Norwegian University
of Life Sciences

Master's Thesis 2019 60 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Exploring host-microbiome interactions in Norwegian Salmon via weighted network analysis

Marius André Strand
Bioinformatics and applied statistics

Acknowledgements

This thesis was written as part of my master's degree in bioinformatics at the Norwegian University of Life Sciences, Department of Chemistry, Biotechnology, and Food Sciences. I would like to thank my primary supervisor Professor Torgeir Rhoden Hvidsten, and my co-supervisors, Associate professor Phillip Pope and Associate professor Simen Rød Sandve for all their help and patience during this thesis.

Thanks to all my friends who have helped me both directly and indirectly. Thank you for all the support you have given me through a difficult part of my life. I am forever grateful to my family, my mother, and my father, who has always been supportive of me. And finally, to my sister – Not a day goes by that I don't think about you and wish you were still with us.

Ås, December 2019

Marius André Strand

Abstract

Invisible to the naked eye, but present everywhere. Microbial organisms live and reproduce in almost any conceivable environment on the planet. In recent years, researchers have started to gain quantified information of the microscopic world, fueled by the technical advances in sequencing technology. From the presence of microbes in environmental samples (metagenomics) to the processes within a cell (transcriptomics), omics data has revolutionized our understanding of the microcosmos.

One of the places that microbial communities inhabit is the intestines of animals. Many studies have shown that host-associated microbes have a crucial role in the development of the host and are vital for critical functions within many animals. Many studies focus their attention on the microbiota of mammals, and there has been less focus on animals in aquatic environments.

The farmed Atlantic salmon is of great importance for the Norwegian economy. The Atlantic salmon is also an exciting study object due to the dramatic environmental change the fish experiences during its lifetime. The transition from freshwater to saltwater would kill most other fish and is an extreme barrier for the proliferation of the salmon's gut microbiota.

In this thesis, we developed a pipeline using the R-package weighted-correlation-network-analysis (WGCNA) to create networks and detect modules in data from a long-term feeding trial of farmed Atlantic salmon. Through the use of representative profiles for each module, this network-based dimensionality reduction approach gives a holistic approach to the discovery of potential host symbionts.

Complex behaviors emerge from the pairwise interactions of individual objects on every level (genes, proteins, metabolites, cells, animals). Network analysis differs from traditional methods of data analysis by analyzing all relations between objects and studying how they behave together.

In this thesis, we analyze two datasets: one of the host gene expression (RNAseq) and another for the microbial abundance (16S rRNA amplicon sequencing), both of which were previously analyzed separately. We found confirmation of the central status of several bacteria as potential host-symbionts but was limited in our ability to link these bacteria to specific sets of genes by the strong effect of the freshwater-saltwater transition.

A method for regressing out principal components was used to remove large effects in the data in the hopes that it would reveal the subtle interactions. Although the method was successful in discovering genes involved in lipid metabolism, the analysis showed no positive correlation between any group of genes and microorganisms. This lack of positive relationship goes against the established literature on the subject; however, there are some limitations of the method. We discuss these challenges and potential improvements to the pipeline for future studies.

Sammendrag

Usynlig for det blotte øye, men til stede overalt. Mikrobielle organismer lever og reproduserer i nesten alle tenkelige omgivelser på planeten. De siste årene har forskere begynt å få kvantifisert informasjon om den mikroskopiske verden, drevet av tekniske fremskritt innen sekvenseringsteknologi. Fra tilstedeværelsen av mikrober i miljøprøver (metagenomikk) til prosessene i en celle (transkriptomikk), omics-data har revolusjonert vår forståelse av mikrokosmos.

Et av stedene hvor mikrobielle samfunn bor i er tarmene til dyr. Mange studier har vist at vertsassosierte mikrober har en avgjørende rolle i utviklingen av verten og er avgjørende for kritiske funksjoner i mange dyr. Mange studier fokuserer oppmerksomheten mot mikrobiota hos pattedyr, og det har vært mindre fokus på dyr i vannmiljøer. Oppdrett av atlanterhavslaks er av stor betydning for norsk økonomi. Atlanterhavslaksen er også et spennende studieobjekt på grunn av den drastiske endringen i miljø som laksen opplever i løpet av sin levetid. Overgangen fra ferskvann til saltvann vil drepe de fleste andre fisker og er en ekstrem barriere for spredning av laksens tarmmikrobiota.

I denne avhandlingen utviklet vi en pipeline ved bruk av R-pakken weighted-correlation-network-analysis (WGCNA) for å lage nettverk og oppdage moduler i data fra et langvarig fôringsforsøk med atlanterhavslaks fra oppdrett. Gjennom bruk av representative profiler for hver modul gir denne nettverksbaserte dimensjonalitetsreduksjonsmetoden en helhetlig tilnærming til oppdagelsen av potensielle vert symbionter.

Kompleks atferd oppstår fra parvise interaksjoner mellom individuelle objekter på alle nivåer (gener, proteiner, metabolitter, celler, dyr). Nettverksanalyse skiller seg fra tradisjonelle metoder for dataanalyse ved å analysere alle forhold mellom objekter og studere hvordan de oppfører seg sammen.

I denne oppgaven analyserer vi to datasett: ett av vertsgenekspresjonen (RNAseq) og ett annen for mikrobiell tallrikhet (16S rRNA amplicon sekvensering), begge har tidligere blitt analysert separat. Vi fant bekreftelse av den sentrale statusen til flere bakterier som potensielle vert-symbionter, men var begrenset i vår evne til å koble disse bakteriene til spesifikke sett med gener av den sterke effekten av overgangen til ferskvann og saltvann.

En metode for å fjerne prinsipalkomponenter ved regresjon ble brukt for å fjerne store effekter i dataene i håp om at de ville avsløre de subtile interaksjonene. Selv om metoden lyktes i å oppdage gener involvert i lipidmetabolisme, viste analysen ingen positiv sammenheng mellom gen-grupper og grupper av mikroorganismer. Denne mangelen på positive forhold går mot den etablerte litteraturen på emnet; Imidlertid er det noen begrensninger i metoden. Vi diskuterer disse utfordringene og potensielle forbedringer til pipeline-en for fremtidige studier.

Contents

Chapter 1

Introduction

Microbes are everywhere. Entire unseen communities that live among us, on us and within us. Some microbes have adapted to live almost anywhere, making microbes a genuinely ubiquitous part of nature.

1.1 Microbiomes

Most, if not all, plants and animals have associated Bacteria, archaea, fungi, protozoa and viruses, which make part of the multicellular host their home (Sommer and Bäckhed 2013). Collectively all these microscopic creatures are referred to as the MICROBIOTA of that host, whereas specific localized communities are known as microbiomes.

Where researchers previously considered microbes for the diseases they caused, today most have come to appreciate the mainly beneficial influence that host-associated microbiomes have (Sommer and Bäckhed 2013).

The formation and preservation of healthy gut microbiota contribute, in addition to the extraction of nutrients, to the normal physiological development of the host; The development of the gastrointestinal tract, other organs, and the immune system are some broad examples (Sommer and Bäckhed 2013). While healthy flora contributes towards normal development, imbalances in the gut microbiome have been associated with everything from obesity and diseases like irritable bowel syndrome, to asthma, arthritis, and even anxiety-like behavior (in mice) (Sommer and Bäckhed 2013).

The "Holobiont" theory of evolution considers the host and its microbiome as a single unit of selection (Zilber-Rosenberg and Rosenberg 2008; Bordenstein and Theis 2015). This view is, however, not without its critics (Moran and Sloan 2015). Nonetheless, the debate regarding the Holobiont theory illustrates the increased awareness of the role that microbes play in shaping both animals and plants.

Although interconnections between host and microbiota has been studied intensively in many model animals and in humans; and many extremely close interconnections have been found between those hosts and their microbiota (Rudman et al. 2019; Dominguez-Bello et al. 2019; Desbonnet et al. 2014). A greater diversity of host organisms is needed to understand general and fundamental principles of beneficial microbe-host-interactions.

It is, for example, not clear-cut that all animals are dependent on their microbiota (Hammer, Sanders, and Fierer 2019). Only a small proportion of species have been shown to depend on a microbiome directly, and as such dependence should not be the default assumption (Hammer, Sanders, and Fierer 2019). While it may be the case that all vertebrae, for example, form important mutualistic relationships with their microbiota, such claims can only come from extensive studies of many species. We still know very little about gut microbiota in aquatic environments (Rudi et al. 2018; Jin et al. 2019) including in fish

(Egerton et al. 2018). To truly understand what is specific and what is general in the world of host-microbe interactions would require examination of many different organisms with very different lifestyles.

1.2 The Atlantic Salmon

The Atlantic salmon is most notable for living in rivers while young, heading to the ocean to mature, and then returning to the same river to breed.

Differences between freshwater and saltwater environments act as a barrier that usually stops any fish from inhabiting both. For fish that live in saltwater, its gills actively get rid of salt. For fish that live in freshwater, gills actively take up salt through its gills. A fish with the wrong osmotic regulation will not be able to maintain a stable internal condition (homeostasis) for the chemical and physiological processes of its body. As a consequence, the fish would die.

The Atlantic salmon, however, does this switch more than once in its life. The process, known as smoltification, involves changing the regulation of its body chemistry to adapt to the different environments.

In addition to changes of osmoregulation, the Atlantic salmon also needs deal with a drastically different availability of dietary lipids. Juvenile fish, called parr, mostly eats invertebrates which are low in long-chained poly unsaturated fatty acids (LC-PUFA) such as omega-3. Because of this salmon has developed an ability to produce this in their own bodies (Gillard et al. 2018). In the post-smolt fish, after smoltification, the fish heads out to sea, where there is a high availability of LC-PUFA in the form of other fish and krill (Gillard et al. 2018).

It is theorized that the abilities of salmon to adapt to these changes has come about because of an ancestral whole-genome duplication event that happened some 80 million years ago (Gillard et al. 2018).

In contrast to mammals, which keep a steady temperature in their bodies, poikilothermic ectotherms or 'cold-blooded' animals, like the Atlantic salmon, do not regulate their body temperature in any way. So their internal temperature fluctuates depending on the environment. Changing temperatures leads to a unique challenge for the gut microbiota of these animals, that is not present in endothermic ('warm-blooded') animals (Kokou et al. 2018a).

All fish does the majority of their digestion in their mid-gut (Egerton et al. 2018). In addition some fish also has a pyloric caeca, a collection of finger like protuberances right before the start of the large intestine, that is thought to help the fish retain more bacteria and increase feed abortion by increasing the surface area (Egerton et al. 2018). Its been known for a long time that the colonisation of the gut by microbes are specific for each species (Egerton et al. 2018) and that the microbial composition of the fish gut is very different from the environmental microbial composition (Lokesh et al. 2019).

It has been found that the most abundant phylum in the gut of fish in late stage freshwater are *Firmicutes* (Lokesh et al. 2019). While bacteria from the phylum *Proteobacteria* gain dominance when the fish enter saltwater.

For farmed Atlantic salmon, fish raised with low LC-PUFA had a microbial community that was dominated by *Firmicutes* just as in the wild Atlantic salmon (Jin et al. 2019). For those that were given high LC-PUFA the community was dominated by *Proteobacteria*. This shows that the diet given to juvenile Atlantic salmon can influence the microbial composition after the fish has entered saltwater, and that the gut microbiome of fish given low LC-PUFA most resemble that of the wild gut microbiome after transition to saltwater (Jin et al. 2019).

