte fields. H3 vs H2 is the transition from pre-helophyte fields to after helophyte fields. H4 vs H3 is the transition from after helophyte fields to after 4th step.

Gene	**H1 vs H2**	**H2 vs H3**	**H3 vs H4**	**Annotation from Ensembl (Human annotation is not specified)**
CAT	0.998	0.966	0.912	Catalase
CHGA	0.930	0.997	0.114	Chromogranin A
CYP17A1	0.574	0.995	0.952	Cytochrome P450 family 17 subfamily A member 1
CYP19A1	0.784	0.990	0.947	Cytochrome P450 family 19 subfamily A member 1
CYP1C1	1	0.999	1	Cytochrome P450 family 1 subfamily C polypeptide 1 (Zebrafish)
ESR1	1	0.999	1	Estragen receptor 1
ESRRG	0.991	0.755	0.881	Estrogen related receptor gamma
GPX1	0.999	0.986	0.982	Glutathione peroxidase 1
GSTA5	0.992	0.937	0.246	Glutathione S-transferase alpha 5
GSTK1	0.997	0.893	0.908	Glutathione S-transferase kappa 1
GSTM3	0.996	0.946	0.800	Glutathione S-transferase mu 3
GSTR1	1	0.089	1	Glutathione S-transferase rho
GSTZ1	0.999	0.982	0.980	Glutathione S-transferase zeta 1
MUC19	1	0.944	0.877	Mucin 19
PGR	1	0.976	0.960	Pregesterone receptor
STAR	0.995	0.353	1	Steroidogenic acute regulatory protein
STARD10	0.998	0.984	0.880	StAR related lipid transfer domein containing 9
STARD3NL	0.997	0.516	0.997	STARD3 N-terminal like
STARD9	0.992	0.985	0.991	StAR related lipid transfer domain containing 9
SULT1ST4	0.996	0.941	0.996	Sulfotransferase family 1, cytosolic sulfotransferase 4 (Zebrafish)
SULT3A1	1	1	0.230	Sulfotransferase family 3A, member 1
SULT4A1	0.992	0.867	0.266	Sulfotransferase family 4A member 1
ZP3	1	0.932	0.997	Zona pellucida glycoprotein 3
ZP4	0.996	0.784	1	Zona pellucida glycoprotein 4
ZPAX	0.998	0.784	0.312	Egg envelope component ZPAZ (Xenopus gene)
ZPD	0.995	0.826	0.660	Zona pellucida protein D (Xenopus gene)
VTG1	1	1	1	Vitellogenin 1 (Zebrafish gene)
VTG2	1	0.972	1	Vitellogenin 2 (Zebrafish gene)
VTG3	1	0.972	1	Vitellogenin 3 (Zebrafish gene)

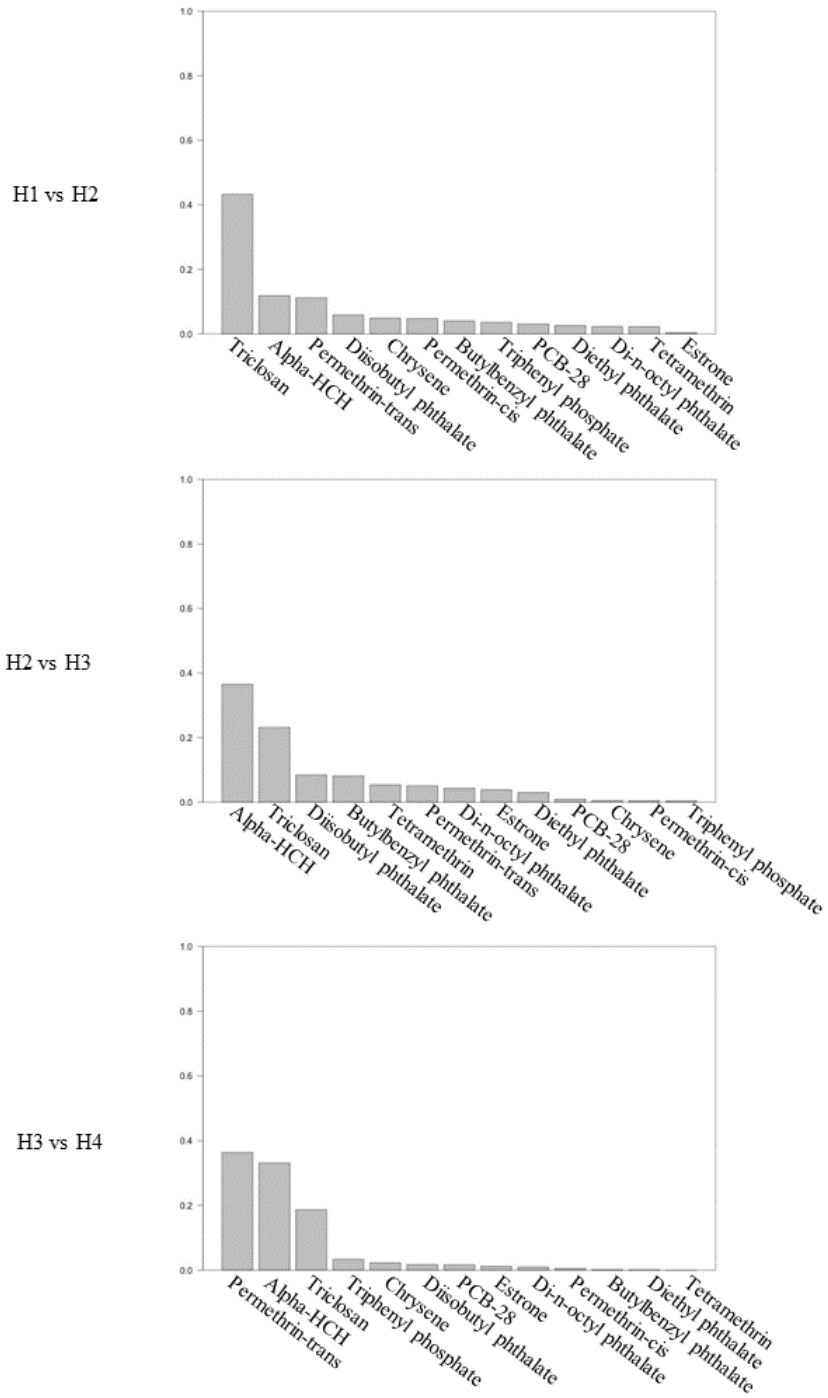


Figure 4.24: Effect of each of the chemicals in transitions H1 vs H2 (pre-sedimentation pond to pre-helophyte fields), H2 vs H3 (pre-helophyte fields to after helophyte fields) and H3 vs H4 (after helophyte fields to after final remediation compartment).

Table 4.12: Mapping of chemicals which have importance in the Bayesian model to the static networks

Chemical	**Most effect**	**Chemicals in same cluster**	**In static network of males**	**In static network of females**
Triclosan	sedimentation, helophytes, 3rd stage	diazinon, endosulfan, caffeine	caffeine (0_58)	diazinon (0_4), caffeine (0_4)
Diisobutyl phthalate	helophytes, sedimentation	pendimethalin, lenacil, estradiol, mineral oils, metamitron, pirimiphos methyl, phantolid, musk xylene	estradiol (0_58), diisobutyl phthalate (0_45), lenacil (0_45), mineral oils (0_45, 0_55)	estradiol (0_4), mineral oil (0_2)
Chrysene	sedimentation, 3rd stage	PAHs (benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, indeno(1,2,3-cd)pyrene, benz(a)anthracene, benzo(ghi)perylene, fluoranthene and pyrene)	PAHs (0_58)	PAHs (0_4)
Butylbenzyl phthalate	helophytes	tri-(2-chloroisopropyl) phosphates a, b and c, nonylphenol and gemfibrozil	Butylbenzyl phthalate (0_58)	Butylbenzyl phthalate (0_2), gemfibrozil (0_2), nonylphenol (0_4)
Tri-phenyl phosphate	3rd stage	DEET, diuron, white mineral oils	white mineral oils (0_45 and 0_55)	diuron (0_4),white mineral oils (0_2 and 0_4)

4.5 Discussion

4.5.1 Systems biology approach highlights the cumulative effect of low-risk chemicals

The analysis of chemical risk demonstrated that Waterharmonica remediation decreases the concentrations of high-risk chemicals. However, this simple evaluation depends on known PNECs which are computational values based on acute exposures and might underestimate the real risks, due to chronic exposures, bioaccumulation and mixture effects in real systems. The integration of chemical concentrations with transcriptomics suggested that even low-risk chemicals might affect gene expression. This effect might be due to one class of chemicals with same targets, with each chemical having low risk individually, but together as a class of chemicals, their cumulative risk might be higher. One such group of chemicals in the static network analysis was PAHs. The Bayesian model further highlight the importance of PAH removal (as chrysene was one of the chemicals that gene expression depended on in the model and many other PAH concentrations were correlated with chrysene). Genes of cytochrome P450 family were also affected in all transitions, but especially transition 2 (H2 vs H3 – before and after helophyte fields) where PPI values were > 0.99 for all representatives of cytochrome P450 family. This is supported by the fact that especially PAHs are biodegraded by organisms symbiotic to the reeds (reviewed in [133, 227]).

Interestingly, from the toxicity tests (algae, MTT, Daphnia), algae TU EC50 was the only one correlated with the modules containing most chemicals. It has been shown that effects on algae are more prominent than on Daphnia or fish [272]. The results of network analysis support this and might even suggests that algae could be a more suitable species for monitoring certain chemicals. However, as many of the chemicals in these modules, in addition to PAHs, were herbicides, the effect on algae might be explained by the nature of these chemicals to target plant-specific pathways, such as photosystem II. Moreover, the same systems has also been shown to be sensitive to metals [176], which is especially

interesting as the modules correlated with toxicity to algae were also enriched in Gene Ontology term "response to cadmium ion". Chemicals correlated with genes in modules linked to survival of algae included several with high risk. For example, chlorpyrifos has been shown to affect algae [12]. There were many PAHs correlated with gene expression in both male and female stickleback, in the modules that are correlated with toxicity to algae. One example is fluoranthene which has been shown to be accumulated and biodegraded in algae [144]. Moreover, the same article suggests that the biodegradation is more effective when there is a mixture of PAHs. Other authors have reported that bioremediation of PAHs by bacteria can be antagonistic or synergistic dependant on mixture [137].

4.5.2 Can systems biology and transcriptomics reveal effects of chemicals which were not measured?

Gene Ontology analysis showed enrichment for the term "Response to cadmium ion" in modules of most chemicals in both male and female stickleback models. Cadmium or any other metals were not measured by the passive samplers as the aim was to detect organic chemicals, but they have been shown to be toxic to algae [74]. Moreover, this result suggests that cadmium, or other chemicals with similar effects, might be in the effluent. In fact, this hypothesis could be further supported by mapping known targets of cadmium into the network.

4.5.3 Stickleback growth and phthalates

We showed that there are network modules for both male and female sticklebacks that are correlated with fish length, weight and condition factor. Although these modules were not correlated with as many chemicals as other modules (F_0_4 and M_0_58), there were phthalates in module 0_2 and also aldicarb, which has risk > 1. Additionally, module F_0_2 was enriched in Gene Ontology term "Xenobiotic metabolism". In fish, phthalate exposure has been linked to growth inhibition [357, 29]. Exposure to phthalates

has also been linked to growth hormone levels and growth in humans [148, 32]. The Gene Ontology term "xenobiotic metabolism" includes several Cytochrome P450 genes which encode proteins involved in the metabolism of xenobiotics. These genes have also been linked to chemical exposure, including exposure to PAHs [33]. Associations between phthalates, thyroid hormone and Cytochrome P450 have been shown in the literature. For example, thyroid hormone, in addition to being linked to growth hormone and growth, also regulates Cytochrome P450 genes and vice versa, cytochrome P450 is also involved in thyroid hormone homeostasis [43], therefore it is reasonable to hypothesize that different phthalate concentrations in the remediation system perturbed thyroid metabolism or cytochrome P450 metabolism in sticklebacks. These affected systems might have also perturbed growth hormone levels and through this, growth of sticklebacks. Interestingly, the module (M_0_41_75) in males which is enriched for "xenobiotic metabolism" does not contain weight, length or any physiological parameters, but is correlated with chemicals. This might indicate that the chemical responses are sex-specific. Other possibility is that the gene expression effects in the modules linked to growth are not because of chemicals but because of other characteristics of wastewater effluent, such as microbial composition of the remediated water. Indeed, both modules correlated with weight, length and condition factor are enriched in GO Biological Process term "antigen processing and presentation". The module of female stickleback is also enriched in GO Biological Process term "defence response to bacterium". Additionally, Module M_0_42 in the network of male sticklebacks is enriched for GO Biological Process terms "cell cycle", "DNA replication initiation" and "spindle organization", "cell division", suggesting that the reduced size might be caused by perturbed mitosis.

4.5.4 Detailed Bayesian model shows triclosan and polycyclic aromatic hydrocarbons as main contributors of gene expression changes

Bayesian model ran with representatives of several chemical clusters has shown that some chemicals have more effect on gene expression. The results further supported the conclusion of static network analysis, showing that especially PAHs affect gene expression, as chrysene, a representative of the cluster of highly correlated PAHs, was shown to affect the gene expression during the transitions. Another chemical shown to be important in most transitions was triclosan. Bayesian model highlighted some genes that change during all remediation steps: ESR1 and VTG1. According to the CTD database [72, 71], ESR1 is affected by triclosan , butylbenzyl phthalate, diisobutyl phthalate, estrone and chrysene – all of these chemicals were found to have effect according to the model. VTG1 and VTG2 are affected by triclosan and estrone according to the CTD database [72, 71].

4.5.5 Data-driven approach has helped interpreting a complex process

We have shown the potential for data-driven approach for interpreting changes during water remediation by constructed wetlands. By integrating several types of data, we have shown that even chemicals with low or unknown risk might have effect on the stickleback and more importantly, that many high-and low risk chemicals correlated with gene expression in both male and female sticklebacks decrease during remediation, especially PAHs. The integration of chemical concentrations, morphology and gene expression has linked several pathways for which support exists in the literature, suggesting that phthalate levels in the remediation system might be linked to growth by perturbed cytochrome P450 or thyroid metabolism. This might show the importance for analysing mixture effects, as phthalates had low risk in our analysis, creating a hypothesis that the gene expression changes associated with growth of sticklebacks might be due to additive or synergistic effects, either by phthalates or in combination with other chemicals. This highlights the

limitations of assessing water quality by quantifying single chemicals and comparing these to PNEC or other constant which has been estimated based on acute exposure to single chemicals. As static networks have shown the importance of sex-differences and although many of the chemicals were associated with gene expression in both male and female sticklebacks, it would be useful to make new Bayesian model for male and female sticklebacks separately and also with a selection of chemicals from the static networks. As the static model has shown that phthalates might be linked to growth, it would be particularly useful to make a small model integrating the cumulative risk of all PAHs the expression of selected genes and also growth parameters. Additionally, instead of using individual genes, principal components of all KEGG pathways could be used to aid biological interpretation and find out which pathways are most affected during the consecutive stages of remediation.

CHAPTER 5

GENERAL DISCUSSION

5.1 Systems biology approach is able to connect molecular changes to physiology

Individual studies on three different organisms in this thesis have shown the potential of systems biology to connect environmental stress, high-throughput omics data and organism's physiological endpoints. I have shown that this approach is able to generate hypotheses which would need to be tested. A recently published study connecting transcriptomics with physiology and using additional information on potential molecular targets has revealed novel endocrine disruptors in largemouth bass [25], further showing the potential of systems biology to generate biological knowledge potentially relevant for future environmental risk assessments.

Currently, the effects of chemicals are mainly studied by either measuring traditional endpoints or analysing omics data on its own, most often by enrichment analysis of differentially expressed genes.

These examples, when considered in the context of the Adverse Outcome Pathway (AOP) framework can represent either only the Adverse Outcome associated with chemical stressor and in the latter case, might connect chemical stressor with Molecular Initiating Event or Key Events, if the hypothesis resulting from the omics studies is confirmed and results in a mechanistic knowledge about the effects of the chemical. However, connecting

these two parts of the Adverse Outcome Pathway, would be more informative and systems biology methods integrating chemical stressor, various omics data and physiological endpoints could help achieving this aim, by providing hypotheses which, when confirmed experimentally, has the potential to contribute to AOP pathways. The importance of using the AOP framework to connect molecular measurements with apical endpoints has also been highlighted in a vision for using omics for environmental monitoring [214].

Of course, for systems biology approaches to be used for integrating all relevant available data in future studies, with the aim of connecting the different parts of AOP, more user-friendly technical tools would need to be implemented with Graphical User Interface (GUI) instead of the need to use multiple R packeges [260], custom scripts and multiple apps of Cytoscape [285]. Additionally, after several approaches have been studied and published, the sharing of workflows, such as Taverna [349], could also be used for environmental systems biology, establishing a standard analysis strategy for many studies contributing to the generation of knowledge. As systems biology methods have the potential to connect environmental and omics data with physiological endpoints, the resulting hypotheses, once tested, can lead to the development of mode AOP-s. Assays can then be developed to target Key Events in the Adverse Outcome Pathway and based on these, the Adverse Outcomes can be predicted, reducing the need for animal testing. For example, AOP-based strategy has been demonstrated in zebrafish, predicting the effect of chemicals on thyroid function [300].

5.2 Challenges in future for risk assessment: timing, exposure length and computational resources

The study of mussel in this thesis has demonstrated that timing is important for the prediction of gender and especially for chemicals that have the potential to affect processes associated with reproductive cycles, the effects might be seen only at certain times of the year. This might mean that some exposures, which have been reported to result in no

effects have been done during a period where these effects are not seen. Another factor to consider is the exposure length. Currently, many chemicals are tested during acute exposure, whereas in the environmental scenario, organisms are exposed chronically and it is not easy to predict, which developmental stage is most crucial for the effects to appear later. It has even been shown that in addition for the effects to be seen in the organism that was exposed, the effects can become visible in the next generation. For example, in medaka, it has been shown that low-dose parental bisphenol A exposure can affect brain development during the embryonic and larval stage and also cause behavioural changes of the larvae [153]. In humans it has been shown that prenatal phthalate exposure is associated with the development of eczema [294].

Therefore, in the future, risk assessment could take into account different times in the reproductive cycle and should consider the effects of chronic exposure.

Technically, to test every chemical of chemical class for the full reproductive cycle would be expensive and against the effort to reduce animal-testing. However, if there was a strategy to develop AOPs for key species that can be used to extrapolate the effects on others, for chemicals which are representative of different mechanism of action, this could be informative and could be used in the models of other species and chemicals. Ideally, as the number of AOP-s grow, there might also be tools available that are able to model chronic exposures and exposures to mixtures. Of course, these developments depend heavily on available data and efforts have already been started to generate resources such as LINCS database and data portal [171] and the Connectivity Map [182, 305] making it possible to connect chemical perturbations to transcriptional signatures for human cell lines. Fish connectivity map has also been developed, linking chemical stressors with transcriptomics profiles [339]. In the AOP community, efforts to integrate different types of data, especially a wide range of databases, with AOP-s has already started with the launch of AOP-DB [250]. These resources and developments, for both human and environmental species provide many opportunities to advance the current state of AOPs.

5.3 Potential of systems biology for early warning signs of environmental stress

As previously seen, when certain methods are mature enough in fields of better knowledge, annotation and experimental control, more challenging problems can be approached with similar methods. For example, after sequencing the human genome and model species, many non-model species were also sequenced. Similarly, systems biology methods have been first demonstrated on either simpler and more easily testable species (*E. coli*, *C. elegans*) or for humans, which although complex, are motivated by understanding human diseases. It has been proposed that in the future, using both healthy population baseline and personal baseline of different markers identified from complex omics studies, human disease can be predicted and prevented [185, 160] and the authors suggest that this kind of personalised approach with complex machine learning can be used for many scenarios. Similarly, as individual (human) profiles are to be used as predictors of disease, maybe there is enough knowledge about the baseline "healthy" ecosystem based on different indicator species, taking into account also seasonal and reproductive systems and omics profiles. At the same time, maybe there is also similar data of "unhealthy" ecosystems, *i.e.* places with higher pollution, where effects are already seen at organism or population level. Using regular monitoring at informative times of the year, such as gonadal development, the shifts of omics profiles from previously healthy ecosystems towards more polluted places can have predictive power before changes might become visible at organism of population levels, similarly to the proposed disease prediction in humans. The importance using available knowledge and continuous baseline monitoring to determine "normal" state of an ecosystem and investigating further when the profile shifts outside the normal range has also been suggested in the context of the environment [214].

Of course, continuous monitoring in humans or environmental species is not cost-effective now, but similarly to the time when the human genome was first sequenced and it seemed unlikely this could be done for thousands of people later, it is possible that in the future, longitudinal omics profiling might be used in real life as early sign of later events,

provided that computational methods and resources are able to manage the amounts of data.

5.4 Systems toxicology and human health

In this thesis, I have demonstrated how dynamical models can relate omics data with environmental parameters and physiological endpoints, creating hypotheses that can be tested and if confirmed, could provide new knowledge for the definition of adverse outcome pathways.

However, the approach here is not only relevant for environmental pollution concerning non-model species such as fish, mussels or earthworms. The demonstration that systems biology integrating omics, physiology and environmental parameters can reveal testable hypotheses could potentially also be used for human studies.

In fact, an Adverse Outcome Pathway describing how chemical exposure can lead to Adverse Outcomes in humans has been used in the OEDC test guidelines for an in vitro skin sensitisation assay [234]. Human exposomics [333] has also already started as a field, where human omics data is integrated with exposome and health parameters. In addition to studies designed for exposomics, there is also potential in the current rise of population-based biobanks to be used for integrating omics data with health and environmental parameters. For example, once there are sufficient numbers of people who have various omics profiles, it is possible to integrate this data with environmental and exposure data, such as smoking status or living in an area with air pollution. Or course, due to life choices available to humans, these studies are more complex, but also provide ways of relating pollution to diseases, especially if electronic health records are also linked to the population biobanks. For example the use of health records and ^{1}H-NMR data has been used for finding markers for predicting all-cause mortality [98]. More recently, air pollution has been associated with incident cardiovascular disease using biobank cohorts from Norway and the UK [50]. The first of these examples has connected molecular

markers with an adverse outcome and the second has connected environmental pollution with and adverse outcome. Ultimately, there might be studies where environmental effects are connected to molecular markers of exposure and these in turn are linked to health outcomes.

These efforts are in a way similar to the assessment of Adverse Drug Reactions [190] and in the future, the frameworks assessing the effects of environmental chemicals on humans or other organisms might be similar to the ones used for evaluating the side effects of medicines.

Moreover, for some cancers, like breast cancer, it has been shown that in addition to genetic risks of an individual and their lifestyle, exposure to persistent organic pollutants might also contribute to the development of cancer [112, 113]. Therefore, it is possible, that in the future, serum concentrations of various pollutants are also incorporated in predictive models of diseases, together with genetic risk scores [174] and other factors currently used, such as age, sex and body mass index [180, 197].

5.5 Conclusions

I have shown the potential of data-driven systems biology in creating models integrating omics data with environmental and physiological parameters. For all three organisms, earthworm, mussel and stickleback, the resulting model provided biologically meaningful results. These results have the potential to contribute to the development of AOPs if experimentally validated.

LIST OF REFERENCES

[1] N. Abe and V. Cavalli. Nerve injury signaling. *Current Opinion in Neurobiology*, 18(3):276–283, 2008.

[2] H. Abusamra. A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Computer Science*, 23:5–14, 2013.

[3] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198, 2003.

[4] Agilent. Method of producing oligonucleotide arrays with features of high purity. *United States Patent*, 2000.

[5] V. Alexandar, P.G. Nayar, R. Murugesan, S. Shajahan, J. Krishnan, and S.S.S.J. Ahmed. A systems biology and proteomics-based approach identifies SRC and VEGFA as biomarkers in risk factor mediated coronary heart disease. *Molecular BioSystems*, 12(8):2594–2604, 2016.

[6] G.T. Ankley, R.S. Bennett, R.J. Erickson, D.J. Hoff, M.W. Hornung, R.D. Johnson, D.R. Mount, J.W. Nichols, C.L. Russom, P.K. Schmieder, et al. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, 29(3):730–741, 2010.

[7] P. Antczak, T. A White, A. Giri, F. Michelangeli, M.R. Viant, M.T.D. Cronin, C. Vulpe, and F. Falciani. Systems biology approach reveals a calcium-dependent mechanism for basal toxicity in *Daphnia magna*. *Environmental Science & Technology*, 49(18):11132–11140, 2015.

[8] I. Apraiz, J. Mi, and S. Cristobal. Identification of proteomic signatures of exposure to marine pollutants in mussels (*Mytilus edulis*). *Molecular & Cellular Proteomics*, 5(7):1274–1285, 2006.

[9] A. Arukwe, F.R. Knudsen, and A. Goksøyr. Fish zona radiata (eggshell) protein: a sensitive biomarker for environmental estrogens. *Environmental Health Perspectives*, 105(4):418–422, 1997.

[10] A. Asai, J. Qiu, Y. Narita, S. Chi, N. Saito, N. Shinoura, H. Hamada, Y. Kuchino, and T. Kirino. High level calcineurin activity predisposes neuronal cells to apoptosis. *Journal of Biological Chemistry*, 274(48):34450–34458, 1999.

[11] V. Asensio, P. Kille, A.J. Morgan, M. Soto, and I. Marigomez. Metallothionein expression and Neutral Red uptake as biomarkers of metal exposure and effect in *Eisenia fetida* and *Lumbricus terrestris* exposed to Cd. *European Journal of Soil Biology*, 43:S233–S238, 2007.

[12] V. Asselborn, C. Fernández, Y. Zalocar, and E.R. Parodi. Effects of chlorpyrifos on the growth and ultrastructure of green algae, *Ankistrodesmus gracilis*. *Ecotoxicology and Environmental Safety*, 120:334–341, 2015.

[13] D. Aygun, Z. Doganay, L. Altintop, H. Guven, M. Onar, T. Deniz, and T. Sunter. Serum acetylcholinesterase and prognosis of acute organophosphate poisoning. *Journal of Toxicology: Clinical Toxicology*, 40(7):903–910, 2002.