1.3 A revolution in molecular biology tools

Since the rise of molecular biology more than 80 years ago, our knowledge of the molecular biology of the cell has revolutionized the way we understand living systems, and how they function.

As the development of tools and subsequent discoveries gave rise to new and improved methods and tools, increasingly detailed observations about the structure and function of the cellular components were made. This research has concluded that the building blocks of life are more or less identical across the kingdoms of life; All cells have DNA, RNA, and proteins, and they all have the same role in all living systems.

It is this universality of building blocks that has allowed us to understand the structure of genomes, the regulation of gene activity (transcription) and the synthesis of proteins (translation), which are the ultimate determinants of cell function.

The revolution in high throughput DNA-sequencing, ever since the completion of the human genome project two decades ago, has truly been a game changer in molecular cell biology research. We can now easily, and at low cost, quantify the activity and composition of the molecular cellular universe that surrounds us, from the presence of microbes in environmental samples to the cellular processes within the cell. However, the scale at which these new sequencing technologies generate data has overwhelmed traditional approaches for handling (storing and retrieving) and analyzing such data types, making applied mathematics, statistics, and computer science increasingly important components in modern molecular biology research. Bioinformatics is at the intersection of all these disciplines – using computational approaches to organize and extract biological insights from (mostly) the vast amounts of sequencing data that quantifies DNA, RNA or proteins.

1.4 Cellular processes

A living cell is a self contained system that can respond to outside stimuli and internal demands. All the genes of a cell control the metabolic processed that happens in a cell. The genes are all part of the gene regulatory network (GRN). Transcription factors are proteins that recognize specific sections of the DNA by binding to them and in so doing initiating the transcription of a nearby gene. As many as 10% of genes are transcription factors (TF) (Alberts et al. 2015, p. 373). Most genes are controlled by more than one TF (Alberts et al. 2015, p. 373). Gene expression can be seen as the output of the GRN.

1.4.1 Transcription and regulation of transcription

A simplistic example of a response to stimuli can start with a cascade, which results in a transcription factor (TF) initiating transcription of a gene into many identical RNA transcripts. These transcripts then get translated into many identical proteins. Each specific protein typically only has one function (Alberts et al. 2015). Still, a collection of different proteins can accomplish more complicated tasks e.g., making a fatty acid, initiating cell division, or organize bilateral symmetry during embryo development. The beginning of the cascade can be initiated by a receptor on the cell surface, detecting a hormone. Intervening many of these steps are mechanisms that regulate the final expression, see figure 1.1, although most regulation happens at the transcriptional stage (Alberts et al. 2015, p. 373). Other factors, like the lifetime of the produced molecules, can also affect the direct relationship between transcript expression and protein abundance. As a result, the expression of a gene does not necessarily correspond to the expressed level of the related protein. It is

estimated that only about 40% of the variation in protein level can be explained by knowing the mRNA transcript level (Vogel and Marcotte 2012). Still, changes in the expression of genes can indicate patterns of gene behavior and give insight into the status of cellular processes (Eisen et al. 1998). And measuring the expression of genes allows us to discover these variational/dynamic groups of genes.

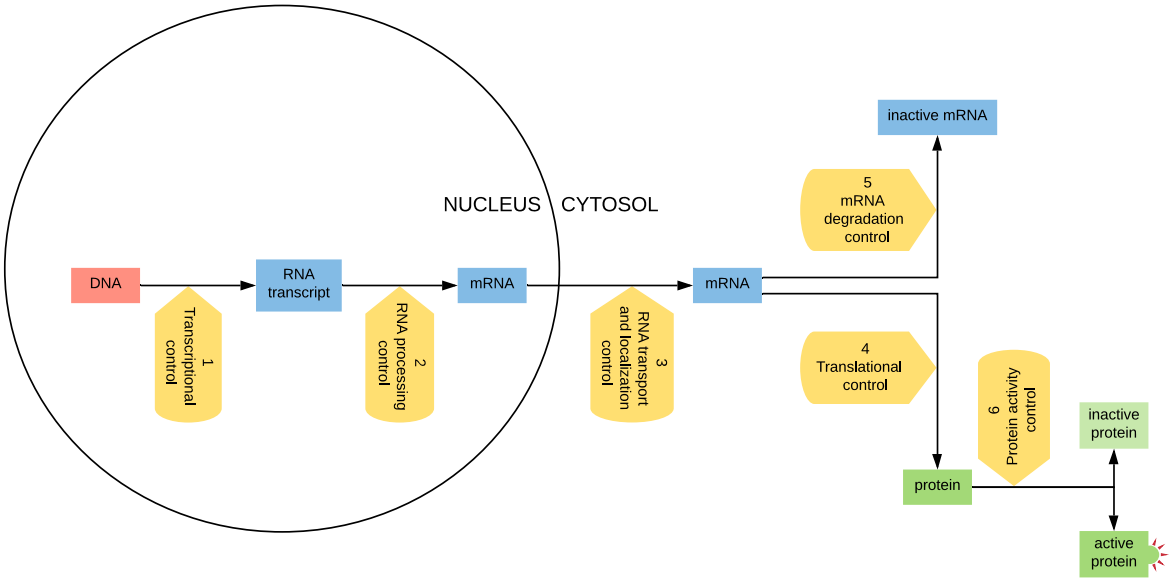

Figure 1.1: Six steps at which eukaryotic gene expression can be controlled. Adapted from Molecular Biology of the Cell, Alberts et al. (2015, p. 373, Figure 7-5).

1.4.2 Modularity

Modularity is a well established phenomenon in biology (Wagner, Pavlicev, and Cheverud 2007). The modular nature of genes, proteins and metabolites, gives rise to complex physiological processes. Gene expression is organized with groups of genes that have a high within-group connection, and lower between-group connection, i.e. Some genes are more connected to each other than they are to the rest of the network. Co-expression patterns are variational modules, a representation of the structure that the gene regulatory network outputs.

1.5 Network analysis

Networks offers a way of precisely describing relations between objects. Network analysis is a conceptually different framework to modelling compared to many other methods. In network analysis the relations between variables are in focus (A. L. Barabási and Oltvai 2004). This separates it from more traditional analysis where variables are considered by themselves or as a combination of a few variables (multivariate analysis).

The main goal of a network analysis is to understand the structure and dynamics of complex systems (A. L. Barabási and Oltvai 2004). One of the earliest and most widespread scientific adopters of network analysis has been sociology, where the relationships between people is studied. It has also increasingly been used to understand complex systems in biology (Dong and Horvath 2007) as well as many other areas of science (A. L. Barabási and Oltvai 2004).

Complex behavior emerges from simple rules of pairwise interactions (A. L. Barabási and Oltvai 2004). An emergent system can be described as a system which is more than

the sum of its parts. As detailed above in section 1.4, even seemingly simple 'machines' that does one thing each, can in conjunction with other machines produce results that are seemingly outside the capabilities of single entities.

Because complex biological networks often have modular properties, a property that they share with other complex networks (A. L. Barabási and Oltvai 2004). Studies of complex systems can gain insights from a wide range of scientific fields. Modularity and the concepts of connectivity provide a useful theoretical framework from which the analysis of interacting parts can be better understood (Dong and Horvath 2007).

1.6 Studying host-microbe systems

The advent of massively parallel sequencing opened up a whole new world of derivative technologies. Technologies that, by and large, have created entirely new ways of studying biological systems.

RNA-seq increased the possible detection range for the quantification of transcript expression, but also reduced the cost and need for specially produced microarrays (Shendure 2008; Wang, Gerstein, and Snyder 2009). These improvements opened up the possibility of studying gene expression for animals and plants that had never been sequenced before (Shendure 2008; Wang, Gerstein, and Snyder 2009).

The reduced cost of sequencing also meant that taxonomic marker gene studies, such as 16S rRNA amplicon sequencing, now could be applied wholesale to environmental samples.

Generally we think it is useful to differentiate between two types of approaches, non-network methods and network methods. Within each approach there are many different assumptions and practises. Approaches differ on the intent end result, some are fine mining for direct interactions, others try to gain a more holistic overview. Many also differ in the extent at which they integrate data from both host and microbiome. For some, studying pathogens the main culprit(s) are usually already known or suspected, and therefore such studies use only data specific for these organisms. However, such singular focus could easily miss important aspects of the system which only a more complete omics approach could identify. Network methods, in contrast to non-network methods also allows for the study of emerging properties which only come into view when studying the system as a whole.

1.6.1 Non network methods

Non-network methods are the more traditional approach to data analysis. Here each variable is treated as if it is independent from all others.

- Dimension reduction, like Principal component analysis (PCA)
- Correlation based methods
- Regression based methods

For plants and animals that cannot regulate their internal temperature, both the host and its microbiome are presented with challenges during temperature changes. Kokou et al. (2018a) used a trans-generational study of cold-resistant blue tilapia (Oreochromis aureus) to show that the genetics of the fish influenced the composition and the response of the gut microbiota to cold. For their study they used, among other methods, PCA to calculate how much of the variance in the microbiota could be explained by the gene expression of the host. Cold-resistant fish had gut microbiota that were more stable (less variation) even under optimal non stressed conditions.

1.6.2 Network approaches

Even though network analysis is relatively new in both transcriptomics and metagenomics studies, there are already many different measures and models (Faust and Raes 2012; Layeghifard, Li, et al. 2019).

Complex models have a higher chance of overfitting data. A detailed model that captures all interesting data in one dataset can be an exceedingly poor predictor for future datasets. This apparent contradiction comes from the fact that complex models more easily can capture particularities and noise from the training dataset. Even though there are many potential complex interactions between microbes, it is thought that most dynamics can be explained through pairwise interactions. Evidence for this mostly-pairwise-interactions hypothesis has been found through systems biology modeling, which showed that the growth and pairwise interaction of ecological driver species exhibit the most substantial impact on the community structure.Venturelli et al. (2018)

- Dissimilarity based methods
- Correlation based methods
- Regression based methods
- Probabilistic Graphical Models *

* (Layeghifard, Hwang, and Guttman 2017)

In many ways, methods within network analysis mirror non-network methods. The reason is simple; Network analysis is really about shifting perspective, it is not how it is measured, but instead what we measure for that matters. As listed above, both pairwise similarity/dissimilarity measures and other more complex measures like multiple regression and rule-based interactions can be used to create networks (Faust and Raes 2012; Layeghifard, Li, et al. 2019).

Weighted co-expression networks

Co-expression networks have been useful for describing pairwise relationships among transcripts (Langfelder and Horvath 2008). WGCNA has been used in several highly cited studies and is well proven for microarray and RNAseq data (IWGSC 2018; Xue et al. 2014; Hawrylycz et al. 2012; Voineagu et al. 2011; Oldham et al. 2008).

In standard WGCNA procedure eigengenes can be correlated to sample traits via simple correlation. The relationships of eigengenes has also been further explored showing that there are meta-structures in transcriptomics data, and that these modules can be compared across datasets (Langfelder and Horvath 2007). This lays the ground work for working with eigengenes across datasets also across the datasets of different omics-types.