[14] E. Bååth, Å. Frostegård, and H. Fritze. Soil bacterial biomass, activity, phospholipid fatty acid pattern, and pH tolerance in an area polluted with alkaline dust deposition. *Applied and Environmental Microbiology*, 58(12):4026–4031, 1992.

[15] M. Bachelot, Z. Li, D. Munaron, P. Le Gall, C. Casellas, H. Fenet, and E. Gomez. Organic UV filter concentrations in marine mussels from French coastal regions. *Science of the Total Environment*, 420:273–279, 2012.

[16] A. Badiou, M. Meled, and L.P. Belzunces. Honeybee *Apis mellifera* acetylcholinesterase – a biomarker to detect deltamethrin exposure. *Ecotoxicology and Environmental Safety*, 69(2):246–253, 2008.

[17] N.V.E. Bagazgoïtia, H.D. Bailey, L. Orsi, B. Lacour, L. Guerrini-Rousseau, A. Bertozzi, P. Leblond, C. Faure-Conter, et al. Maternal residential pesticide use during pregnancy and risk of malignant childhood brain tumors: A pooled analysis of the ESCALE and ESTELLE studies (SFCE). *International Journal of Cancer*, 142(3):489–497, 2018.

[18] P.A. Bahamonde, A. Feswick, M.A. Isaacs, K.R. Munkittrick, and C.J. Martyniuk. Defining the role of omics in assessing ecosystem health: Perspectives from the Canadian environmental monitoring program. *Environmental Toxicology and Chemistry*, 35(1):20–35, 2016.

[19] M. Banni, A. Negri, F. Mignone, H. Boussetta, A. Viarengo, and F. Dondero. Gene expression rhythms in the mussel *Mytilus galloprovincialis* (Lam.) across an annual cycle. *PloS One*, 6(5):e18904, 2011.

[20] A. Barabási. Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413, 2009.

[21] A. Barabási and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[22] A. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101, 2004.

[23] J. Baršienė, A. Rybakovas, G. Garnaga, and L. Andreikėnaitė. Environmental genotoxicity and cytotoxicity studies in mussels before and after an oil spill at the marine oil terminal in the Baltic Sea. *Environmental Monitoring and Assessment*, 184(4):2067–2078, 2012.

[24] S.L. Bartelt-Hunt, D.D. Snow, T. Damon, J. Shockley, and K. Hoagland. The occurrence of illicit and therapeutic pharmaceuticals in wastewater effluent and surface waters in Nebraska. *Environmental Pollution*, 157(3):786–791, 2009.

[25] D. Basili, J. Zhang, J. Herbert, K. Kroll, N.D. Denslow, C.J. Martyniuk, F. Falciani, and P Antczak. In silico computational transcriptomics reveals novel endocrine disruptors in largemouth bass (*Micropterus salmoides*). *Environmental Science & Technology*, 52(13):7553–7565, 2018.

[26] A.J. Bednarska, M. Choczyński, R. Laskowski, and M. Walczak. Combined effects of chlorpyriphos, copper and temperature on acetylcholinesterase activity and toxicokinetics of the chemicals in the earthworm *Eisenia fetida*. *Environmental Pollution*, 220:567–576, 2017.

[27] A. Ben-Hur, H.T. Siegelmann, D. Horn, and V. Vapnik. A support vector clustering method. In *Proceedings 15th International Conference on Pattern Recognition*, page 2724. IEEE, 2000.

[28] J.P Berninger, D. Martinović-Weigelt, N. Garcia-Reyero, L. Escalon, E.J. Perkins, G.T. Ankley, and D.L. Villeneuve. Using transcriptomic tools to evaluate biological effects across effluent gradients at a diverse set of study sites in Minnesota, USA. *Environmental Science & Technology*, 48(4):2404–2412, 2014.

[29] H. Bhatia, A. Kumar, J.C. Chapman, and M.J. McLaughlin. Long-term exposures to di-n-butyl phthalate inhibit body growth and impair gonad development in juvenile Murray rainbowfish (*Melanotaenia fluviatilis*). *Journal of Applied Toxicology*, 35(7):806–816, 2015.

[30] A.D. Biales, M.S. Kostich, A.L. Batt, M.J. See, R.W. Flick, D.A. Gordon, J.M. Lazorchak, and D.C. Bencic. Initial development of a multigene 'omics-based exposure biomarker for pyrethroid pesticides. *Aquatic Toxicology*, 179:27–35, 2016.

[31] J.P. Bignell, M.J. Dodge, S.W. Feist, B. Lyons, P.D. Martin, N.G.H. Taylor, D. Stone, L. Travalent, and G.D. Stentiford. Mussel histopathology: effects of season, disease and species. *Aquatic Biology*, 2(1):1–15, 2008.

[32] M. Boas, H. Frederiksen, U. Feldt-Rasmussen, N.E. Skakkebæk, L. Hegedüs, L. Hilsted, A. Juul, and K.M. Main. Childhood exposure to phthalates: associations with thyroid function, insulin-like growth factor I, and growth. *Environmental Health Perspectives*, 118(10):1458–1464, 2010.

[33] S. Bogovski, B. Sergeyev, V. Muzyka, and S. Karlova. Cytochrome P450 system and heme synthesis enzymes activity in flounder liver as biomarkers of marine environments pollution. *Marine Environmental Research*, 46(1-5):13–16, 1998.

[34] M. Boulais, P. Soudant, N. Le Goïc, C. Quéré, P. Boudry, and M. Suquet. ATP content and viability of spermatozoa drive variability of fertilization success in the Pacific oyster (*Crassostrea gigas*). *Aquaculture*, 479:114–119, 2017.

[35] M. Bourgin, B. Beck, M. Boehler, E. Borowska, J. Fleiner, E. Salhi, R. Teichler, U. Von Gunten, et al. Evaluation of a full-scale wastewater treatment plant upgraded with ozonation and biological post-treatments: Abatement of micropollutants, formation of transformation products and oxidation by-products. *Water Research*, 129:486–498, 2018.

[36] N.G. Bowery and T.G. Smart. GABA and glycine as neurotransmitters: a brief history. *British Journal of Pharmacology*, 147(S1):S109–S119, 2006.

[37] R.A. Branch and E. Jacqz. Is carbaryl as safe as its reputation? Does it have a potential for causing chronic neurotoxicity in humans? *The American Journal of Medicine*, 80(4):659–664, 1986.

[38] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, et al. Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics*, 29(4):365, 2001.

[39] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[40] E.K. Brockmeier, G. Hodges, T.H. Hutchinson, E. Butler, M. Hecker, K. E. Tollefsen, N. Garcia-Reyero, P. Kille, et al. The role of omics in the application of adverse outcome pathways for chemical risk assessment. *Toxicological Sciences*, 158(2):252–262, 2017.

[41] E.K. Brockmeier, P.D. Scott, N.D. Denslow, and F.D.L. Leusch. Transcriptomic and physiological changes in eastern mosquitofish (*Gambusia holbrooki*) after exposure to progestins and anti-progestagens. *Aquatic Toxicology*, 179:8–17, 2016.

[42] R.D. Brook, S. Rajagopalan, C.A. III Pope, J.R. Brook, A. Bhatnagar, A.V. Diez-Roux, F. Holguin, Y. Hong, et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation*, 121(21):2331–2378, 2010.

[43] J. Brtko and Z. Dvorak. Role of retinoids, rexinoids and thyroid hormone in the expression of cytochrome P450 enzymes. *Current Drug Metabolism*, 12(2):71–88, 2011.

[44] G.W. Bryan, P.E. Gibbs, L.G. Hummerstone, and G.R. Burt. The decline of the gastropod *Nucella lapillus* around South-West England: evidence for the effect of tributyltin from antifouling paints. *Journal of the Marine Biological Association of the United Kingdom*, 66(3):611–640, 1986.

[45] T.D. Bucheli and K. Fent. Induction of cytochrome P450 as a biomarker for environmental contamination in aquatic ecosystems. *Critical Reviews in Environmental Science and Technology*, 25(3):201–268, 1995.

[46] A. Bugrim, T. Nikolskaya, and Y. Nikolsky. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discovery Today*, 9(3):127–135, 2004.

[47] J.G. Bundy, J.K. Sidhu, F. Rana, D.J. Spurgeon, C. Svendsen, J.F. Wren, S.R. Stürzenbaum, A.J. Morgan, and P. Kille. 'Systems toxicology' approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm *Lumbricus rubellus*. *BMC Biology*, 6(1):25, 2008.

[48] G. Burnstock. Historical review: ATP as a neurotransmitter. *Trends in Pharmacological Sciences*, 27(3):166–176, 2006.

[49] D. Butina, M.D. Segall, and K. Frankcombe. Predicting ADME properties in silico: methods and models. *Drug Discovery Today*, 7(11):S83–S88, 2002.

[50] Y. Cai, S. Hodgson, M. Blangiardo, J. Gulliver, D. Morley, D. Fecht, D. Vienneau, K. de Hoogh, et al. Road traffic noise, air pollution and incident cardiovascular disease: A joint analysis of the HUNT, EPIC-Oxford and UK Biobank cohorts. *Environment International*, 114:191–201, 2018.

[51] A. Calisi, N. Zaccarelli, M.G. Lionetto, and T. Schettino. Integrated biomarker analysis in the earthworm *Lumbricus terrestris*: application to the monitoring of soil heavy metal pollution. *Chemosphere*, 90(11):2637–2644, 2013.

[52] P. Carvalho, C. Arias, and H. Brix. Constructed wetlands for water treatment: new developments. *Water*, 9(6), 2017.

[53] A. Cassese, M. Guindani, P. Antczak, F. Falciani, and M. Vannucci. A Bayesian model for the identification of differentially expressed genes in *Daphnia magna* exposed to munition pollutants. *Biometrics*, 71(3):803–811, 2015.

[54] J. Chen, M. Saleem, C. Wang, W. Liang, and Q. Zhang. Individual and combined effects of herbicide tribenuron-methyl and fungicide tebuconazole on soil earthworm *Eisenia fetida*. *Scientific Reports*, 8(1):2967, 2018.

[55] A. Chiffre, F. Degiorgi, A. Buleté, L. Spinner, and P. Badot. Occurrence of pharmaceuticals in WWTP effluents and their impact in a karstic rural catchment of Eastern France. *Environmental Science and Pollution Research*, 23(24):25427–25441, 2016.

[56] M.A. Cikutovic, L.C. Fitzpatrick, A.J. Goven, B.J. Venables, M.A. Giggleman, and E.L. Cooper. Wound healing in earthworms *Lumbricus terrestris*: a cellular-based biomarker for assessing sublethal chemical toxicity. *Bulletin of Environmental Contamination and Toxicology*, 62(4):508–514, 1999.

[57] C. Colinas, E. Ingham, and R. Molina. Population responses of target and non-target forest soil organisms to selected biocides. *Soil Biology and Biochemistry*, 26(1):41–47, 1994.

[58] European Commission. Technical guidance document on risk assessment in support of Commission Directive 93/67/EEC on risk assessment for new notified substances Commission regulation (EC) No 1488/94 on risk assessment for Existing Substances Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. *Ispra (IT): European Commission Joint Research Centre. EUR*, 20418, 2003.

[59] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M Talón, and M. Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

[60] K.M. Connor and A.Y. Gracey. Circadian cycles are the dominant transcriptional rhythm in the intertidal mussel *Mytilus californianus*. *Proceedings of the National Academy of Sciences*, 108(38):16110–16115, 2011.

[61] K.M. Connor and A.Y. Gracey. High resolution analysis of metabolic cycles in the intertidal mussel *Mytilus californianus*. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 302(1):R103–R111, 2012.

[62] Gene Ontology Consortium. Gene Ontology annotations and resources. *Nucleic Acids Research*, 41(D1):D530–D535, 2012.

[63] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[64] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 41(D1):D8–D20, 2012.

[65] M.G. Coppolino, M.J. Woodside, N. Demaurex, S. Grinstein, R. St-Arnaud, and S. Dedhar. Calreticulin is essential for integrin-mediated calcium signalling and cell adhesion. *Nature*, 386(6627):843, 1997.

[66] M.A. Cousin. Synaptic vesicle endocytosis. *Molecular Neurobiology*, 22(1-3):115–128, 2000.

[67] A.M. Cowie, K.I. Sarty, A. Mercer, J. Koh, K.A. Kidd, and C.J. Martyniuk. Molecular networks related to the immune system and mitochondria are targets for the pesticide dieldrin in the zebrafish (*Danio rerio*) central nervous system. *Journal of Proteomics*, 157:71–82, 2017.

[68] Y. Dai, R. Tao, Y. Tai, N.F. Tam, A. Dan, and Y. Yang. Application of a full-scale newly developed stacked constructed wetland and an assembled bio-filter for reducing phenolic endocrine disrupting chemicals from secondary effluent. *Ecological Engineering*, 99:496–503, 2017.

[69] P. Dalgaard. Introductory statistics with R, 2002.

[70] A. Dan, D. Fujii, S. Soda, T. Machimura, and M. Ike. Removal of phenol, bisphenol A, and 4-tert-butylphenol from synthetic landfill leachate by vertical flow constructed wetlands. *Science of the Total Environment*, 578:566–576, 2017.

[71] A.P. Davis, C.J. Grondin, R.J. Johnson, D. Sciaky, B.L. King, R. McMorran, J. Wiegers, T.C. Wiegers, and C.J. Mattingly. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Research*, 45(D1):D972–D978, 2016.

[72] A.P. Davis, C.G. Murphy, C.A. Saraceni-Richards, M.C. Rosenstein, T.C. Wiegers, and C.J. Mattingly. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Research*, 37(suppl_1):D786–D792, 2008.

[73] A. De los Ríos, L. Pérez, M. Ortiz-Zarragoitia, T. Serrano, M.C. Barbero, B. Echavarri-Erasun, J.A. Juanes, A. Orbea, and M.P. Cajaraville. Assessing the effects of treated and untreated urban discharges to estuarine and coastal waters applying selected biomarkers on caged mussels. *Marine Pollution Bulletin*, 77(1-2):251–265, 2013.

[74] B. Debelius, J. M Forja, Á. DelValls, and L.M. Lubián. Toxicity and bioaccumulation of copper and lead in five marine microalgae. *Ecotoxicology and Environmental Safety*, 72(5):1503–1513, 2009.

[75] K. Dettmer, P.A. Aronov, and B.D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, 2007.

[76] N.M. Dheilly, C. Lelong, A. Huvet, K. Kellner, M. Dubos, G. Riviere, P. Boudry, and P. Favrel. Gametogenesis in the Pacific oyster *Crassostrea gigas*: a microarrays-based analysis identifies sex and stage specific genes. *PloS One*, 7(5):e36353, 2012.

[77] Water Framework Directive. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. *Official Journal of the European Communities*, 22(2000):L327, 2000.

[78] J. Dittman and T.A. Ryan. Molecular circuitry of endocytosis at nerve terminals. *Annual Review of Cell and Developmental*, 25:133–160, 2009.

[79] F. Dominici, R.D. Peng, M.L. Bell, L. Pham, A. McDermott, S.L. Zeger, and J.M. Samet. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10):1127–1134, 2006.

[80] F. Dondero, A. Negri, L. Boatti, F. Marsano, F. Mignone, and A. Viarengo. Transcriptomic and proteomic effects of a neonicotinoid insecticide mixture in the marine mussel (*Mytilus galloprovincialis*, Lam.). *Science of the Total Environment*, 408(18):3775–3786, 2010.

[81] J.W. Doran. Soil health and global sustainability: translating science into practice. *Agriculture, Ecosystems & Environment*, 88(2):119–127, 2002.

[82] J.W. Doran and M.R. Zeiss. Soil health and sustainability: managing the biotic component of soil quality. *Applied Soil Ecology*, 15(1):3–11, 2000.

[83] M. Douziech, I.R. Conesa, A. Benítez-López, A. Franco, M. Huijbregts, and R. van Zelm. Quantifying variability in removal efficiencies of chemicals in activated sludge wastewater treatment plants–a meta-analytical approach. *Environmental Science: Processes & Impacts*, 20(1):171–182, 2018.

[84] W.B. Dunn and D.I. Ellis. Metabolomics: current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24(4):285–294, 2005.

[85] ECHA. REACH, 2007. https://echa.europa.eu/regulations/reach/understanding-reach; Accessed June 8, 2018.

[86] ECHA. 21551 chemicals on EU market now registered, 2018. https://echa.europa.eu/-/21-551-chemicals-on-eu-market-now-registered; Accessed: June 8th, 2018.

[87] ECHA. OECD and EU test guidelines, 2018. https://echa.europa.eu/support/oecd-eu-test-guidelines; Accessed June 8th, 2018.

[88] R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[89] R. Ekblom and J. Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1, 2011.

[90] S. Ekins, Y. Nikolsky, and T. Nikolskaya. Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends in Pharmacological Sciences*, 26(4):202–209, 2005.

[91] R.P. Ellis, J.I. Spicer, J.J. Byrne, U. Sommer, M.R. Viant, D.A. White, and S. Widdicombe. 1H NMR metabolomics reveals contrasting response by male and female mussels exposed to reduced seawater pH, increased temperature, and a pathogen. *Environmental Science & Technology*, 48(12):7044–7052, 2014.

[92] A.J. Enright, S. Van Dongen, and C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[93] European Union. Regulation (EC) No 1907/2006 - Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), 2006. https://osha.europa.eu/en/legislation/directives/regulation-ec-no-1907-2006-of-the-european-parliament-and-of-the-council; Accessed June 8th 2018.

[94] Eurostat. The REACH baseline study 5 years update, 2012.

[95] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, and T.S. Gardner. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):e8, 2007.

[96] D. Fattorini, A. Notti, R. Di Mento, A.M. Cicero, M. Gabellini, A. Russo, and F. Regoli. Seasonal, spatial and inter-annual variations of trace metals in mussels from the Adriatic sea: a regional gradient for arsenic and implications for monitoring the impact of off-shore activities. *Chemosphere*, 72(10):1524–1533, 2008.

[97] O. Fiehn, D. Robertson, J. Griffin, M. van der Werf, B. Nikolau, N. Morrison, L.W. Sumner, R. Goodacre, N.W. Hardy, C. Taylor, et al. The metabolomics standards initiative (MSI). *Metabolomics*, 3(3):175–178, 2007.

[98] K. Fischer, J. Kettunen, P. Würtz, T. Haller, A.S. Havulinna, A.J. Kangas, P. Soininen, T. Esko, M. Tammesoo, R. Mägi, et al. Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. *PLoS Medicine*, 11(2):e1001606, 2014.

[99] F. Fonnum. Glutamate: a neurotransmitter in mammalian brain. *Journal of Neurochemistry*, 42(1):1–11, 1984.

[100] D.M. Fry. Reproductive effects in birds exposed to pesticides and industrial chemicals. *Environmental Health Perspectives*, 103(suppl 7):165–171, 1995.

[101] A. Fujiwara, Y. Kamata, K. Asami, and I. Yasumasu. Relationship between ATP level and respiratory rate in sea urchin embryos. *Development, Growth & Differentiation*, 42(2):155–165, 2000.

[102] F. Gagné, B. Bouchard, C. André, E. Farcy, and M. Fournier. Evidence of feminization in wild *Elliptio complanata* mussels in the receiving waters downstream of a municipal effluent outfall. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 153(1):99–106, 2011.

[103] N. Gambi, A. Pasteris, and E. Fabbri. Acetylcholinesterase activity in the earthworm *Eisenia andrei* at different conditions of carbaryl exposure. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 145(4):678–685, 2007.

[104] N. Garcia-Reyero, B.L. Escalon, P. Loh, J.G. Laird, A.J. Kennedy, B. Berger, and E.J. Perkins. Assessment of chemical mixtures and groundwater effects on *Daphnia magna* transcriptomics. *Environmental Science & Technology*, 46(1):42–50, 2011.

[105] N. Garcia-Reyero, T. Habib, M. Pirooznia, K.A. Gust, P. Gong, C. Warner, M. Wilbanks, and E. Perkins. Conserved toxic responses across divergent phylogenetic lineages: a meta-analysis of the neurotoxic effects of RDX among multiple species using toxicogenomics. *Ecotoxicology*, 20(3):580, 2011.

[106] N. Garcia-Reyero and E.J. Perkins. Systems biology: leading the revolution in ecotoxicology. *Environmental Toxicology and Chemistry*, 30(2):265–273, 2011.

[107] G. García-Santos and K. Keller-Forrer. Avoidance behaviour of *Eisenia fetida* to carbofuran, chlorpyrifos, mancozeb and metamidophos in natural soils from the highlands of Colombia. *Chemosphere*, 84(5):651–656, 2011.

[108] E. Gauthier, I. Fortier, F. Courchesne, P. Pepin, J. Mortimer, and D. Gauvreau. Environmental pesticide exposure as a risk factor for Alzheimer's disease: a case-control study. *Environmental Research*, 86(1):37–45, 2001.

[109] F. Geoghegan, I. Katsiadaki, T.D. Williams, and J.K. Chipman. A cDNA microarray for the three-spined stickleback, *Gasterosteus aculeatus* L., and analysis of the interactive effects of oestradiol and dibenzanthracene exposures. *Journal of Fish Biology*, 72(9):2133–2153, 2008.

[110] A.C. Gerecke, M. Schärer, H.P. Singer, S.R. Müller, R.P. Schwarzenbach, M. Sägesser, U. Ochsenbein, and G. Popow. Sources of pesticides in surface waters in Switzerland: pesticide load through waste water treatment plants—-current situation and reduction potential. *Chemosphere*, 48(3):307–315, 2002.

[111] D. Gershon. Microarray technology: an array of opportunities. *Nature*, 416(6883):885, 2002.

[112] M. Ghisari, H. Eiberg, M. Long, and E.C. Bonefeld-Jørgensen. Polymorphisms in Phase I and Phase II genes and breast cancer risk and relations to persistent organic pollutant exposure: a case–control study in Inuit women. *Environmental Health*, 13(1):19, 2014.

[113] M. Ghisari, M. Long, D.M. Røge, J. Olsen, and E.C. Bonefeld-Jørgensen. Polymorphism in xenobiotic and estrogen metabolizing genes, exposure to perfluorinated compounds and subsequent breast cancer risk: a nested case-control study in the Danish National Birth Cohort. *Environmental Research*, 154:325–333, 2017.

[114] E. Giarratano, M.N. Gil, and G. Malanga. Seasonal and pollution-induced variations in biomarkers of transplanted mussels within the Beagle Channel. *Marine Pollution Bulletin*, 62(6):1337–1344, 2011.

[115] M. Giltrap, J. Ronan, S. Hardenberg, G. Parkes, B. McHugh, E. McGovern, and J.G. Wilson. Assessment of biomarkers in *Mytilus edulis* to determine good environmental status for implementation of MSFD in Ireland. *Marine Pollution Bulletin*, 71(1-2):240–249, 2013.

[116] L. Gipperth. The legal design of the international and European Union ban on tributyltin antifouling paint: direct and indirect effects. *Journal of Environmental Management*, 90:S86–S95, 2009.

[117] D.J. Goldberg, S.T. Green, D. Nathwani, J. McMenamin, N. Hamlet, and D.H. Kennedy. RDX intoxication causing seizures and a widespread petechial rash mimicking meningococcaemia. *Journal of the Royal Society of Medicine*, 85(3):181, 1992.

[118] E.D. Goldberg. The mussel watch: a first step in global marine monitoring. *Marine Pollution Bulletin*, 6:111–114, 1975.

[119] J. A. Goldstein, L.A. Bastarache, J.C. Denny, J.M. Pulley, and D.M Aronoff. PregOMICS – Leveraging systems biology and bioinformatics for drug repurposing in maternal–child health. *American Journal of Reproductive Immunology*, 80(2):e12971, 2018.

[120] L. Gong, J. Wu, S. Zhou, Y. Wang, J. Qin, B. Yu, X. Gu, and C. Yao. Global analysis of transcriptome in dorsal root ganglia following peripheral nerve injury in rats. *Biochemical and Biophysical Research Communications*, 478(1):206–212, 2016.

[121] P. Gong, X. Guan, L.S. Inouye, Y. Deng, M. Pirooznia, and E.J. Perkins. Transcriptomic analysis of RDX and TNT interactive sublethal effects in the earthworm *Eisenia fetida*. *BMC Genomics*, 9(1):S15, 2008.

[122] P. Gong, X. Guan, M. Pirooznia, C. Liang, and E.J. Perkins. Gene expression analysis of CL-20-induced reversible neurotoxicity reveals GABAA receptors as potential targets in the earthworm *Eisenia fetida*. *Environmental Science & Technology*, 46(2):1223–1232, 2012.

[123] P. Gong, L.S. Inouye, and E.J. Perkins. Comparative neurotoxicity of two energetic compounds, hexanitrohexaazaisowurtzitane and hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine, in the earthworm *Eisenia fetida*. *Environmental Toxicology and Chemistry*, 26(5):954–959, 2007.

[124] P. Gong and E.J. Perkins. Earthworm toxicogenomics: A renewed genome-wide quest for novel biomarkers and mechanistic insights. *Applied Soil Ecology*, 104:12–24, 2016.

[125] P. Gong, M. Pirooznia, X. Guan, and E.J. Perkins. Design, validation and annotation of transcriptome-wide oligonucleotide probes for the oligochaete annelid *Eisenia fetida*. *PloS One*, 5(12):e14266, 2010.

[126] S. Gorbi, C.V. Lamberti, A. Notti, M. Benedetti, D. Fattorini, G. Moltedo, and F. Regoli. An ecotoxicological protocol with caged mussels, *Mytilus galloprovincialis*,

for monitoring the impact of an offshore platform in the Adriatic sea. *Marine Environmental Research*, 65(1):34–49, 2008.

[127] S. Götz, J.M. García-Gómez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, M. Robles, M. Talón, J. Dopazo, and A. Conesa. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10):3420–3435, 2008.

[128] N.W. Green and J. Knutzen. Organohalogens and metals in marine fish and mussels and some relationships to biological variables at reference localities in Norway. *Marine Pollution Bulletin*, 46(3):362–374, 2003.

[129] T.F. Grilo and R. Rosa. Intersexuality in aquatic invertebrates: prevalence and causes. *Science of the Total Environment*, 592:714–728, 2017.