WGCNA has been applied to both metabolomics and proteomics (Pei, L. Chen, and W. Zhang 2017; DiLeo et al. 2011). The R-package has also been applied to 16S rRNA amplicon sequencing data (16S-seq) (McHardy et al. 2013; Duran-Pinedo et al. 2011; Geng et al. 2016).

Many approaches that have used weighted networks for the integration of host and microbes have either used very limited data correlating only a few OTUs at a time. Previous studies using WGCNA and associating single bacteria, or using a measure of bacterial community complexity like beta-diversity to asses the influence that the host transcriptome has on these factors and vice versa. There has also been studies where WGCNA has been used on bacteria and associated those with some trait of the host. But as far as we know none

have used WGCNA network analysis on both host transcriptomics data and metagenomics abundance data simultaneously and correlated the resulting modules to each other.

This is exemplified by a study on pulmonary fibrosis where microbial traits and phenotypic traits of interest were correlated to MEs found on DE genes related to the disease (Molyneaux et al. 2017).

Some that do use weighted network approach and WGCNA on bacterial data often associate to only one trait, or a limited number of traits. Likewise studies into the transcriptome usually try to find genes correlated to bacteria, but focus on a limited number of predefined bacteria .

Huang et al. (2017) found gene modules and correlated MEs with bacterial community variables and individual OTUs. But did not perform any network analysis on the OTUs.

The studies mentioned above fall short of any 'complete' exploratory data analysis about the connections between host and microbiomes. As far as I can tell, few network approaches rely on both transcriptomics data and metagenomics data or use them to their full potential.

All these previously mentioned papers, use in some way the WGCNA package, and the weighted network methodology it provides. But the most common type of network analysis is unweighted.

The Transkingdom pipeline uses unweighted network analysis to associate bacterial genes with genes from the host. It also depends on the use of Differential Expression to reduce the number of host genes and microbial genes. The approach integrates networks but is dependent on different programs for the selection of modules (Rodrigues, Shulzhenko, and Morgun 2018). Genes from the gene network are then correlated to the microbial gene network. A measure of bipartite betweenness centrality is used to find bottleneck microbes genes (Rodrigues, Shulzhenko, and Morgun 2018). It has also been used on microbial abundances from 16S data (Greer et al. 2016). The Transkingdom pipeline relies heavily on setting thresholds for inclusion throughout the pipeline. Unweighted networks are highly dependent on the threshold for which they are constructed (Horvath 2011). The pipeline also involves manual work to get all parts set together.

In this thesis the focus will be on the application of weighted network analysis for integrating host-microbiome data.

Earlier work on host-microbe interactions using weighted network analysis has often been limited in scale either focusing on a few microbes, or a few host traits. Methods that do integrate and look at large scale interactions have to my knowledge only used unweighted networks and depend on limiting the analysis on just the most differentially expressed genes and the most differentially abundant microbes.

1.7 Study aims

The first aim of this thesis is to develop a pipeline for the network-based dimension reduction of host-microbe multi-omics data. Relating modules from the host organism with modules from the microbiome and identifying potential interaction between them. A sub-goal is to find good ways to represent this data and give a holistic overview of the data at hand.

The second aim is to apply this pipeline to discover interactions between the Atlantic salmon and its microbiome.

The third aim is to discover potential challenges and propose future improvements.

Additionally, as our intention is for the improvement and reuse of this pipeline, the code will be made available as a GitLab repository. Finally, for reproducibility, the code will be written almost exclusively in R-markdown.

Chapter 2

Background

2.1 16S rRNA amplicon sequencing – Quantified Taxonomic profiling

The 16S rRNA gene is essential for protein synthesis within the prokaryotic cells. Because of this, the gene is highly conserved allowing phylogenetic (what is related to what) comparisons between very different prokaryotes (Weisburg et al. 1991). The 16S rRNA gene is essentially >97% identical within prokaryotic species (Woo et al. 2008).

Of course, the 16S rRNA gene is not the only gene that can be used for comparing microbial species. As there are many genes that can be used as marker genes. For example, for nitrogen-fixing prokaryotes, the nifH gene is highly conserved, and is therefore used to identify and compare nitrogen-fixing bacterial and archaeal species (Gaby and Buckley 2012).

Polymerase Chain Reaction (PCR) is a method that allows for the amplification of DNA sequences. A repeatedly copied genetic sequence is referred to as an amplicon.

The use of 16S amplicon sequencing has made possible the discovery of many new species and also resulted in many taxonomic re-classifications of already known species(Woo et al. 2008). For species that have unusual phenotypes, grow slowly in laboratory conditions or simply cannot be cultured at all, using sequencing based technologies have made their discovery and classification possible (Woo et al. 2008).

16S amplicon sequencing, illumina protocol(Illumina n.d.):

1. Data that is sampled from a site must first be made ready for sequencing, by first breaking up the cells and amplify part of the 16S rRNA sequence using degenerate primers.

2. 1. Amplicon Primers (interest specific, choice of primers introduce bias, there are no perfectly universal primers available)

3. 2. PCR amplify

4. 3. Add barcodes and adapters.

5. 4. Sequence using pair end reads

Data processing

1. Demultiplex, samples are usually sequences together because of cost.

2. Remove redundancy but counting identical reads and keeping only one representative.

3. Look for and remove wrong hybridisation's, called chimeras, that happened during the PCR step.

The most common approach to dealing with 16S amplicon sequencing data is based on comparing and dividing these sequences based on their sequence identity. Because of the high degree of noise both due to the necessary laboratory steps for preparing samples and for sequencing, the output data from such efforts are highly diverse but this diversity does not represent the true underlying biological variation. One thing that defines amplicon data, apart from the fact that it allows for discovery of unculturable species of bacteria and archaea, is that it has very large levels of systematic and random noise. To combat such noise, a cutoff of 97% sequence identity, i.e. 3% dissimilarity, is used to define what is called Operational Taxonomic Units (OTUs). Why 97%? Most likely because it corresponded to already defined species classifications. As such an OTU should roughly correspond to a prokaryotic species. Classification of such OTUs are usually done by matching centroid sequences to sequences in one or more databases that has associated sequences with taxonomic classifications. It is the OTUs that are the basis for all downstream analysis within this thesis.

2.2 RNAseq – Capturing a snapshot of cell expression

The transcriptome of a cell is all RNA transcripts expressed in that cell. As previously discussed, the transcription and regulation of transcription in a cell is a highly dynamic process because it is the cells responses to internal and external stimuli. As such, any measurement of the transcriptome is a snapshot of the processes that are happening in that cell in response to some condition.

RNA-seq is a method for measuring the transcriptome, which started replacing older methods like microarrays around 2008 (Shendure 2008).

One of the drawbacks of using microarrays, especially for non-model organisms, was the need to know the genome sequence to design probesets for which the transcripts could hybridize. The reliance on hybridisation also meant that microarrays had fundamental problems with background noise, because the probes, and especially short probes, tended to hybridize to more than one gene product (Wang, Gerstein, and Snyder 2009). This cross-hybridization and the use of light intensity to measure abundance meant that microarrays had a limited dynamic range (Wang, Gerstein, and Snyder 2009), i.e. they could not measure either very low expression or very high expression. RNA-seq in contrast works by sequencing transcripts, in the form of cDNA (Wang, Gerstein, and Snyder 2009). This results in very low background noise and no upper limit on quantification (Wang, Gerstein, and Snyder 2009).

For microarrays, many normalisation methods became increasingly difficult and complicated, and initial proclamations of the ability of RNA-seq to effectively measure absolute quantity lead some like Wang, Gerstein, and Snyder (2009) to believe that RNA-seq would itself require little to no normalization. This however, turned out not to be the case. RNA-seq also has to deal with technical and biological biases like library preparation, laboratory specific practices, sequencing depth, gene length, GC content, transcript lengths etc. (Ballouz, Verleyen, and Gillis 2015).

RNAseq: Illumina protocol:

1. Isolate the RNA

2. Break into smaller segments
3. Convert to double stranded DNA via reverse transcription
4. Add sequencing adaptors
5. PCR amplify
6. Verify concentration and fragment length
7. Sequence on an illumina sequencer

Data processing

Raw sequencing results require computationally intensive mapping to either a reference genome or by de novo reconstruction to be useful as quantitative data on the transcriptome.

For complex genomes, such as multicellular eukaryotic species, the presence of introns, non coding segments of genes that facilitate alternative splicing of genes into gene isoforms, different versions of the same gene. This complicates the quantification process for such organisms, by making mapping reads to gene ambiguous. Many sequence mappers deal with this problem in various ways.

With the advent of longer sequencing read technologies, such problems might one day become a problem of the past. Longer reads that span entire exons and introns, or even multiple exon-intron barriers would be needed to get non-ambiguous results.

2.3 Pre-processing of high throughput sequencing data

2.3.1 Normalization

The idea of normalization is to remove any systematic bias from the data. Data appropriate normalization is a critical part of any analysis, and network analysis is no exception.

TSS and TPM

A common way of normalizing is to divide each sample by their total sum. This often the simplest method for normalizing any sample. An additional requirement for genes is to also to account for gene length by the same procedure. The order of operation separates Reads Per Kilobase Million (RPKM) from Transcript per million (TPM).

TMM

Trimmed mean of M values (TMM) is a normalization method for gene expression data. It works by estimating global fold change and adjusting samples under the assumption that most genes should not be differentially expressed (Robinson and Oshlack 2010).

CSS

CSS was developed specifically for gene marker data, based on the use of quantile normalization approaches in RNAseq (Paulson et al. 2013). One of the defining characteristics of marker gene data is as noted earlier the extreme sparsity: most variables are rare because of a combination of true zeros and under-sampling (Paulson et al. 2013). CSS normalizes the raw data by dividing the raw counts by the 'Cumulative sum of counts' up to a percentile

that is determined directly from the data (Paulson et al. 2013). It has worked well when tested for differential abundance effects, but depends on a large set of samples to work, and fails for low sample data (Pereira et al. 2018). Pereira et al. (2018) found that CSS normalization was particularly good at controlling the false discovery rate (FDR) when other normalization techniques had problems with unbalanced effects, i.e. one group has a large amount of differential abundances compared to other groups.

2.3.2 Transformations

Both low variance and low expression is usually filtered out before further processing, because it is thought to mostly represent noise. Transformations are usually done to reduce the variance of highly expressed genes.

2.4 Correlation

Correlation is a measure of association between two variables that is independent of the unit-scale. There are many types of correlation coefficients, but the most common and most straightforward is the Pearson correlation. Pearson correlation, however, has some underlying assumptions that are not always met. Pearson correlation is meant to measure linear relationships between independent continuous and normally distributed variables.

The strength of the measured relationship is between -1 and 1, where a correlation of 0 means no association and a correlation of either -1 or 1 means a complete negative or positive correlation, respectively.

Spearman correlation is a rank-based version of the Pearson correlation. It is, therefore, able to measure non-linear relationships and can handle that one or both of the variables are skewed and non-normally distributed (Mukaka M.M. 2012). It does, however, require that the relationship between the variables is monotonic.

Mutual information can detect relationships between variables that are non-monotonic. There also exist other types of correlation coefficients, one of which will be discussed later in this thesis.

2.5 PCA – Principal component analysis

PCA is a simple and therefore commonly used method in modern data analysis that can extract relevant information from complex and confusing datasets without the need for tweaking parameters(Shlens 2014).

For statistical learning methods we distinguish between two types; The supervised kind were we already know what groups we are looking for, and the unsupervised kind were the method discovers groups without any prior knowledge (also referred to as clustering). PCA is the latter.

PCA works by first finding which linear combination ($y = w_1 \cdot gene_1 + w_2 \cdot gene_2 + \cdots$) of the original variables captures the largest possible variance (Shlens 2014). This new variable, called the first principal component (PC), now captures the most variance in the data of all such possible linear combinations. After this variation has been "captured", the second PC is found that captures the most variance not captured by PC1, and so on until all the variation of the data has been captured. Each of these new latent variables are orthogonal to each other, which means that they are uncorrelated i.e. don't capture the same variance. If the underlying structure is related to the property we are interested in, the first couple

of variables will hopefully capture all needed signal and the remaining structure is noise (Shlens 2014).

If the property we are interested in examining is the reason for this variance, this hidden (latent) variable now reveal a structure that is related to this property, but fails to reveal the desired structure if the largest variance corresponds to some thing else (Shlens 2014).

2.6 Hierarchical clustering

Hierarchical clustering is a greedy algorithm for clustering objects based on measured dissimilarity. The algorithm can either be agglomerative, where all objects are separate and then combined until all objects are clustered; Or, divisive, where all objects start out in one large cluster that gets divided into smaller groups.

The hierarchical clustering algorithm needs to have a dissimilarity (i.e. distance) measure to compare objects. Some common distance measures include Euclidean, Manhattan and binary, but there are many more. It is also possible to create distance measure from a similarity measures like Pearson and Spearman correlation by subtracting one from the result: 1-correlation.

Comparing single objects is trivial, but comparing groups depends on using a clustering rule, which can be either: Single linkage, closest members of each group are compared, Complete linkage, the most distant members of each group, or Average distance, which is the distance between the arithmetic mean also called the centroid of each group.

An agglomerative hierarchical clustering algorithm starts by comparing all objects, lets now call them clusters, to every other object/cluster. The two clusters that are closest together, as determined by the distance measure and the clustering-rule, are merged to become a new cluster. Then the process repeats until there is only one cluster left. The result is a hierarchical grouping of clusters which can be viewed as a tree-like structure, where connections between groups split into two for each step. Such a structure is appropriately called a dendrogram. The number of clusters you get is defined by cutting the dendrogram at different heights.

The algorithm is deterministic, which means that repeating the process with the same data results in the same dendrogram. However, because the algorithm only chooses the first and best merging, it can miss better mergers that result in more optimal groups (Black 2005). It also means that adding or removing objects can have a big impact on the final dendrogram and therefore the resulting groups. Changing between distance measures and clustering rules also change the resulting dendrogram and therefore also the resulting groups.

2.7 Graph theory and Networks

Nodes and edges

A graph is a mathematical structure, consisting of both vertices, often called nodes, and edges. More precisely, a graph is an ordered pair, see eq. 2.1, consisting of a nonempty set of vertices V eq. 2.2 , and a set of edges E eq. 2.3 that contain two-element subsets of V. (Levin n.d.; Freeman 1978)

$$G = (V, E) \tag{2.1}$$

$$V = \{A, B, C, D, E\} \tag{2.2}$$

$$E = \Big\{\{A,B\},\{A,C\},\{A,D\},\{A,E\},\{B,C\},\\ \{B,D\},\{B,E\},\{C,D\},\{C,E\},\{D,E\}\Big\} \tag{2.3}$$

Two nodes are adjacent if they are connected by an edge. An adjacency matrix is the collection of all pairwise relations between nodes. Such a matrix is called a square matrix because it has an equal number of rows and columns. Because the order of each row and each column can be changed there exists many adjacency matrices that describe the same graph. Likewise, a graph will also have many graphical representations that are all equally valid.

Directed and undirected graphs

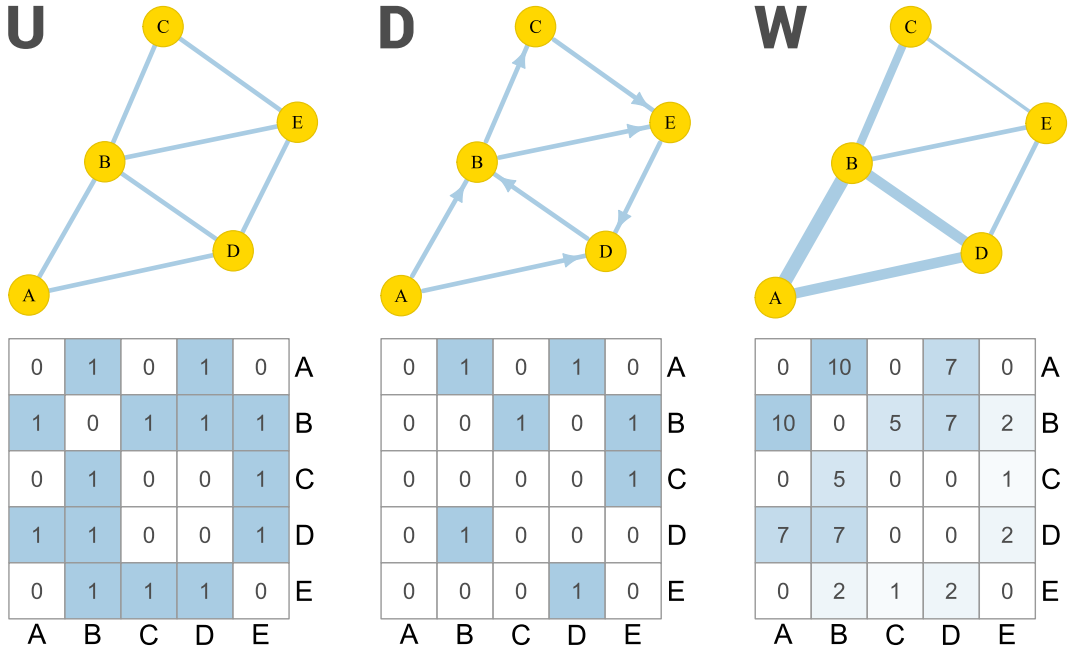

Figure 2.1: Examples of graphs: **U** - Unweighted and undirected graph. **D** - Unweighted and directed graph. **W** - Weighted and undirected graph.

Figure 2.1 shows three different graphs with different properties. Graph **U** is what is called a simple or strict graph, because it is unweighted, has no direction, contains no loops and has no more than one edge between nodes. In undirected graphs like **U** and **W**, the direction of the relationship is not specified, and the corresponding adjacency matrix is symmetrical. For directed graphs, the direction of the relationship does matter, and the corresponding adjacency matrix is unsymmetrical. All graphs drawn are also *connected* which means that it is possible to reach any node from any other node. (Levin n.d.)

Paths

The basis for graph theory was established by mathematician Leonhard Euler when he tried to answer the Seven bridges of Königsberg problem: Can you devise a walk through the city that would cross each bridge once and only once? In abstract terms: Can you traverse through every edge in a graph exactly once. Traversing from node to node via edges is called a path. Calculating paths can be useful for measures of centrality as we will see below.

The shortest path between two nodes is the path that includes the fewest number of edges. In a weighted graph, it is the path that has the lowest sum of edges from one node to another.

Centrality measures and scale free networks

There are three fundamental measures of centrality (Freeman 1978):

- The DEGREE CENTRALITY: The number of edges that are connected to a given node.
- BETWEENNESS CENTRALITY: Counts the number of times a node is part of the shortest path between other nodes.
- CLOSENESS CENTRALITY: Reciprocal of the sum (i.e $\frac{1}{\text{sum}}$) of the length of the shortest path between a node and all other nodes in the graph.

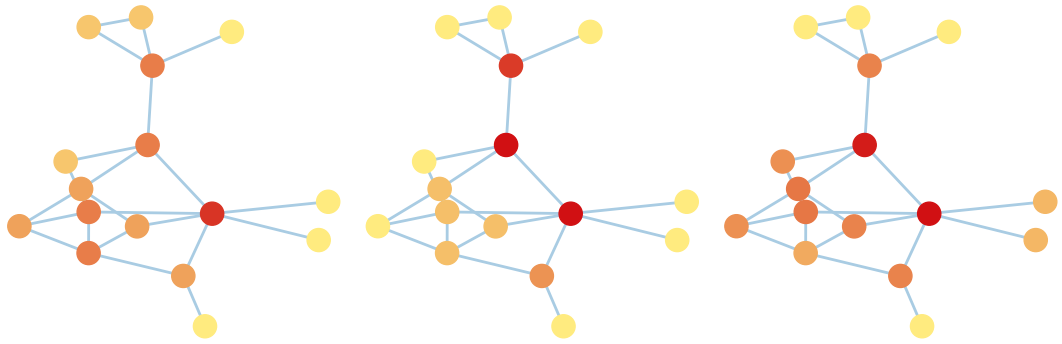

Figure 2.2: Centrality measures. From the left: Degree, Betweenness & Closeness. These centrality measures are a measure of node importance. Many nodes that have a high score in one centrality measure also has a high score in another. However, there are some nodes that are deemed important by one centrality measure that do no receive equally high score in another.

The graph in figure 2.2 is a random graph. A random graph follows a Poisson distribution, assuming that the probability of connecting two nodes is random and uniform. Traditionally, many theoretical properties of graphs were described by the random graph theory of Erdős and Rényi (A.-L. Barabási and Albert 1999). With the increased availability of large real world networks A.-L. Barabási and Albert (1999) showed that real world large-scale properties of complex networks have a high degree of self-organisation, and that independent of the system and what it described, the degree distribution of the nodes followed a power law distribution, see equation 2.4.

$$\mathrm{P}(k) \sim k^{-\gamma} \tag{2.4}$$

A network with this degree distribution is referred to as being scale-free, because the distribution of the local connectivity (Degree) is free of scale (A.-L. Barabási and Albert 1999). The degree distribution naturally generalizes to a weighted graph where k is a non-negative weight (B. Zhang et al. 2005).

For scale free networks, a majority of nodes have few connections, while a minority has many. The origin of such networks can best be described by i) continual growth, that

the number of nodes increases throughout the lifetime of the network, and ii) preferential attachment, that new nodes are more likely to attach to nodes that already have many connections (A.-L. Barabási and Albert 1999). An example of preferential attachment is that a new author is more likely to cite a well known paper that is already heavily cited, than that author is likely to cite a less known and less cited paper (A.-L. Barabási and Albert 1999).