[130] R. Gupta, A. Stincone, P. Antczak, S. Durant, R. Bicknell, A. Bikfalvi, and F. Falciani. A computational framework for gene regulatory network inference that combines multiple methods and datasets. *BMC Systems Biology*, 5(1):52, 2011.

[131] E. Hahlbeck, I. Katsiadaki, I. Mayer, M. Adolfsson-Erici, J. James, and B. Bengtsson. The juvenile three-spined stickleback (*Gasterosteus aculeatus* L.) as a model organism for endocrine disruption II-kidney hypertrophy, vitellogenin and spiggin induction. *Aquatic Toxicology*, 70(4):311–326, 2004.

[132] S. Hardinger. Proton nuclear magnetic resonance spectroscopy (H-NMR), http:www.chem.ucla.edu/~harding/notes/notes_14C_nmr02.pdf, accessed June 1, 2018.

[133] A.K. Haritash and C.P. Kaushik. Biodegradation aspects of polycyclic aromatic hydrocarbons (PAHs): a review. *Journal of Hazardous Materials*, 169(1-3):1–15, 2009.

[134] K. Haug, R.M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated metadata. *Nucleic Acids Research*, 41(D1):D781–D786, 2012.

[135] J. Hellou and R.J. Law. Stress on stress response of wild mussels, *Mytilus edulis* and *Mytilus trossulus*, as an indicator of ecosystem health. *Environmental Pollution*, 126(3):407–416, 2003.

[136] P. Hennersdorf, S. Kleinertz, S. Theisen, M.A. Abdul-Aziz, G. Mrotzek, H.W. Palm, and H.P. Saluz. Microbial diversity and parasitic load in tropical fish of different environmental conditions. *PloS One*, 11(3):e0151594, 2016.

[137] C.T. Hennessee and Q.X. Li. Effects of polycyclic aromatic hydrocarbon mixtures on degradation, gene expression, and metabolite production in four *Mycobacterium species*. *Applied and Environmental Microbiology*, 82(11):3357–3369, 2016.

[138] T.B. Henry, J.T. McPherson, E.D. Rogers, T.P. Heah, S.A. Hawkins, A.C. Layton, and G.S. Sayler. Changes in the relative expression pattern of multiple vitellogenin genes in adult male and larval zebrafish exposed to exogenous estrogens. *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative physiology*, 154(1):119–126, 2009.

[139] E.S. Herald. Effects of DDT-oil solutions upon amphibians and reptiles. *Herpetologica*, 5(6):117–120, 1949.

[140] A. Hines, F.J. Staff, J. Widdows, R.M. Compton, F. Falciani, and M.R. Viant. Discovery of metabolic signatures for predicting whole organism toxicology. *Toxicological Sciences*, 115(2):369–378, 2010.

[141] A. Hines, W.H. Yeung, J. Craft, M. Brown, J. Kennedy, J. Bignell, G.D. Stentiford, and M.R. Viant. Comparison of histological, genetic, metabolomics, and lipid-based methods for sex determination in marine mussels. *Analytical Biochemistry*, 369(2):175–186, 2007.

[142] I.D. Hodkinson and J.K. Jackson. Terrestrial and aquatic invertebrates as bioindicators for environmental monitoring, with particular reference to mountain ecosystems. *Environmental Management*, 35(5):649–666, 2005.

[143] G. Hoek, B. Brunekreef, P. Fischer, and J. van Wijnen. The association between air pollution and heart failure, arrhythmia, embolism, thrombosis, and other cardiovascular causes of death in a time series study. *Epidemiology*, 12(3):355–357, 2001.

[144] Y. Hong, D. Yuan, Q. Lin, and T. Yang. Accumulation and biodegradation of phenanthrene and fluoranthene by the algae enriched from a mangrove aquatic ecosystem. *Marine Pollution Bulletin*, 56(8):1400–1405, 2008.

[145] L. Hood. Systems biology: new opportunities arising from genomics, proteomics and beyond. In *Experimental Hematology*, volume 26, pages 681–681, 1998.

[146] Leroy Hood, James R Heath, Michael E Phelps, and Biaoyang Lin. Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696):640–643, 2004.

[147] D.A. Hosack, G. Dennis, B.T. Sherman, H.C. Lane, and R.A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10):R70, 2003.

[148] T. Huan, E.M. Forsberg, D. Rinehart, C.H. Johnson, J. Ivanisevic, H.P. Benton, M. Fang, A. Aisporna, B. Hilmers, F.L. Poole, et al. Systems biology guided by XCMS Online metabolomics. *Nature Methods*, 14(5):461, 2017.

[149] D.W. Huang, B.T. Sherman, and R.A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44, 2008.

[150] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.

[151] J.L. Hurley-Sanders, J.F. Levine, S.A.C. Nelson, J.M. Law, W.J. Showers, and M.K. Stoskopf. Key metabolites in tissue extracts of *Elliptio complanata* identified using 1 H nuclear magnetic resonance spectroscopy. *Conservation Physiology*, 3(1), 2015.

[152] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372, 2001.

[153] T. Inagaki, N.L. Smith, K.M. Sherva, and S. Ramakrishnan. Cross-generational effects of parental low dose BPA exposure on the Gonadotropin-Releasing Hormone3 system and larval behavior in medaka (*Oryzias latipes*). *Neurotoxicology*, 57:163–173, 2016.

[154] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[155] C.R. Jiménez, F.J. Stam, K.W. Li, Y. Gouwenberg, M.P. Hornshaw, F. De Winter, J. Verhaagen, and A.B. Smit. Proteomics of the injured rat sciatic nerve reveals protein expression dynamics during regeneration. *Molecular & Cellular Proteomics*, 4(2):120–132, 2005.

[156] C. Kampichler, R. Wieland, S. Calmé, H. Weissenberger, and S. Arriaga-Weiss. Classification in conservation biology: a comparison of five machine-learning methods. *Ecological Informatics*, 5(6):441–450, 2010.

[157] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab–an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

[158] I. Katsiadaki, A.P. Scott, and I. Mayer. The potential of the three-spined stickleback (*Gasterosteus aculeatus* L.) as a combined biomarker for oestrogens and androgens in European waters. *Marine Environmental Research*, 54(3-5):725–728, 2002.

[159] I. Katsiadaki, T.D. Williams, J.S. Ball, T.P. Bean, M.B. Sanders, H. Wu, E.M. Santos, M.M. Brown, et al. Hepatic transcriptomic and metabolomic responses in the Stickleback (*Gasterosteus aculeatus*) exposed to ethinyl-estradiol. *Aquatic Toxicology*, 97(3):174–187, 2010.

[160] R.A. Kellogg, J. Dunn, and M.P. Snyder. Personal omics for precision health. *Circulation Research*, 122(9):1169–1171, 2018.

[161] S.C. Kessler, E.J. Tiedeken, K.L. Simcock, S. Derveau, J. Mitchell, S. Softley, A. Radcliffe, J.C. Stout, and G.A. Wright. Bees prefer foods containing neonicotinoid pesticides. *Nature*, 521(7550):74, 2015.

[162] K. Kim, E. Kabir, and S.A. Jahan. Exposure to pesticides and the associated human health effects. *Science of the Total Environment*, 575:525–535, 2017.

[163] Un-Jung Kim, Jung Keun Oh, and Kurunthachalam Kannan. Occurrence, removal, and environmental emission of organophosphate flame retardants/plasticizers in a wastewater treatment plant in New York State. *Environmental Science & Technology*, 51(14):7872–7880, 2017.

[164] R.J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011, 2011.

[165] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[166] A.K. Kivaisi. The potential for constructed wetlands for wastewater treatment and reuse in developing countries: a review. *Ecological Engineering*, 16(4):545–560, 2001.

[167] P.J. Kloepper-Sams and J.W. Owens. Environmental biomarkers as indicators of chemical exposure. *Journal of Hazardous Materials*, 35(2):283–294, 1993.

[168] T. Knigge, T. Monsinjon, and O. Andersen. Surface-enhanced laser desorption/ionization-time of flight-mass spectrometry approach to biomarker discovery in blue mussels (*Mytilus edulis*) exposed to polyaromatic hydrocarbons and heavy metals under field conditions. *Proteomics*, 4(9):2722–2727, 2004.

[169] M. Köck-Schulmeyer, M. Villagrasa, M.L. de Alda, R. Céspedes-Sánchez, F. Ventura, and D. Barceló. Occurrence and behavior of pesticides in wastewater treatment plants and their environmental impact. *Science of the Total Environment*, 458:466–476, 2013.

[170] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y.A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, et al. ArrayExpress update – simplifying data submissions. *Nucleic Acids Research*, 43(D1):D1113–D1116, 2014.

[171] A. Koleti, R. Terryn, V. Stathias, C. Chung, D.J. Cooper, J.P. Turner, D. Vidović, M. Forlin, et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Research*, 46(D1):D558–D566, 2017.

[172] A. Krämer, J. Green, J. Pollard Jr., and S. Tugendreich. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4):523–530, 2013.

[173] V.J. Kramer, M.A. Etterson, M. Hecker, C.A. Murphy, G. Roesijadi, D.J. Spade, J.A. Spromberg, M. Wang, and G.T. Ankley. Adverse outcome pathways and ecological risk assessment: Bridging to population-level effects. *Environmental Toxicology and Chemistry*, 30(1):64–76, 2011.

[174] N.T. Krarup, A. Borglykke, K.H. Allin, C.H. Sandholt, J.M. Justesen, E.A. Andersson, N. Grarup, T. Jørgensen, O. Pedersen, and T. Hansen. A genetic risk score of 45 coronary artery disease risk variants associates with increased risk of myocardial infarction in 6041 Danish individuals. *Atherosclerosis*, 240(2):305–310, 2015.

[175] P. Krzeminski, C. Schwermer, A. Wennberg, K. Langford, and C. Vogelsang. Occurrence of UV filters, fragrances and organophosphate flame retardants in municipal WWTP effluents and their removal during membrane post-treatment. *Journal of Hazardous Materials*, 323:166–176, 2017.

[176] K.S. Kumar, H.U. Dahms, J.S. Lee, H.C. Kim, W.C. Lee, and K.H. Shin. Algal photosynthetic responses to toxic metals and herbicides assessed by chlorophyll a fluorescence. *Ecotoxicology and Environmental Safety*, 104(1):51–71, 2014.

[177] M. Kutmon, T. Kelder, P. Mandaviya, C.T.A. Evelo, and S.L. Coort. CyTargetLinker: a Cytoscape app to integrate regulatory interactions in network analysis. *PloS One*, 8(12):e82160, 2013.

[178] Y. Kwon, Y. Jung, J. Park, J. Seo, M. Choi, and G. Hwang. Characterizing the effect of heavy metal contamination on marine mussels using metabolomics. *Marine Pollution Bulletin*, 64(9):1874–1879, 2012.

[179] K.P. Lai, J.C. Lee, H.T Wan, J.W. Li, A.Y.M. Wong, T.F. Chan, C. Oger, J. Galano, T. Durand, K.S. Leung, et al. Effects of in utero PFOS exposure on transcriptome, lipidome, and function of mouse testis. *Environmental Science & Technology*, 51(15):8782–8794, 2017.

[180] K. Läll, R. Mägi, A. Morris, A. Metspalu, and K. Fischer. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics in Medicine*, 19(3):322, 2017.

[181] C.A. LaLone, D.L. Villeneuve, L.D. Burgoon, C.L. Russom, H.W. Helgen, J.P. Berninger, J.E. Tietge, M.N. Severson, J.E. Cavallin, and G.T. Ankley. Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk of chemicals with known modes of action. *Aquatic Toxicology*, 144:141–154, 2013.

[182] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J. Lerner, J. Brunet, A. Subramanian, K.N. Ross, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006.

[183] C.J. Langdon, T.G. Piearce, A.A. Meharg, and K.T. Semple. Survival and behaviour of the earthworms *Lumbricus rubellus* and *Dendrodrilus rubidus* from arsenate-contaminated and non-contaminated sites. *Soil Biology and Biochemistry*, 33(9):1239–1244, 2001.

[184] G.R. Langley, I.M. Adcock, F. Busquet, K.M. Crofton, E. Csernok, C. Giese, T. Heinonen, K. Herrmann, M. Hofmann-Apitius, B. Landesmann, et al. Towards a 21st-century roadmap for biomedical research and drug discovery: consensus report and recommendations. *Drug Discovery Today*, 22(2):327–339, 2017.

[185] E. Lau and J.C. Wu. Omics, big data, and precision medicine in cardiovascular sciences. *Circulation Research*, 122(9):1165–1168, 2018.

[186] S. Lèbre, J. Becq, F. Devaux, M.P.H. Stumpf, and G Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(1):130, 2010.

[187] D. Lee, P. M. Lind, D.R. Jacobs, S. Salihovic, B. Van Bavel, and L. Lind. Polychlorinated biphenyls and organochlorine pesticides in plasma predict development of type 2 diabetes in the elderly: the prospective investigation of the vasculature in Uppsala Seniors (PIVUS) study. *Diabetes Care*, 34(8):1778–1784, 2011.