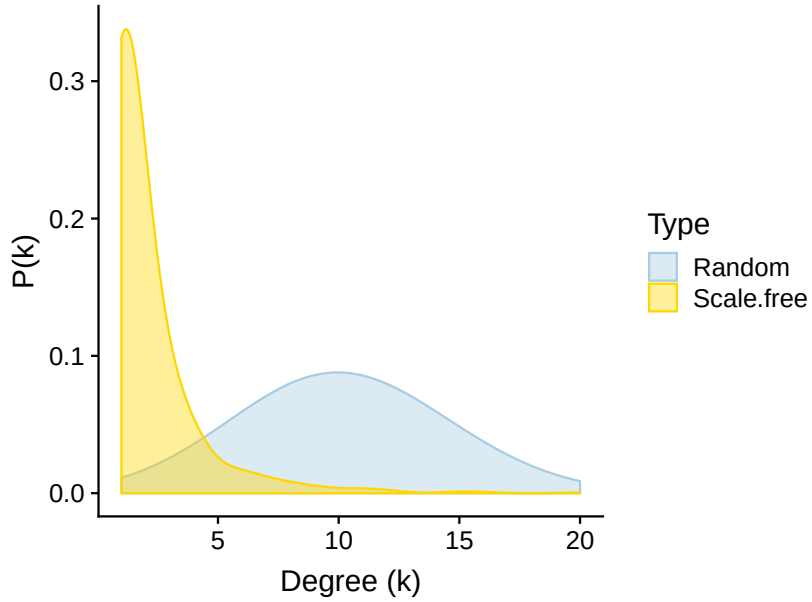

Figure 2.3: Degree distributions, generated in R and visualized in R with igraph.

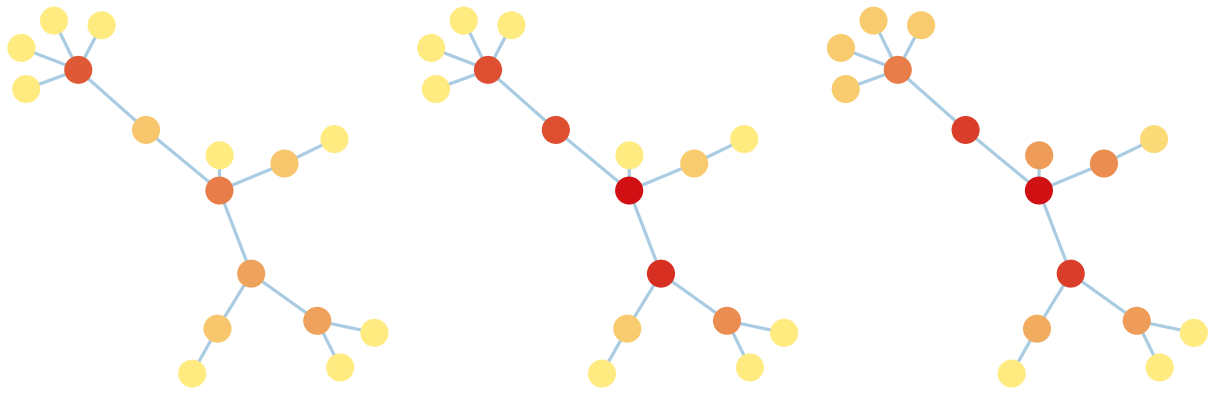

Figure 2.4: Centrality measures. Degree, Betweenness & Closeness

As you can see in the figure 2.4, only a few nodes are highly connected.

2.8 WGCNA – Weighted Gene Co-expression Network Analysis

WGCNA exists as a stand-alone R-package, with available tutorials and proper documentation. WGCNA was developed for microarray data but has also been used for RNA-seq data (Langfelder and Horvath 2014). The WGCNA R-package has functions that can be used to create a network and detect modules within that network. WGCNA can be used as a biologically-motivated network-based reduction method by using a representative expression profile for each module (Langfelder and Horvath 2008; Horvath 2011). This representative can either be a central hub gene or a sort of weighted average called the module eigengene, i.e., the first principal component of the expression in a module. WGCNA has primarily been used to analyze genomics data, from brain cancer to yeast cell cycle to plants (Langfelder and Horvath 2008).

The authors of the WGCNA R-package usually use specific words like gene expression, eigengene, etc. in their tutorials and articles. Here, however, because I will also include microbial abundances, I will try to use more general words; node profile, eigennode, etc.

2.8.1 Making a correlation network

We start out with an M x N data matrix, where M is the number of samples, and N is the number of variables. In WGCNA, we want to create a graph from this data by relating (correlating) every variable to every other variable. The collection of all pairwise correlations are stored in a N x N similarity matrix, and since correlation does not depend on the order of variables, i.e., $cor(A, B) = cor(B, A)$, the relations do not have any direction, and the resulting similarity matrix is symmetric.

- s_{ij} is the similarity between node i and node j
- $S = [s_{ij}]$ is the entire similarity matrix

2.8.2 Setting a threshold

In WGCNA, there are two different options for how you would create an adjacency matrix from this similarity matrix. Option one: Use a hard threshold τ, or option two: Use a soft threshold β. Both of these options make use of a scale-free criterion for choosing a value for their respective parameters. When applying the scale-free criterion, the assumption is that the underlying network has a scale-free topology. As previously mentioned, this means that only a few nodes are highly connected while the majority have few connections. The assumption of scale-free topology as a commonly shared property of complex networks is not without its critics; And in reality, there might be other degree distributions that describe the network better (Broido and Clauset 2019). For our purposes, however, even if many complex networks only satisfies the scale-free topology approximately, simple and close enough is good enough (B. Zhang et al. 2005).

How close the distribution is to a scale-free distribution can be calculated as the square of the correlation between log P(k) and log k (B. Zhang et al. 2005), where k is the degree and p(k) is the degree distribution. A distribution is approximately scale-free when this value is close to one (B. Zhang et al. 2005).

For the hard threshold, getting an adjacency matrix means setting a value τ were all correlations above or equal to that value is 1, and everything below is 0.

$$a_{ij} = \begin{cases} 1 & \text{if } s_{ij} \geq \tau; \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

- a_{ij} is the adjacency between node i and node j
- $A = [a_{ij}]$ is the entire adjacency matrix

A question worth asking is how good a network with discrete on-off connections is at describing the underlying biological system (B. Zhang et al. 2005). What would, for example, be the biological difference between a correlation of 0.79 and 0.8 (B. Zhang et al. 2005)? The binary nature of a hard threshold also means that the resulting network, although very intuitive, is highly dependent on the choice of this threshold and and that setting an 'incorrect' threshold could entail information loss (B. Zhang et al. 2005), e.g., missing genes/bacteria interaction because it was just below the threshold.

Weighted correlation networks avoid this distinction by applying a function that scales the correlations continuously. Because the transformation is gradual, the soft threshold is more forgiving with regards to choosing slightly 'incorrect' thresholds (B. Zhang et al. 2005). In WGCNA, using a soft threshold means finding a value β for which all correlations should be raised.

Use a Signed or Unsigned Weighted Network?

$$a_{ij} = |s_{ij}|^{\beta} \tag{2.6}$$

$$a_{ij} = (0.5 + 0.5 \cdot s_{ij})^{\beta} \tag{2.7}$$

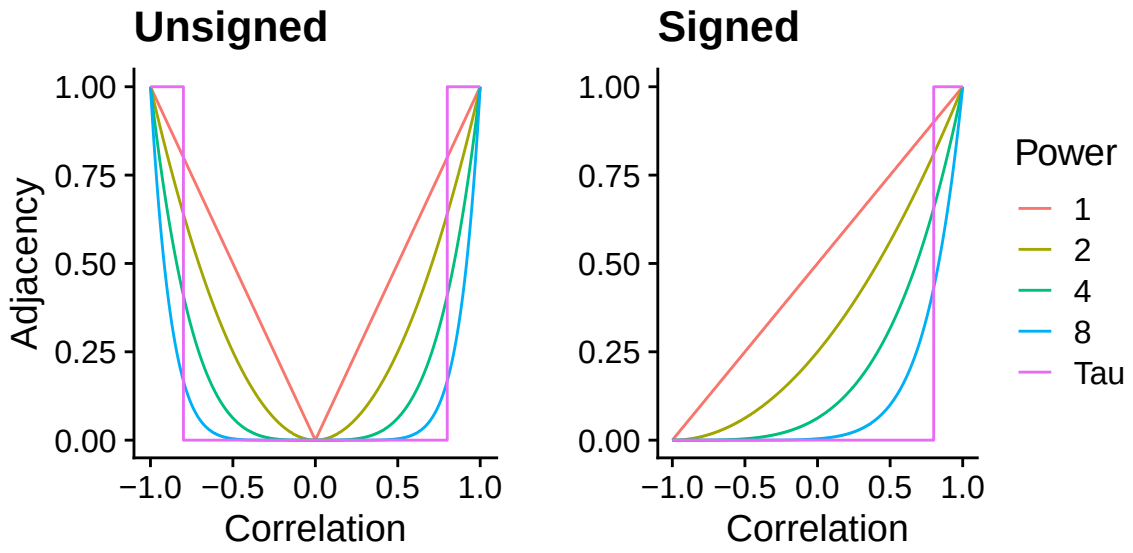

Figure 2.5: Effect of raising a correlation [-1,1] to a power β and the effect of a hard threshold $\tau = 0.8$. Adapted from figure 5.1 in (Horvath 2011).

In an unsigned weighted network, see equation 2.6 and figure 2.5, both high positive and high negative correlations result in a high adjacency. In a signed weighted network, see equation 2.7 and figure 2.5, negative correlations gets low to no adjacency, with a correlation of -1 giving an adjacency of 0 (Horvath 2011). By raising the correlation to a power, low correlations get penalized harder than high correlation values.

So what is best? A signed or unsigned network? A signed network groups only those genes/OTUs that are positively correlated, which is beneficial for biological interpretation, because positive and negative correlation have different meaning in biological systems e.g. a beneficial symbiosis can only be a positive correlation; Negative correlations also tends to come from other biological categories (Langfelder 2018). So, for gene ontology enrichment (detailed in section 2.9.1 on page 20) a signed network makes more sense. It is therefore recommended by the authors of the WGCNA R-package themselves to use a signed network (Langfelder 2018). As we will see later modules can be inter-correlated to connect negative and positively correlated genes/OTUs later.

A weighted network approach has the advantage of preserving the continuous nature of the connections, and can therefore avoid the loss of information that selecting a hard threshold brings (Horvath 2011). B. Zhang et al. (2005) show, using empirical and simulated data, that weighted networks can give more robust results than unweighted networks.

2.8.3 A measure of interconnectedness

It is possible to cluster and detect modules directly on the adjacency matrix, but standard procedure in WGCNA is to first transform this adjacency matrix into a Topological Overlap Measure (TOM) matrix.

The Topological Overlap defined by Ravasz et al. (2002) gives a measure of how close two nodes are to each other by comparing how many 1-step neighbours they have in common (Yip and Horvath 2007). Ravasz et al. (2002) found that a high topological overlap between substrates in an *Escherichia coli* metabolic network indicated an increased likelihood of belonging to the same functional class. Yip and Horvath (2007) and B. Zhang et al. (2005) generalized this measure for use with weighted networks. Weighted topological overlap measure leads to more cohesive modules than its unweighted counterpart (B. Zhang et al. 2005)

$$\mathrm{TOM}_{ij} = \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 + a_{ij}} \tag{2.8}$$

$$\mathrm{DistTOM}_{ij} = 1 - TOM_{ij} \tag{2.9}$$

2.8.4 Clustering & Module detection

The module detection in WGCNA is unsupervised based on the hierarchical clustering of the dissimilarity TOM see equation 2.9. The resulting branches of the dendrogram corresponds to network modules (Langfelder and Horvath 2008). There are many ways to proceed from here, either cutting at a static height, or as is the default in WGCNA, use the Dynamic Tree Cut method implemented in the package (Langfelder and Horvath 2008). The dynamic Tree cut method uses the shape of the branches as guides for how to best cut them (Langfelder, B. Zhang, and Horvath 2008).

The function *blockwiseModules* is a higher-level function that implements a fast and crude clustering method to split variables into a predefined number of clusters. These subsets (blocks) can then be analysed by complete pairwise correlation independently, as opposed to complete pairwise correlation between all genes. After this is done similar modules across the blocks are merged and module membership recalculated. So for the user there is no difference. Except for that there are now as many dendrograms as there are blocks. Although this results in non-optimal modules, the benefit is dramatically reduced computational demand; Making it possible to analyze 30000 + genes with under 4GB of RAM in under 2 hours. Something that would normally use 18-20 GB of RAM and >24

hours. This is the difference between running the analysis on a standard laptop or needing a high performance stationary computer or server.

After module discovery there are two options for summarizing the collective behaviour of the nodes in the module. 1) Find a highly connected node via the intramodular connectivity measure and use that as the modules representative, or 2) calculate the module eigennode, the first principal component of the expression/abundance of the nodes. The eigengene can be thought of as a weighted average expression/abundance profile (Langfelder and Horvath 2008). The intramodular hub nodes are highly correlated to the module eigennode (Horvath and Dong 2008).

2.8.5 Correlating modules to each other and to traits

How connected a node is in a module, its membership status, is called the intramodular connectivity of that node, k_{IM} (Langfelder and Horvath 2008). But the membership to a module can be defined for all nodes, not only those that were original members (Langfelder and Horvath 2008).

How significant an association between a gene and a trait is, is called the gene significance (Langfelder and Horvath 2008). Modules significance is the absolute average of the gene significance within that module (Langfelder and Horvath 2008). When we use correlation the eigengene/OTU tends to be highly related to this module significance and can therefore be used as an approximation (Langfelder and Horvath 2008). This also means that modules can be related to one another by correlating the module eigennodes (Langfelder and Horvath 2008). Those modules that are highly correlated can be merged (Langfelder and Horvath 2008).

Genes/OTUs that have a high membership, i.e. correlates highly to the eigengene/OTU they are candidates for further validation (Langfelder and Horvath 2008).

2.9 Gene Ontology

The observation that many organisms have very similar genes (orthologs), and that these genes share the same function, lead to the development of a shared central resource with structured and controlled vocabulary for describing gene function (Plessis, Skunca, and Dessimoz 2011). By using Gene Ontology (GO) terms it is possible to use knowledge gained in one organism to infer the function in another (Plessis, Skunca, and Dessimoz 2011). The GO project is the largest and most widely used resource for cataloguing gene function and gene products. Equally, or arguably of more importance, its structure allows for computational inference – paramount for modern biological research (Ashburner et al. 2000; The Gene Ontology Consortium 2019).

Gene ontology terms are structured in a directed acyclic graph that starts with three non-overlapping ontologies:

- CC: Cellular Component which represent locations in the cell where the gene product is active.
- BP: Biological Processes describes series of interacting physiological and biochemical events.
- MF: Molecular Function describes what mechanisms or roles a gene product may perform.

2.9.1 GO enrichment analysis

GO enrichment is a statistical analysis of over-representation of GO terms. E.g. How many genes from a specific term is observed compared to how many was expected to be observed. In addition to the classic over-representation test, R-packages like topGO include many other methods that use the structure of the directed acyclic graph, to inform the method and make better over-representation estimates (Alexa and Rahnenfuhrer 2018).

2.10 PC-correction – Regressing out global effect variables

The PC-correction method of Parsana et al. (2019) from the R-package *sva* is intended to remove all latent confounding artifacts from a co-expression dataset so that better co-expression networks can be made. How many such latent variables that should be removed is estimated by an permutation based approach in the *num.sv* function (Parsana et al. 2019). The function *sva_network* is then used to remove that number of estimated latent variable from the data. PCA has been previously used to correct batch effects and other unwanted sources of variation (Leek et al. 2012). Note that this is from the same research group.

The PC-correction method depends on the idea that most broad correlation, i.e those that affect many genes, is due to confounding variables, and that the majority of genes are only connected to a limited number of other genes i.e scale-free topology. Because of this, principal components can safely be removed without removing any real relationship between genes (Parsana et al. 2019). Parsana et al. (2019) find that this approach can reduced the number of false positives when constructing gene correlation networks.

The effect is achieved through 'regressing out' the PC. Because only a number, how many PCs to remove is given to the $sva_network$-function it can also be used to manually remove the most influential variables.

Chapter 3

Material and methods

3.1 Materials

The data used in this thesis comes from a long term feeding trial of farmed Atlantic salmon (*Salmo Salar*) see figure 3.1. The salmon were raised on two contrasting diets, one with low amounts of long chained polyunsaturated fatty acids (LC-PUFA), containing a 1.8:1 ratio of linseed oil and palm oil, and one high in LC-PUFA, based on North Atlantic fish oil (Gillard et al. 2018; Rudi et al. 2018). The feeds will from here on out be the referred to as vegetable oil feed (VO) and fish oil feed (FO).

There are two types of omics data from this feeding trial. From the Atlantic Salmon host we have RNA-seq from the gut and liver. Salmon (Patro et al. 2017) was used to quantify the expression of transcripts from the RNAseq reads. Details about the sampling and sequencing, as well as data processing can be found in Gillard et al. (2018).In this thesis we will only make use of the gut transcripts.

From the gut microbiota we have metagenomics – 16S rRNA amplicon sequencing (16S-seq, V3-V4). Details about the collection of the samples and sequencing can be found in Rudi et al. (2018). The 16S sequences were made into OTUs with the USEARCH pipeline using 97% identity. NB! OTU-construction was not done by me but by one of the co-authors of the Rudi et al. (2018)-paper; Classification was done with the R-package *microclass*.

Links to the raw data can be found in both their their respective papers.

Figure 3.2 shows which time-points (Day) in combination with feed that was sampled. For each combination there several replicates, but not all replicates are available for every combination. Figure 3.2 shows the available samples.

Previous analysis of the transcriptomics (RNA-seq) has shown that the host's lipid metabolism is plastic and adaptable to changes in dietary lipid when the fish is pre smolt (before smoltification, see section 1.2 on page 2), but that the fish loses this plasticity after smoltification (Gillard et al. 2018).

Previous analysis of the metagenomics (16S-seq) has shown that there is a dramatic shift in microbial composition as a consequence of the salmon's habitat transition from freshwater to saltwater, but that there is no significant effect of feed on the microbiota (Rudi et al. 2018). A majority of operational taxonomic units (OTUs) showed decreases in saltwater ($>$ 75%), while OTUs with a high relative quantity increased (Rudi et al. 2018). Rudi et al. (2018) conclude that it is unlikely that the gut microbiota has a role in compensating for the low levels of LC-PUFA in freshwater. But they do describe a set of four OTUs that were stable across the freshwater-saltwater divide. All these OTUs were placed in the phylum *Firmicutes* (Rudi et al. 2018).

Dominance of Firmicutes at both parr and postsmolt stages while Wild salmon is dominated by Proteobacteria (Rudi et al. 2018). Freshwater and saltwater environments share

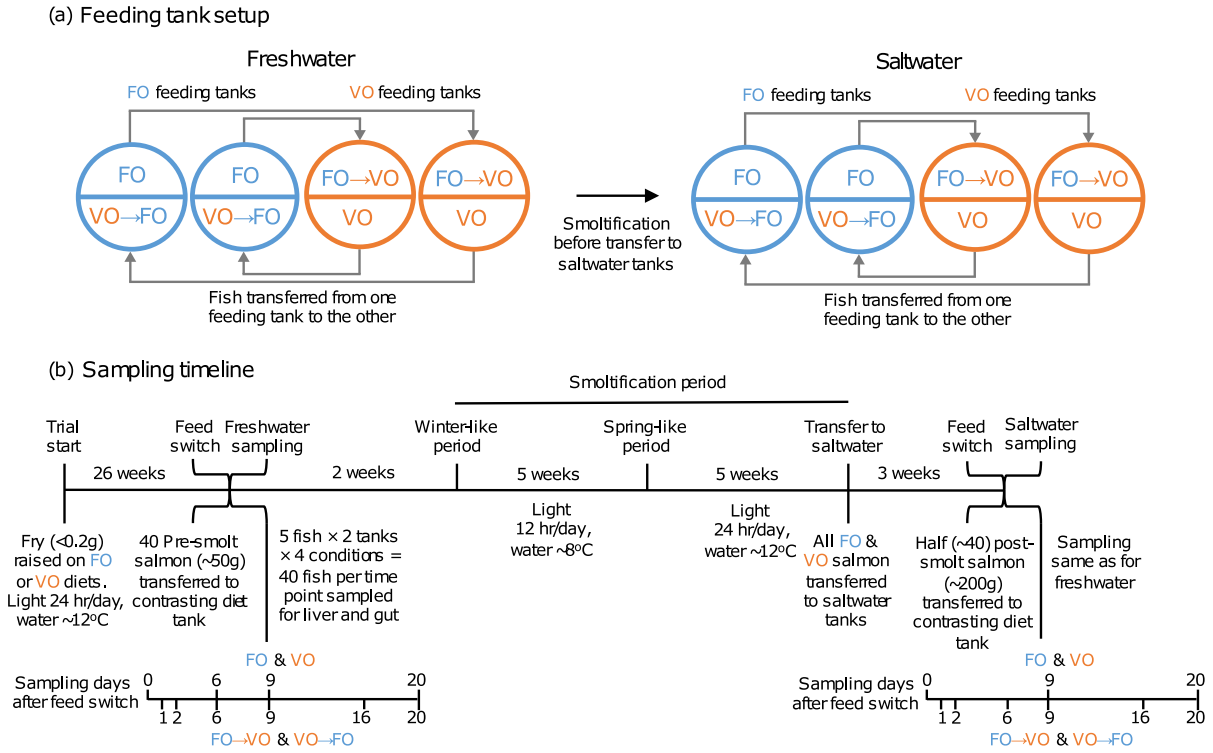

Figure 3.1: Overview of experimental set up and timeline (a) Atlantic salmon fry was reared in four different tanks which each had two compartments. Fish in two tanks halves were continuously fed feed with vegetable oil as its lipid source (VO), fish in the the other two tank halves were fed feed with fish oil (FO). Switching feed involved taking fish from VO and FO tanks and moving them to empty tank haves with the opposite feed regimen. Only fish from the non-switched tanks were transferred to new tanks containing saltwater after smoltification – There the same process was repeated. (b) Experiment and sampling timeline. In total the experiment lasted ∼47 weeks. There are two sampling times each lasting 20 days, one in freshwater 26 weeks after hatching and one three weeks after introduction to saltwater. Figure taken from Gillard et al. (2018).

818 OTUs, roughly 69% of total (Rudi et al. 2018). There was a higher number of unique OTUs in freshwater than for saltwater (Rudi et al. 2018).

Traits/external variables of the samples are as follows:

- Environment: transition from freshwater (FW) to saltwater (SW).
- Day: Time in days since start of the sampling. Starts at 0 (D0) and ends at 20 (D20) for freshwater and saltwater independently.
- Feed: There are two different feeds; Fish oil (FO) a diet rich in long chained poly unsaturated fatty acids (LC-PUFA); Vegetable oil (VO) a diet with low amounts of LC-PUFA.