[188] D. Lee, M.W. Steffes, A. Sjödin, R.S. Jones, L.L. Needham, and D.R. Jacobs Jr. Low dose of some persistent organic pollutants predicts type 2 diabetes: a nested case–control study. *Environmental Health Perspectives*, 118(9):1235–1242, 2010.

[189] I. Lee, P. Eriksson, A. Fredriksson, S. Buratovic, and H. Viberg. Developmental neurotoxic effects of two pesticides: behavior and biomolecular studies on chlorpyrifos and carbaryl. *Toxicology and Applied Pharmacology*, 288(3):429–438, 2015.

[190] J. Lee, S.C. Ji, B. Kim, S. Yi, K.H. Shin, J.Y. Cho, K.S. Lim, S.H. Lee, S.H. Yoon, J.Y. Chung, et al. Exploration of biomarkers for amoxicillin/clavulanate-induced liver injury: multi-omics approaches. *Clinical and Translational Science*, 10(3):163–171, 2017.

[191] K.M.Y. Leung. Joining the dots between omics and environmental management. *Integrated Environmental Assessment and Management*, 14(2):169–173, 2018.

[192] H. Li. *Dynamic Bayesian networks for gene regulatory network reconstruction.* PhD thesis, The University of Southern Missisippi, 2013.

[193] M. Li, L. Qiu, L. Wang, W. Wang, L. Xin, Y. Li, Z. Liu, and L. Song. The inhibitory role of γ-aminobutyric acid (GABA) on immunomodulation of Pacific oyster *Crassostrea gigas*. *Fish & Shellfish Immunology*, 52:16–22, 2016.

[194] M. Li, L. Wang, L. Qiu, W. Wang, L. Xin, J. Xu, H. Wang, and L. Song. A glutamic acid decarboxylase (CgGAD) highly expressed in hemocytes of Pacific oyster *Crassostrea gigas*. *Developmental & Comparative Immunology*, 63:56–65, 2016.

[195] S. Li, S.M. Assmann, and R. Albert. Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. *PLoS Biology*, 4(10):e312, 2006.

[196] Y. Li, L. Zhang, Y. Sun, X. Ma, J. Wang, R. Li, M. Zhang, S. Wang, X. Hu, and Z. Bao. Transcriptome sequencing and comparative analysis of ovary and testis identifies potential key sex-related genes and pathways in scallop *Patinopecten yessoensis*. *Marine Biotechnology*, 18(4):453–465, 2016.

[197] J. Lindström and J. Tuomilehto. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care*, 26(3):725–731, 2003.

[198] M. Linxweiler, B. Schick, and R. Zimmermann. Let's talk about Secs: Sec61, Sec62 and Sec63 in signal transduction, oncology and personalized medicine. *Signal Transduction and Targeted Therapy*, 2:17002, 2017.

[199] X. Liu, Q. Song, Y. Tang, W. Li, J. Xu, J. Wu, F. Wang, and P.C. Brookes. Human health risk assessment of heavy metals in soil–vegetable system: a multi-medium analysis. *Science of the Total Environment*, 463:530–540, 2013.

[200] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675, 1996.

[201] B.L. Lockwood and G.N. Somero. Transcriptomic responses to salinity stress in invasive and native blue mussels (genus *Mytilus*). *Molecular Ecology*, 20(3):517–529, 2011.

[202] R. Loos, R. Carvalho, D.C. António, S. Comero, G. Locoro, S. Tavazzi, B. Paracchini, M. Ghiani, T. Lettieri, L. Blaha, et al. EU-wide monitoring survey on emerging polar organic contaminants in wastewater treatment plant effluents. *Water Research*, 47(17):6475–6487, 2013.

[203] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee. Transcriptomics technologies. *PLoS Computational Biology*, 13(5):e1005457, 2017.

[204] C. Ludwig, J.M. Easton, A. Lodi, S. Tiziani, S.E. Manzoor, A.D. Southam, J.J. Byrne, L.M. Bishop, et al. Birmingham Metabolite Library: a publicly accessible database of 1-D 1 H and 2-D 1 H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics*, 8(1):8–18, 2012.

[205] C. Ludwig and M.R. Viant. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques*, 21(1):22–32, 2010.

[206] T. Lukkari, M. Taavitsainen, M. Soimasuo, A. Oikari, and J. Haimi. Biomarker responses of the earthworm *Aporrectodea tuberculata* to copper and zinc exposure: differences between populations with and without earlier metal exposure. *Environmental Pollution*, 129(3):377–386, 2004.

[207] J.C. Madden, V. Rogiers, and M. Vinken. Application of in silico and in vitro methods in the development of adverse outcome pathway constructs in wildlife. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1656):20130584, 2014.

[208] S. Mahadevan, S.L. Shah, T.J. Marrie, and C.M. Slupsky. Analysis of metabolomic data using support vector machines. *Analytical Chemistry*, 80(19):7562–7570, 2008.

[209] D. Majeti, K. Kwon, M.E. Ahmed, B. Uzzi, E. Akleman, and I. Pavlidis. Scholar PLotR, https://www.csullender.com/scholar/, accessed Feb 6, 2018.

[210] D. Marbach, J.C. Costello, R. Küffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, A. Aderhold, R. Bonneau, Y. Chen, et al. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796, 2012.

[211] D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286–6291, 2010.

[212] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC Bioinformatics*, volume 7, page S7. BioMed Central, 2006.

[213] P. Marjan, C.J. Martyniuk, M.L.M. Fuzzen, D.L. MacLatchy, M.E. McMaster, and M.R. Servos. Returning to normal? Assessing transcriptome recovery over time in male rainbow darter (*Etheostoma caeruleum*) liver in response to wastewater-treatment plant upgrades. *Environmental Toxicology and Chemistry*, 36(8):2108–2122, 2017.

[214] C.J. Martyniuk. Are we closer to the vision? A proposed framework for incorporating omics into environmental assessments. *Environmental Toxicology and Pharmacology*, 59:87–93, 2018.

[215] P. Matthiessen and P.E. Gibbs. Critical appraisal of the evidence for tributyltin-mediated endocrine disruption in mollusks. *Environmental Toxicology and Chemistry*, 17(1):37–43, 1998.

[216] G. Mayer, H.J.L. Heerspink, C. Aschauer, A. Heinzel, G. Heinze, A. Kainz, J. Sunzenauer, P. Perco, et al. Systems biology–derived biomarkers to predict progression of renal function decline in type 2 diabetes. *Diabetes Care*, 40(3):391–397, 2017.

[217] S. McGinnis and T.L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(suppl_2):W20–W25, 2004.

[218] J. Meyer and K. Bester. Organophosphate flame retardants and plasticisers in wastewater treatment plants. *Journal of Environmental Monitoring*, 6(7):599–605, 2004.

[219] P.E. Meyer, F. Lafitte, and G. Bontempi. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9(1):461, 2008.

[220] S.A. Meyer, A.J. Marchand, J.L. Hight, G.H. Roberts, L.B. Escalon, L.S. Inouye, and D.K. MacMillan. Up-and-down procedure (UDP) determinations of acute oral toxicity of nitroso degradation products of hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine (RDX). *Journal of Applied Toxicology: An International Journal*, 25(5):427–434, 2005.

[221] C. Miege, J.M. Choubert, L. Ribeiro, M. Eusèbe, and M. Coquery. Fate of pharmaceuticals and personal care products in wastewater treatment plants–conception of a database and first results. *Environmental Pollution*, 157(5):1721–1726, 2009.

[222] V. Mommaerts, S. Reynders, J. Boulet, L. Besard, G. Sterk, and G. Smagghe. Risk assessment for side-effects of neonicotinoids against bumblebees with and without impairing foraging behavior. *Ecotoxicology*, 19(1):207, 2010.

[223] S. Mompelat, B. Le Bot, and O. Thomas. Occurrence and fate of pharmaceutical products and by-products, from resource to drinking water. *Environment International*, 35(5):803–814, 2009.

[224] J.H. Morris, L. Apeltsin, A.M. Newman, J. Baumbach, T. Wittkop, G. Su, G.D. Bader, and T.E. Ferrin. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, 12(1):436, Nov 2011.

[225] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621, 2008.

[226] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(1):S4, 2008.

[227] A. Muratova, T. Hübner, S. Tischer, O. Turkovskaya, M. Möder, and P. Kuschk. Plant–rhizosphere-microflora association during phytoremediation of PAH-contaminated soil. *International Journal of Phytoremediation*, 5(2):137–151, 2003.

[228] Nagalakshmi, U. and Wang, Z. and Waern, K. and Shou, C. and Raha, D. and Gerstein, M. and Snyder, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

[229] K. Nagasawa, H. Oouchi, N. Itoh, K.G. Takahashi, and M. Osada. In vivo administration of scallop GnRH-like peptide influences on gonad development in the Yesso scallop, *Patinopecten yessoensis*. *PloS One*, 10(6):e0129571, 2015.

[230] E.F. Neuhauser and C.A. Callahan. Growth and reproduction of the earthworm *Eisenia fetida* exposed to sublethal concentrations of organic chemicals. *Soil Biology and Biochemistry*, 22(2):175–179, 1990.

[231] J. Nielsen. Systems biology of metabolism: a driver for developing personalized and precision medicine. *Cell Metabolism*, 25(3):572–579, 2017.

[232] M.J.M. Notten, A.J.P. Oosthoek, J. Rozema, and R. Aerts. Heavy metal concentrations in a soil–plant–snail food chain along a terrestrial soil pollution gradient. *Environmental Pollution*, 138(1):178–190, 2005.

[233] P. Nuurai, S.F. Cummins, N.A. Botwright, and P. Sobhon. Characterization of an abalone gonadotropin-releasing hormone and its effect on ovarian cell proliferation. *Aquaculture*, 450:116–122, 2016.

[234] OECD. Test No. 442E: In Vitro Skin Sensitisation assays addressing the Key Event on activation of dendritic cells on the Adverse outcome pathway for Skin Sensitisation, 2017. http://www.oecd-ilibrary.org/environment/test-no-442e-in-vitro-skin-sensitisation_9789264264359-en; Accessed June 8th, 2018.

[235] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[236] A. Orbea, L. Garmendia, I. Marigómez, and M.P. Cajaraville. Effects of the 'Prestige' oil spill on cellular biomarkers in intertidal mussels: results of the first year of studies. *Marine Ecology Progress Series*, 306:177–189, 2006.

[237] P. Ostapczuk, J.D. Schladot, H. Emons, K. Oxynos, K. Schramm, G. Grimmer, and J. Jacob. Environmental monitoring and banking of marine pollutants by using common mussels. *Chemosphere*, 34(9-10):2143–2151, 1997.

[238] J. Owen, B.A. Hedley, C. Svendsen, J. Wren, M.J. Jonker, P.K. Hankard, L.J. Lister, S.R. Stürzenbaum, et al. Transcriptome profiling of developmental and xenobiotic responses in a keystone soil animal, the oligochaete annelid *Lumbricus rubellus*. *BMC Genomics*, 9(1):266, 2008.

[239] M.G. Paoletti. The role of earthworms for assessment of sustainability and as bioindicators. *Agriculture, Ecosystems & Environment*, 74(1-3):137–155, 1999.

[240] V.A. Papaevangelou, G.D. Gikas, V.A. Tsihrintzis, M. Antonopoulou, and I.K. Konstantinou. Removal of endocrine disrupting chemicals in HSF and VF pilot-scale constructed wetlands. *Chemical Engineering Journal*, 294:146–156, 2016.

[241] H. Parsons and M. Viant. Variance stabilising transformations for NMR metabolomics data. *BMC Systems Biology*, 1(Suppl 1):22, 2007.

[242] B.B. Peñalver, L. Cralle, and J.A. Gilbert. Systems biology of the human microbiome. *Current Opinion in Biotechnology*, 51:146–153, 2018.

[243] G. Pérès, F. Vandenbulcke, M. Guernion, M. Hedde, T. Beguiristain, F. Douay, S. Houot, D. Piron, et al. Earthworm indicators as tools for soil monitoring, characterization and risk assessment. An example from the national Bioindicator programme (France). *Pedobiologia*, 54:S77–S87, 2011.

[244] A. Peyvandipour, N. Saberian, A. Shafi, M. Donato, and S. Draghici. A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, 34(16):2817–2825, 2018.

[245] D. Pimentel, S. Cooperstein, H. Randell, D. Filiberto, S. Sorrentino, B. Kaye, C. Nicklin, J. Yagi, et al. Ecology of increasing diseases: Population growth and environmental degradation. *Human Ecology*, 35(6):653–668, 2007.

[246] R.C. Pires-Neto, A.J. Lichtenfels, S.R. Soares, M. Macchione, P.H.N. Saldiva, and M. Dolhnikoff. Effects of São Paulo air pollution on the upper airways of mice. *Environmental Research*, 101(3):356–361, 2006.

[247] M. Pirooznia, P. Gong, X. Guan, L.S. Inouye, K. Yang, E.J. Perkins, and Y. Deng. Cloning, analysis and functional annotation of expressed sequence tags from the earthworm *Eisenia fetida*. In *BMC Bioinformatics*, volume 8, page S7. BioMed Central, 2007.

[248] M. Pirooznia, J.Y. Yang, M.Q. Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, 9(1):S13, 2008.

[249] L.W. Pisa, V. Amaral-Rogers, L.P. Belzunces, J. Bonmatin, C.A. Downs, D. Goulson, D.P. Kreutzweiser, C. Krupke, et al. Effects of neonicotinoids and fipronil on non-target invertebrates. *Environmental Science and Pollution Research*, 22(1):68–102, 2015.

[250] M.E. Pittman, S.W. Edwards, C. Ives, and H.M. Mortensen. AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks. *Toxicology and Applied Pharmacology*, 343:71–83, 2018.

[251] K.S. Pollard and M.J. Van Der Laan. Bioconductor's hopach package. pages 1–7, 2010.

[252] V. Ponesakki, S. Paul, D.K.S. Mani, V. Rajendiran, P. Kanniah, and S. Sivasubramaniam. Annotation of nerve cord transcriptome in earthworm *Eisenia fetida*. *Genomics Data*, 14:91–105, 2017.

[253] R.D. Porter and S.N. Wiemeyer. Dieldrin and DDT: effects on sparrow hawk eggshells and reproduction. *Science*, 165(3889):199–200, 1969.

[254] C. Potting, T. Tatsuta, T. König, M. Haag, T. Wai, M.J. Aaltonen, and T. Langer. TRIAP1/PRELI complexes prevent apoptosis by mediating intramitochondrial transport of phosphatidic acid. *Cell Metabolism*, 18(2):287–295, 2013.

[255] J. Pougnet, E. Toulme, A. Martinez, D. Choquet, E. Hosy, and E. Boué-Grabot. ATP P2X receptors downregulate AMPA receptor trafficking and postsynaptic efficacy in hippocampal neurons. *Neuron*, 83(2):417–430, 2014.

[256] H.C. Poynton, W.E. Robinson, B.J. Blalock, and R.E. Hannigan. Correlation of transcriptomic responses and metal bioaccumulation in *Mytilus edulis* L. reveals early indicators of stress. *Aquatic Toxicology*, 155:129–141, 2014.

[257] A. Pujol, R. Mosca, J. Farrés, and P. Aloy. Unveiling the role of network and systems biology in drug discovery. *Trends in Pharmacological Sciences*, 31(3):115–123, 2010.

[258] D. Quercioli, A. Roli, E. Morandi, S. Perdichizzi, L. Polacchini, F. Rotondo, M. Vaccari, M. Villani, et al. The use of omics-based approaches in regulatory toxicology: An alternative approach to assess the no observed transcriptional effect level. *Microchemical Journal*, 136:143–148, 2018.

[259] QUIAGEN Inc. Ingenuity Pathway Analysis. analysis performed with Ingenuity Pathway Analysis (IPA) in 2011.

[260] R Core Team. R: A Language and Environment for Statistical Computing, 2017.

[261] J.V. Rao, Y.S. Pavan, and S.S. Madhavendra. Toxic effects of chlorpyrifos on morphology and acetylcholinesterase activity in the earthworm, *Eisenia foetida*. *Ecotoxicology and Environmental Safety*, 54(3):296–301, 2003.

[262] N.V. Reo. NMR-based metabolomics. *Drug and Chemical Toxicology*, 25(4):375–382, 2002.

[263] J.R. Richardson, A. Roy, S.L. Shalat, R.T. Von Stein, M.M. Hossain, B. Buckley, M. Gearing, A.I. Levey, and D.C. German. Elevated serum pesticide levels and risk for Alzheimer disease. *JAMA Neurology*, 71(3):284–290, 2014.

[264] H.U. Riisgård, P.P. Egede, and I.B. Saavedra. Feeding behaviour of the mussel, *Mytilus edulis*: new observations, with a minireview of current knowledge. *Journal of Marine Biology*, 2011, 2011.

[265] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, and G.K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[266] P. Y. Robidoux, J. Hawari, G. Bardai, L. Paquet, G. Ampleman, S. Thiboutot, and G. I. Sunahara. TNT, RDX, and HMX decrease earthworm (*Eisenia andrei*) life-cycle responses in a spiked natural forest soil. *Archives of Environmental Contamination and Toxicology*, 43(4):379–388, 2002.

[267] P.Y. Robidoux, C. Svendsen, J. Caumartin, J. Hawari, G. Ampleman, S. Thiboutot, J.M. Weeks, and G.I. Sunahara. Chronic toxicity of energetic compounds in soil determined using the earthworm (*Eisenia andrei*) reproduction test. *Environmental Toxicology and Chemistry: An International Journal*, 19(7):1764–1773, 2000.

[268] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[269] E. Roex, F. Smedes, H. Beeltje, and E. Foekema. The added value of passive sampling in determining the removal efficiency of organic micropollutants in constructed wetlands. Technical report, Waternet, Netherlands, 2015.

[270] W. Sanchez, I. Katsiadaki, B. Piccini, J. Ditche, and J. Porcher. Biomarker responses in wild three-spined stickleback (*Gasterosteus aculeatus* L.) as a useful tool for freshwater biomonitoring: a multiparametric approach. *Environment International*, 34(4):490–498, 2008.

[271] J.C. Sanchez-Hernandez and B.M. Sanchez. Lizard cholinesterases as biomarkers of pesticide exposure: enzymological characterization. *Environmental Toxicology and Chemistry*, 21(11):2319–2325, 2002.

[272] H. Sanderson, D.J. Johnson, C.J. Wilson, R.A. Brain, and K.R. Solomon. Probabilistic hazard assessment of environmentally occurring pharmaceuticals toxicity to fish, daphnids and algae by ECOSAR screening. *Toxicology Letters*, 144(3):383–395, 2003.

[273] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[274] E.M. Santos, J.S. Ball, T.D. Williams, H. Wu, F. Ortega, R. Van Aerle, I. Katsiadaki, F. Falciani, et al. Identifying health impacts of exposure to copper using transcriptomics and metabolomics in a fish model. *Environmental Science and Technology*, 44(2):820–826, 2010.

[275] A. Sarkar, D. Ray, A.N. Shrivastava, and S. Sarker. Molecular Biomarkers: Their significance and application in marine pollution monitoring. *Ecotoxicology*, 15(4):333–340, 2006.

[276] L.D. Scanlan, A.V. Loguinov, Q. Teng, P. Antczak, K.P. Dailey, D.T. Nowinski, J. Kornbluh, X.X. Lin, et al. Gene transcription, metabolite and lipid profiling in eco-indicator *Daphnia magna* indicate diverse mechanisms of toxicity by legacy and emerging flame-retardants. *Environmental Science and Technology*, 49(12):7400–7410, 2015.

[277] L.D. Scanlan, R.B. Reed, A.V. Loguinov, P. Antczak, A. Tagmount, S. Aloni, D.T. Nowinski, P. Luong, C. Tran, N. Karunaratne, et al. Silver nanowire exposure results in internalization and toxicity to *Daphnia magna*. *Acs Nano*, 7(12):10681–10694, 2013.

[278] M. Schena and D. Shalon. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[279] W.A. Jr. Schmitt, R.M. Raab, and G. Stephanopoulos. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Research*, 2(3):1654–1663, 2004.

[280] T.I. Schwarz, I. Katsiadaki, B.H. Maskrey, and A.P. Scott. Mussels (*Mytilus* spp.) display an ability for rapid and high capacity uptake of the vertebrate steroid, estradiol-17β from water. *The Journal of Steroid Biochemistry and Molecular Biology*, 165:407–420, 2017.

[281] A.P. Scott. Do mollusks use vertebrate sex steroids as reproductive hormones? Part I: Critical appraisal of the evidence for the presence, biosynthesis and uptake of steroids. *Steroids*, 77(13):1450–68, 11 2012.

[282] A.P. Scott. Do mollusks use vertebrate sex steroids as reproductive hormones? II. Critical review of the evidence that steroids have biological effects. *Steroids*, 78(2):268–81, 2 2013.

[283] A.P. Scott. Is there any value in measuring vertebrate steroids in invertebrates? *General and Comparative Endocrinology*, (April):0–1, 2018.

[284] J.J. Scott-Fordsmand, J.M. Weeks, and S.P. Hopkin. Toxicity of nickel to the earthworm and the applicability of the neutral red retention assay. *Ecotoxicology*, 7(5):291–295, 1998.

[285] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, (13):2498–2504, 2003.

[286] B. Shi, Z. Huang, X. Xiang, M. Huang, W. Wang, and C. Ke. Transcriptome analysis of the key role of GAT2 gene in the hyper-accumulation of copper in the oyster *Crassostrea angulata*. *Scientific Reports*, 5(May):17751, 2015.

[287] K. Smolarz, A. Hallmann, S. Zabrzańska, and A. Pietrasik. Elevated gonadal atresia as biomarker of endocrine disruptors: Field and experimental studies using *Mytilus trossulus* (L.) and 17-alpha ethinylestradiol (EE2). *Marine Pollution Bulletin*, 120(April):58–67, 2017.

[288] G.K. Smyth, J. Michaud, and H.S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005.

[289] G.K. Smyth, N. Thorne, and J. Wettenhall. Limma: linear models for microarray data user's guide. *Bioconductor*, 2003.

[290] E.M. Sogin, P. Anderson, P. Williams, C. Chen, and R.D. Gates. Application of 1H-NMR metabolomic profiling for reef-building corals. *PloS One*, 9(10):e111274, 2014.

[291] M. Solé, C. Porte, X. Biosca, C.L. Mitchelmore, J.K. Chipman, D.R. Livingstone, and J. Albaigés. Effects of the "Aegean Sea" oil spill on biotransformation enzymes, oxidative stress and DNA-adducts in digestive gland of the mussel (*Mytilus edulis* L.). *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology*, 113(2):257–265, 1996.

[292] Q. Song, H. Chen, Y. Li, H. Zhou, Q. Han, and X. Diao. Toxicological effects of benzo(a)pyrene, DDT and their mixture on the green mussel *Perna viridis* revealed by proteomic and metabolomic approaches. *Chemosphere*, 144:214–224, 2016.

[293] C. Sonne, P.S. Leifsson, R. Dietz, E.W. Born, R.J. Letcher, L. Hyldstrup, F.F. Riget, M. Kirkegaard, and D.C.G. Muir. Xenoendocrine pollutants may reduce size of sexual organs in East Greenland polar bears (*Ursus maritimus*). *Environmental Science and Technology*, 40(18):5668–5674, 2006.

[294] M.H. Soomro, N. Baiz, C. Philippat, C. Vernet, V. Siroux, Nichole M.C., S. Sanyal, R. Slama, C. Bornehag, and I. Annesi-Maesano. Prenatal exposure to phthalates and the development of eczema phenotypes in male children: results from the EDEN Mother–Child Cohort Study. *Environmental Health Perspectives*, 126(2):027002, 2018.

[295] D.J. Spurgeon, S.P. Hopkin, and D.T. Jones. Effects of cadmium, copper, lead and zinc on growth, reproduction and survival of the earthworm *Eisenia fetida* (Savigny): assessing the environmental impact of point-source metal contamination in terrestrial ecosystems. *Environmental Pollution*, 84(2):123–130, 1994.

[296] A. Statnikov, L. Wang, and C.F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(1):319, 2008.

[297] J. J. Steffan, E. C. Brevik, L. C. Burgess, and A. Cerdà. The effect of soil on human health: an overview. *European Journal of Soil Science*, 69(1):159–171, 2018.

[298] J. Stenersen. Action of pesticides on earthworms. Part I: The toxicity of cholinesterase-inhibiting insecticides to earthworms as evaluated by laboratory tests. *Pesticide Science*, 10(1):66–74, 1979.

[299] Stenerson, J. Uptake and metabolism of xenobiotics by earthworms. *Ecotoxicology of Earthworms*, pages 129–138, 1992.

[300] E. Stinckens, L. Vergauwen, G.T. Ankley, R. Blust, V.M. Darras, D.L. Villeneuve, H. Witters, D.C. Volz, and D. Knapen. An AOP-based alternative testing strategy to predict the impact of thyroid hormone disruption on swim bladder inflation in zebrafish. *Aquatic Toxicology*, 200(April):1–12, 2018.

[301] G. Stolovitzky, D. Monroe, and A. Califano. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, 1115(1):1–22, 2007.

[302] S.J. Sturla, A.R. Boobis, R.E. Fitzgerald, J. Hoeng, R.J. Kavlock, K. Schirmer, M. Whelan, M.F. Wilks, and M.C. Peitsch. Systems toxicology: From basic research to risk assessment. *Chemical Research in Toxicology*, 27(3):314–329, 2014.

[303] S.R. Stürzenbaum, P. Kille, and A.J. Morgan. Identification of heavy metal induced changes in the expression patterns of the translationally controlled tumour protein (TCTP) in the earthworm *Lumbricus rubellus*. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1398(3):294–304, 1998.

[304] G. Su, A. Kuchinsky, J.H. Morris, D.J. States, and F. Meng. GLay: Community structure analysis of biological networks. *Bioinformatics*, 26(24):3135–3137, 2010.

[305] A. Subramanian, R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, J. Gould, J.F. Davis, A.A. Tubelli, J.K. Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[306] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44(D1):D463–D470, 2016.

[307] T. Sun, X. Wu, J. Xu, B.D. McNeil, Z.P. Pang, W. Yang, L. Bai, S. Qadri, J.D. Molkentin, D.T. Yue, et al. The role of calcium/calmodulin-activated calcineurin in rapid and slow endocytosis at central synapses. *Journal of Neuroscience*, 30(35):11838–11847, 2010.