3.2 Methods

The main aim of this thesis was to apply the WGCNA framework as a way to integrate multi-omics data to analyse host-microbiota data. Although WGCNA was developed for analysing gene expression, and more specifically microarray data, there is nothing that fundamentally stops this approach from working with other types of data.

It is important to note here that causal relationship can not be inferred with this approach. Only correlations. This approach also deals with the collective behavior of objects,

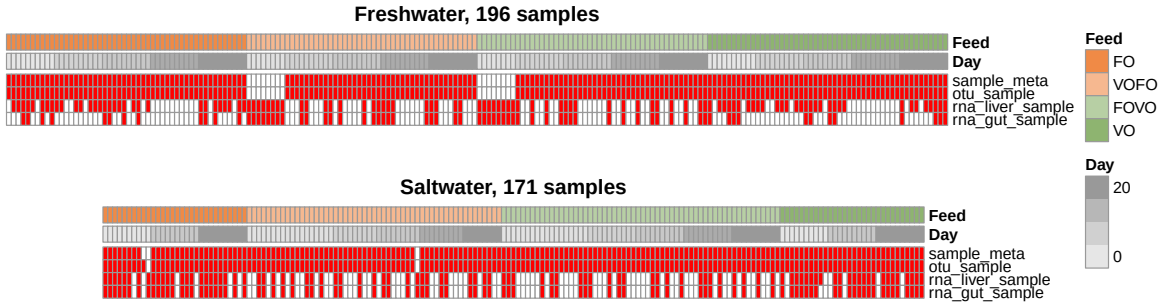

Figure 3.2: Sample overview. VO: Vegetable oil, low LC-PUFA. FO: Fish oil, high LC-PUFA. Red indicates that there is a sample, white indicates that there is not.

in this case gut microbes and host genes. This allows us to gain a more global view of the data.

The WGCNA package contains higher level functions for network creation and the detection of modules, but there also exists lower level functions for fine tuning parameters, and functions that allows you more direct control of similarity measures.

The function *pickSoftThreshold*, gives a data frame output for which the choice of soft threshold can be picked. The power at which the distribution is ≥ 0.8 or more without the mean connectivity becoming to low is chosen as the appropriate threshold (Langfelder and Horvath 2008).

The function *blockwiseModules*, wraps all needed functions from data matrix to module classification. It also implements a way of crudely pre-clustering the data into a predefined number of groups. Full calculation of similarity can then be calculated separately on each group dramatically reducing the computational resources needed. Modules will not be as good as when all data is analyzed together, but are often close enough to the result you would get doing the calculation on the entire dataset.

For both the RNAseq and the marker gene data a signed network was used, with "bicor" as correlation function.

3.2.1 Applying PC-correction

Here the PC-correction method described in section 2.10 on page 20 was applied to try to regress out dominant features of the data that is thought to not be related to host-microbe interactions. Two variations of the approach was 1) The removal of only one PC 2) Estimate the number of PCs to remove from each dataset (RNA-seq, 16S) and remove those. The validity of such efforts is discussed in the discussion.

3.2.2 Outlier detection

Samples where first normalized with a min-max normalization, known to be very sensitive to outliers (Cao, Stojkovic, and Obradovic 2016), then clustered with average distance to identify possible outliers.

There might sometimes be no obvious technical reason for the outliers, but because such samples can have undue influence on the results it does not matter what they represent, they should be removed before continuing the analysis.

3.2.3 Intersection of samples

Samples from the RNA-seq and the 16S-seq was intersected so that there was a one-to-one correspondence.

3.3 Description of the general pipeline

The figure below 3.3 illustrates the general pipeline of this thesis. The pipeline is intended as a frame work

Part 1. Pre-processing

Part 2. Data appropriate processing

Part 3. Network construction and module detection

Part 4. Pre-processing of non-omics data

Part 5. Visualization of correlations

Part 6. Detailed exploration of selected modules.

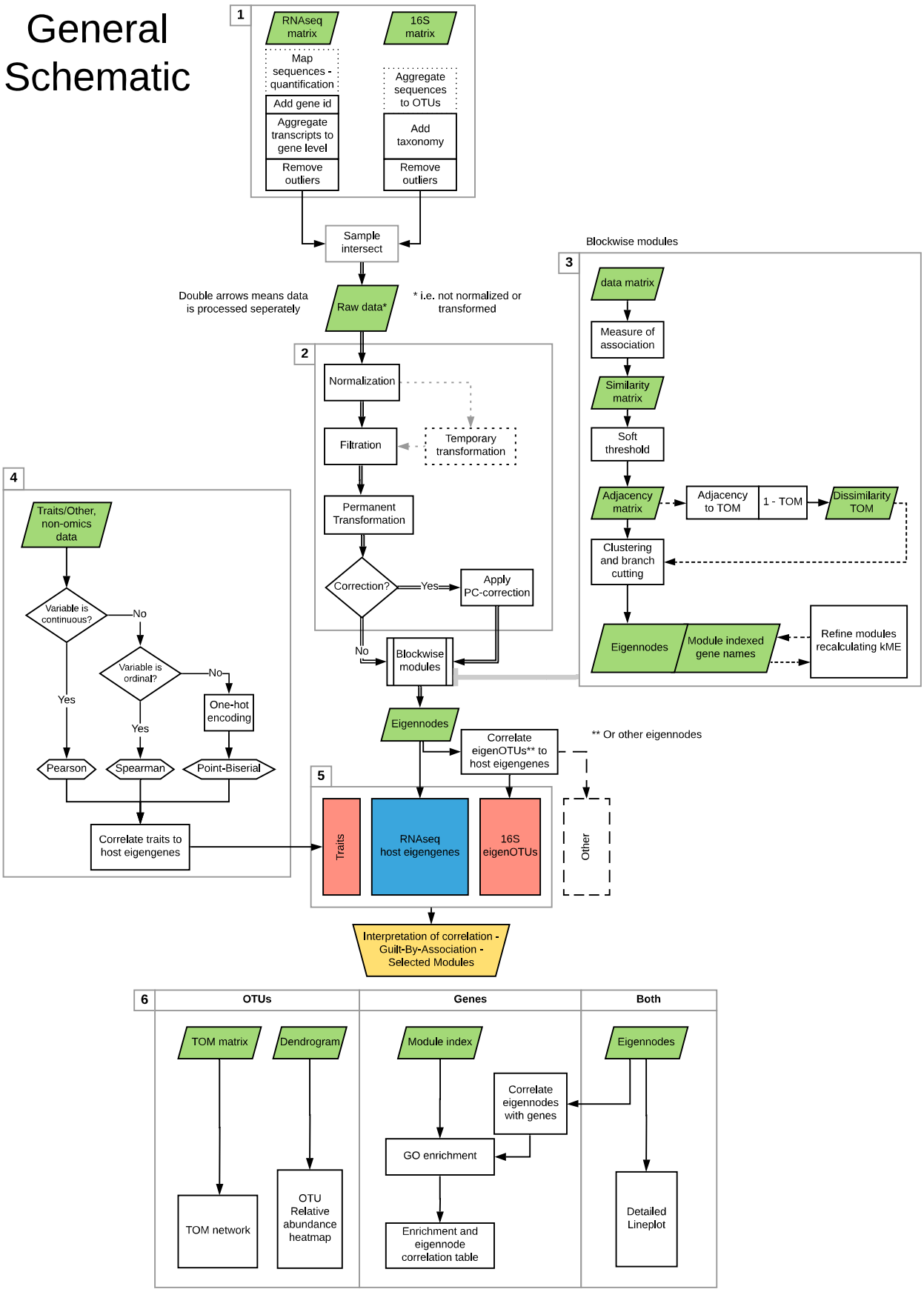

Figure 3.3: **General pipeline schematic**. Numbers indicate the progression of the pipeline. 1. Input and preparation of data, 2. Normalization and transformation, 3. Block-wise construction of a weighted network or customized weighted network construction, 4. Processing of non-omics data. 5. Visualization of correlations between modules and sample traits, 6. Detailed exploration, or selected modules.

3.4 Application of pipeline

3.4.1 Part 1-3 – RNAseq – Host transcriptomics

Part 1 – Set up

The quantified transcripts were reduced to gene level.

A few reasons why reduction to gene-level is preferable:

- Quantification of gene isoforms is still uncertain; Especially for non-model organisms like the Atlantic salmon (*Salmo Salar*).
- It will make using and interpreting Gene ontology enrichment easier.
- It will reduce the number of variables in the data and speed up data processing.

For the Atlantic salmon, conversion between transcript ids and gene ids was done with the R package Ssa.RefSeq.db, an R package that provides annotations for the Atlantic salmon genome ICSASG_2.

There are three options regarding the reduction to gene-level; Either use the longest transcript, Ssa.RefSeq.db has this option as default; Take the average/median, etc., or summarize genes by adding together the expression of each gene isoform. For this pipeline summation to gene level was chosen.

Part 2 – Data appropriate processing

Filtration

Filtration of low counts is important because low counts can increase the number of spurious correlations. Genes and OTUs with low variance are less likely to have any biological importance.

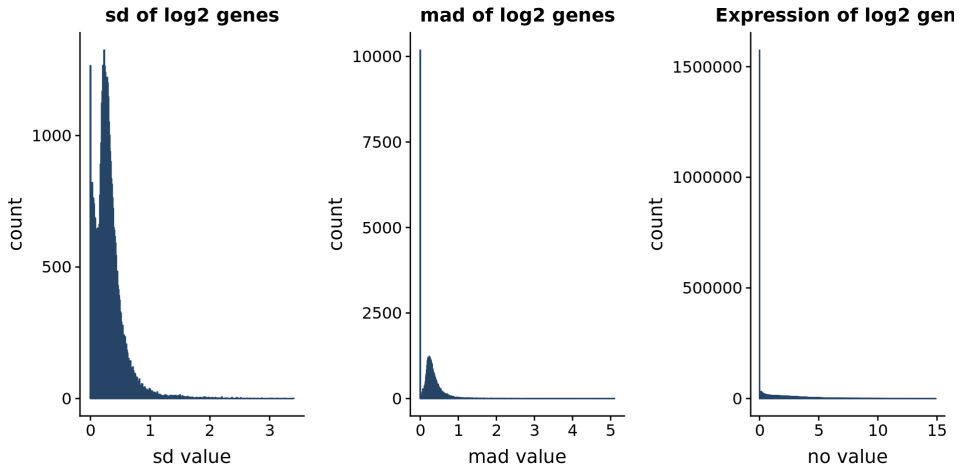

Figure 3.4: Histograms of the standard deviation of each gene, Median absolute deviation of each gene and the expression of all genes before filtering

For the RNA-seq data, genes with expression less than a maximum of 1 and then those with a standard deviation less than 0.15 was removed. The thresholds came from visualization of the expression distribution of counts on a +1 shifted log 2 scale. See figure 3.4. The effect of removing those genes can be seen in figure 3.5.