[308] C. Svendsen, J. Owen, P. Kille, J. Wren, M.J. Jonker, B.A. Headley, A.J. Morgan, M. Blaxter, et al. Comparative transcriptomic responses to chronic cadmium, fluoranthene, and atrazine exposure in *Lumbricus rubellus*. *Environmental Science & Technology*, 42(11):4208–4214, 2008.

[309] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N.T. Doncheva, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, page gkw937, 2016.

[310] G.R. Tetreault, C.J. Bennett, K. Shires, B. Knight, M.R. Servos, and M.E. McMaster. Intersex and reproductive impairment of wild fish exposed to multiple municipal wastewater discharges. *Aquatic Toxicology*, 104(3-4):278–290, 2011.

[311] ThermoFisherScientific. GeneChip Microarrays. Activity #3 - Manufacturing of GeneChip Microarrays and Building Models, 2005. https://tools.thermofisher.com/content/sfs/brochures/activity3_manufacturing_background.p accessed June 8th, 2018.

[312] R. Tibshirani, G. Chu, B. Narasimhan, and J. Li. samr: SAM: Significance analysis of microarrays. *CRAN*, 2, 2011.

[313] G. Tini, L. Marchetti, C. Priami, and M. Scott-Boyer. Multi-omics integration – a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics*, (January):1–11, 2017.

[314] D.R. Tobergte and S. Curtis. Continuation of residue reviews. *Reviews of Environmental Contamination and Toxicology*, 188(9):1–30, 2013.

[315] L. Tomanek. Environmental proteomics of the mussel *Mytilus*: implications for tolerance to stress and change in limits of biogeographic ranges in response to climate change. *Integrative and Comparative Biology*, 52(5):648–64, 11 2012.

[316] A.F. Toro-Vélez, C.A. Madera-Parra, M.R. Peña-Varón, W.Y. Lee, J.C. Bezares-Cruz, W.S. Walker, H. Cárdenas-Henao, S. Quesada-Calderón, et al. BPA and NP removal from municipal wastewater by tropical horizontal subsurface constructed wetlands. *Science of the Total Environment*, 542:93–101, 2016.

[317] V. Trevino and F. Falciani. GALGO: An R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, 22(9):1154–1156, 2006.

[318] E. Tuck and V. Cavalli. Roles of membrane trafficking in nerve repair and regeneration. *Communicative & Integrative Biology*, 3(3):209–214, 2010.

[319] W. Tuffnail, G.A. Mills, P. Cary, and R. Greenwood. An environmental ^{1}H NMR metabolomic study of the exposure of the marine mussel *Mytilus edulis* to atrazine, lindane, hypoxia and starvation. *Metabolomics*, 5(1):33–43, 2009.

[320] R. Turja, A. Soirinsuo, H. Budzinski, M.H. Devier, and K.K. Lehtonen. Biomarker responses and accumulation of hazardous substances in mussels (*Mytilus trossulus*) transplanted along a pollution gradient close to an oil terminal in the Gulf of Finland (Baltic Sea). *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 157(1):80–92, 2013.

[321] V. Turusov, V. Rakitsky, and L. Tomatis. Dichlorodiphenyltrichloroethane (DDT): ubiquity, persistence, and risks. *Environmental Health Perspectives*, 110(2):125–128, 2002.

[322] M. Udovic and D. Lestan. *Eisenia fetida* avoidance behavior as a tool for assessing the efficiency of remediation of Pb, Zn and Cd polluted soil. *Environmental Pollution*, 158(8):2766–2772, 2010.

[323] United States Congress. Federal Water Pollution Control Act. page 234, 2002.

[324] U.S.E.P.A. Guidelines for Ecological Risk Assessment (EPA/630/R-95/002F), 1998. https://www.epa.gov/sites/production/files/2014-11/documents/eco_risk_assessment1998.pdf, Accessed June8, 2018.

[325] H. Uwizeyimana, M. Wang, W. Chen, and K. Khan. The eco-toxic effects of pesticide and heavy metal mixtures towards earthworms in soil. *Environmental Toxicology and Pharmacology*, 55:20–29, 2017.

[326] G. Van Aggelen, G.T. Ankley, W.S. Baldwin, D.W. Bearden, W.H. Benson, J.K. Chipman, T.W. Collette, J.A. Craft, et al. Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. *Environmental Health Perspectives*, 118(1):1–5, 2009.

[327] M.J. Van der Laan and K.S. Pollard. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference*, 117(2):275–303, 2003.

[328] C.A.M. Van Gestel, W.A. Van Dis, E.M. Van Breemen, and P.M. Sparenburg. Development of a standardized reproduction toxicity test with the earthworm species *Eisenia fetida andrei* using copper, pentachlorophenol, and 2, 4-dichloroaniline. *Ecotoxicology and Environmental Safety*, 18(3):305–312, 1989.

[329] T. Vandenbrouck, A. Soetaert, K. van der Ven, R. Blust, and W. De Coen. Nickel and binary metal mixture responses in *Daphnia magna*: molecular fingerprints and (sub) organismal effects. *Aquatic Toxicology*, 92(1):18–29, 2009.

[330] N. Verboven, J. Verreault, R.J. Letcher, G.W. Gabrielsen, and N. P. Evans. Nest temperature and parental behaviour of Arctic-breeding glaucous gulls exposed to persistent organic pollutants. *Animal Behaviour*, 77(2):411–418, 2009.

[331] D.L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T.H. Hutchinson, C.A. LaLone, B. Landesmann, T. Lettieri, et al. Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicological Sciences*, 142(2):312–320, 2014.

[332] D.L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T.H. Hutchinson, C.A. LaLone, B. Landesmann, T. Lettieri, S. Munn, M. Nepelska, et al. Adverse outcome pathway development II: best practices. *Toxicological Sciences*, 142(2):321–330, 2014.

[333] P. Vineis, M. Chadeau-Hyam, H. Gmuender, J. Gulliver, Z. Herceg, J. Kleinjans, M. Kogevinas, S. Kyrtopoulos, et al. The exposome in practice: Design of the EXPOsOMICS project. *International Journal of Hygiene and Environmental Health*, 220(2):142–151, 2017.

[334] C. Vogelsang, M. Grung, T.G. Jantsch, K.E. Tollefsen, and H. Liltved. Occurrence and removal of selected organic micropollutants at mechanical, chemical and advanced wastewater treatment plants in Norway. *Water Research*, 40(19):3559–3570, 2006.

[335] C. Von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(suppl_1):D433–D437, 2005.

[336] C. J. Vörösmarty, P. B. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S. E. Bunn, et al. Global threats to human water security and river biodiversity. *Nature*, 467(7315):555–561, 2010.

[337] C.J. Vörösmarty, P. Green, J. Salisbury, and R.B. Lammers. Global water resources: vulnerability from climate change and population growth. *Science*, 289(5477):284–288, 2000.

[338] N. Voulvoulis, K.D. Arpon, and T. Giakoumis. The EU Water Framework Directive: From great expectations to problems with implementation. *Science of the Total Environment*, 575:358–366, 2017.

[339] D. Wang, L. Yang, P. Zhang, J. LaBaer, H. Hermjakob, D. Li, and X. Yu. AAgAtlas 1.0: a human autoantigen database. *Nucleic Acids Research*, page gkw946, 2016.

[340] Y.P. Wang, J.Y. Shi, H. Wang, Q. Lin, X.C. Chen, and Y.X. Chen. The influence of soil heavy metals pollution on soil microbial biomass, enzyme activity, and community composition near a copper smelter. *Ecotoxicology and Environmental Safety*, 67(1):75–81, 2007.

[341] G.R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. Huber, A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, et al. gplots: Various R programming tools for plotting data. *CRAN R package*, 2(4):1, 2009.

[342] M. Watanabe, A. Fukuda, and J. Nabekura. The role of GABA in the regulation of GnRH neurons. *Frontiers in Neuroscience*, 8(November):1–9, 2014.

[343] M. Watanabe, K.A. Meyer, T.M. Jackson, T.B. Schock, W.E. Johnson, and D.W. Bearden. Application of NMR-based metabolomics for environmental assessment in the Great Lakes using zebra mussel *(Dreissena polymorpha)*. *Metabolomics*, 11(5):1302–1315, 2015.

[344] M.R. Wenk. The emerging field of lipidomics. *Nature Reviews Drug Discovery*, 4(7):594–610, 2005.

[345] P.R. Whitehorn, S. O'Connor, F.L. Wackers, and D. Goulson. Neonicotinoid pesticide reduces bumble bee colony growth and queen production. *Science*, 336(6079):351–352, 2012.

[346] J. Whyte, J.J. Whyte, and D.E. Tillitt. Ethoxyresorufin-O-deethylase (EROD) activity in fish as a biomarker of chemical exposure. *Critical Reviews in Toxicology*, 30(October):347–569, 2015.

[347] L.R. Williams, V. Aroniadou-Anderjaska, F. Qashu, H. Finne, V. Pidoplichko, D.I. Bannon, and M.F.M. Braga. RDX binds to the GABAA receptor–convulsant site and blocks GABAa receptor–mediated currents in the amygdala: a mechanism for RDX-induced seizures. *Environmental Health Perspectives*, 119(3):357–363, 2010.

[348] T.D. Williams, N. Turan, A.M. Diab, H. Wu, C. Mackenzie, K.L. Bartie, O. Hrydziuszko, B.P. Lyons, et al. Towards a system level understanding of non-model organisms sampled from the environment: a network biology approach. *PLoS Computational Biology*, 7(8):e1002126, 8 2011.

[349] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S.N. Soiland-Reyes, I. Dunlop, et al. The Taverna workflow suite: designing and execut-

ing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(Web Server issue):557–561, 2013.

[350] B. A. Woodcock, J. M. Bullock, R. F. Shore, M. S. Heard, M. G. Pereira, J. Redhead, L. Ridding, H. Dean, et al. Country-specific effects of neonicotinoid pesticides on honey bees and wild bees. *Science*, 356(6345):1393–1395, 2017.

[351] H. Wu, A.D. Southam, A. Hines, and M.R. Viant. High-throughput tissue extraction protocol for NMR- and MS-based metabolomics. *Analytical Biochemistry*, 372(2):204–212, 2008.

[352] X.S. Wu, B.D. McNeil, J. Xu, J. Fan, L. Xue, E. Melicoff, R. Adachi, L. Bai, and L.G. Wu. Ca2+ and calmodulin initiate all forms of endocytosis during depolarization at a nerve terminal. *Nature Neuroscience*, 12(8):1003–1010, 2009.

[353] S. Yi, H. Zhang, L. Gong, J. Wu, G. Zha, S. Zhou, X. Gu, and B. Yu. Deep sequencing and bioinformatic analysis of lesioned sciatic nerves after crush injury. *PloS One*, 10(12):1–18, 2015.

[354] K.Y. Yip, R.P. Alexander, K. Yan, and M. Gerstein. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PloS One*, 5(1):e8121, 2010.

[355] L. Yu, X. Ma, L. Zhang, J. Zhang, and L. Gao. Prediction of new drug indications based on clinical data and network modularity. *Scientific Reports*, 6(August):32530, 2016.

[356] K. Yugi, H. Kubota, A. Hatano, and S. Kuroda. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic Layers. *Trends in Biotechnology*, 34(4):276–290, 2016.

[357] V.R.T. Zanotelli, S.C.F. Neuhauss, and M.U. Ehrengrubera. Long-term exposure to bis(2-ethylhexyl) phthalate (DEHP) inhibits growth of guppy fish (Poecilia reticulata). *Journal of Applied Toxicology*, 30(1):29–33, 2010.

[358] B. Zhang, X. Pan, G.P. Cobb, and T.A. Anderson. Uptake, bioaccumulation, and biodegradation of hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) and its reduced metabolites (MNX and TNX) by the earthworm (*Eisenia fetida*). *Chemosphere*, 76(1):76–82, 2009.

[359] X. Zhang, E. Szabo, M. Michalak, and M. Opas. Endoplasmic reticulum stress during the embryonic development of the central nervous system in the mouse. *International Journal of Developmental Neuroscience*, 25(7):455–463, 2007.

[360] C. Zhao, Z. Tang, J. Yan, J. Fang, H. Wang, and Z. Cai. Bisphenol S exposure modulate macrophage phenotype as defined by cytokines profiling, global metabolomics and lipidomics analysis. *Science of the Total Environment*, 592:357–365, 2017.

[361] Y. Zhou and N.C. Danbolt. Glutamate as a neurotransmitter in the healthy brain. *Jornal of Neural Transmission*, 121:799–817, 2014.

[362] P. Zoppoli, S. Morganella, and M. Ceccarelli. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *Computational Intelligence Methods for Bioinformatics and Biostatistics*, 11(1):97–111, 2010.

[363] I. Zorita, I. Apraiz, M. Ortiz-Zarragoitia, A. Orbea, I. Cancio, M. Soto, I. Marigómez, and M.P. Cajaraville. Assessment of biological effects of environmental pollution along the NW Mediterranean Sea using mussels as sentinel organisms. *Environmental Pollution*, 148(1):236–250, 2007.