Normalization and transformation

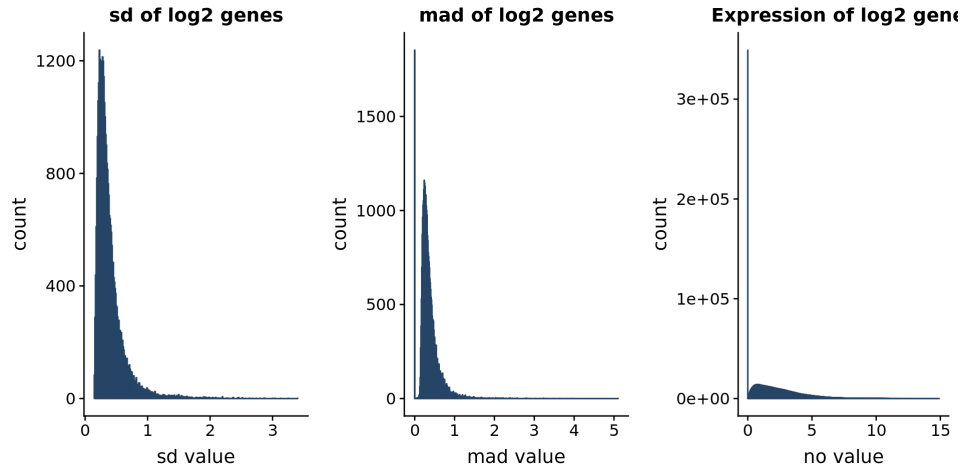

Figure 3.5: Histograms of the standard deviation of each gene, Median absolute deviation of each gene and the expression of all genes after filtering

For the RNA-seq data, the results returned from the mapper is output as transcripts per million (TPM), which is adjusted for the length of genes and library sizes. For using other normalization techniques that are based on raw counts, the R-package tximport can recalculate "raw" counts from the salmon-mapper output (Soneson, Love, and Robinson 2016). Here, tximport was used to create "raw" counts, while still adjusting for the length of each gene (lengthscaledTPM). From this "raw" count the normalization method TMM was used 2.3

Part 3 – Network construction and module detection

Standard procedure using the function *blockwiseModules*. Defaults are used if not specified below. Refer back to Background section 2.8.2 for details about signed or unsigned networks.

- networkType = "signed"
- TOMType = "signed"
- maxPOutliers = 0.05
- pamStage = F

maxPOutliers was set to 0.05, to allow for correlation to binary traits.

3.4.2 Part 1-3 – 16S rRNA - Microbiome marker gene survey data

Part 1 – Set up

It is preferable to "fill in" the taxonomy table with all available information leaving no taxonomic rank Not Available (NA). Making the taxonomy table complete helps keep all OTUs whether they are classified or not. The reason for this is that R remove NA "values" for some plots, even though R often gives a warning. Removing an OTU just because it is not classified is not logical choice in an exploratory analysis.

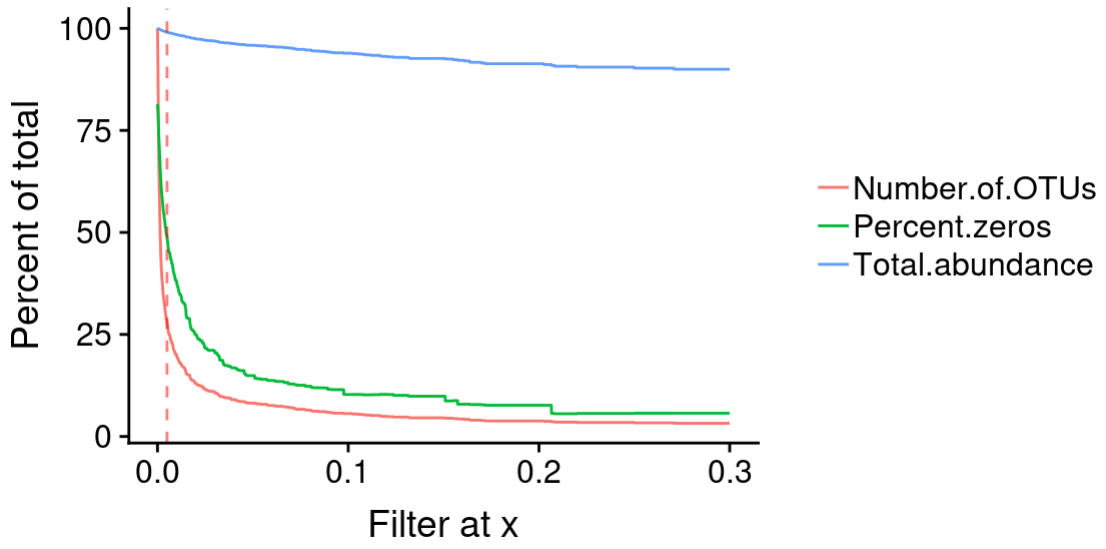

Figure 3.6: Drop off of number of OTUs, total abundance and percentage of zeros in the dataset as a consequence of filtering. The red vertical line shows the cut off at 0.005% of total relative abundance.

Part 2 – Data appropriate processing

Pre-filtering:

The OTU data contains a lot of zero abundances, OTUs that are present in one sample not in another, in fact, in relative percentage over 75% of the entries in the data matrix are zeros.

In figure 3.6, it is clear that many OTUs contain a lot of zeros and contribute little to the total abundance. By choosing a cutoff where we only keep OTUs that contribute 0.005% or more of the total abundance, the percentage of zeros is reduced to ∼50%. And only around 25% of the total number of bacteria remain, corresponding to 296 OTUs – down from 1152 in the original OTU-table.

Normalization and Transformation:

The remaining OTU data matrix was normalized with Cumulative Sum Scaling (CSS) from the Bioconductor package metagenomeSeq which by default also log2 transforms the data.

Part 3 – Network construction and module detection

Some adjustments have been made to accommodate the OTU data. i) The minimum module size has been reduced from the default of 20, to 5. ii) Pam stage is set to FALSE

The community structure, or what you can see from 16S data, is smaller with OTUs, and there does not seem to be any benefit of using the partitioning around medoids (PAM) step. In using this step, the eigenOTU became very influenced by OTUs that were not very central to the module. This is in contrast to the gene data, where PAM seemed to have no real influence, most likely because the eigengene there was more stable to the inclusion of a smaller number of peripheral genes.

- networkType = "signed"
- TOMType = "signed"
- maxPOutliers = 0.05
- pamStage = F

All other parameters were kept at the default. maxPOutliers is set to 0.05, to allow for correlation to binary traits.

3.4.3 Part 4 - Correlation of module eigennodes

The quantified expression/abundance of the transcriptomics and metagenomics data are discrete, but because both have a large range and have been log2 transformed and averaged they can be treated as if they are continuous.

The resulting eigengene/OTU modules are centered and scaled, with "normal" distribution.

Choosing the correct correlation measure

All values will be averaged to avoid underestimating the p-value (overestimating significance of correlation).

The base R function *cor.test* was used with default settings unless otherwise noted.

The null hypothesis can be stated as such: The correlation coefficient is not significantly different from 0, that the co-variation is likely to happen due to chance. The alternative hypothesis is that there is a sufficiently large positive or sufficiently large negative coefficient that not not likely to happen due to chance.

Module eigenOTUs are correlated to the module eigengenes. Both are continuous with normal distributions. Therefore Pearson correlation was the best choice as correlation coefficient:

Let hME_i be the i-th host (gene expression) Module Eigengene.
Let mME_j be the j-th microbe (Abundance) Module EigenOTU.

$$cor(\text{hME}_i, \text{mME}_j)$$

for all i in i = 1,2, ..., N_i, and for all j in j = 1, 2, ..., N_j.

Traits/external variables are correlated to the module eigengenes. When it comes to these non-omics variables there are more considerations to be made. To obtain correct results we need to be aware of the assumptions that underlie the correlation test. To correctly measure any correlation on the non-omics data we first identify what type of variable they are.

We can separate our external variables and traits into categories.

There are four categorical variables: Day, host_sex, environment and Feed. Categorical variables are variables that can be divided into categories, either naturally (apple, orange, ...) or imposed through subdivision of a continuous variable e.g. price groups: cheap, average or expensive.

Of those four categorical variables, three are nominal and one ordinal. The difference between nominal and ordinal is that the ordinal variable has a natural ordering. Day is one such variable, e.g., day one comes before day two, which comes before day three etc.

Of the three that do not have any natural ordering two are naturally dichotomous i.e. there are only two levels possible. The sex of the host fish is either female or male. The environment, or 'Water' as it will later be called, is either freshwater or saltwater. These variables can be directly encoded by replacing either of the levels with 0 or 1. What gets

encoded as a 0 or 1 will change whether correlations with other variables are positive or negative.

The last categorical nominal variable is Feed. Although there are only two possible feed types for this experiment vegetable oil (VO) and fish feed (FO), some fish also experienced perturbations in their feed changing to the opposite feed at different time-points. This variable can to simplify be divided in four different categories VO and FO, including change from VO to FO and FO to VO. For variables like this, the simplest approaches to calculate a correlation coefficient, and one that is often used in machine learning, is to dichotomize the variable with a one hot encoding. A one hot encoding considers each level against every other level, encoding one level as 1 and all others as 0. This means that we now instead of one Feed variable have four, which can then be correlated in the same way as the naturally dichotomous variables above.

Last but not least we have two continuous variables, host weight and host length. These variables are however highly correlated with each other and including to Day and Water. Instead of adding host weight and length directly we can instead calculate the condition factor (CF) of the fish (Barnham and Baxter 1998; Alne et al. 2011). The new variable CF is also a continuous variable. The CF can be thought of as an indirect measure of fish fatness (Barnham and Baxter 1998).

$$\mathrm{CF} = \frac{10^{\mathrm{N}} \cdot \mathrm{W}}{\mathrm{L}^3}$$

Where: W is the weight of the fish in grams, L is the length of the fish in millimeters and N is a constant used to bring the range of output values close to 1. The value of N differs from N = 2 in Alne et al. (2011) to N = 5 in Barnham and Baxter (1998). When calculating the CF we find that only a value of N = 4 gives the expected range, while the values used in Alne et al. (2011) and Barnham and Baxter (1998) do not. Note here that this scaling is irrelevant for the correlation and as such any scaling is purely cosmetic.

In summation, all module eigengenes are correlated with every trait/external variable:

Let T_k be the k-th trait or external variable.

$$cor(hME_i, T_k)$$

for all i in i = 1,2, ..., N_i, and for all k in k = 1, 2, ..., N_k.

- All hME_i are continuous.
- For cases where T_K is also continuous $\longrightarrow$ Pearson correlation
- For cases where T_K is ordinal $\longrightarrow$ Spearman correlation
- For cases where T_K is nominal:
  - And naturally dichotomous $\longrightarrow$ Point biserial
  - Not dichotomous $\longrightarrow$ one hot encoding $\longrightarrow$ Point biserial

Point biserial is a special case of the pearson correlation and as such all these variables can use *cor.test* with the default Pearson correlation. Spearman correlation uses $cor.test(method = "spearman")$.

3.4.4 Part 5 - Visualization of correlations

Correlations from the previous section are then gathered into one figure consisting of three heatmaps. This "Split heatmap" places the host gene expression module eigengenes at the center, with correlation values of the non-omics variables to the left and correlation values of the microbial abundance module eigenOTUs to the right. In this way the figure gives an overview of the entire dataset. The center heatmap is the expression of the host gene modules eigengenes (hME), this represent the general behaviour of the host genes in each module.

To get common dimensions with the Host Module Eigengenes (hME), sample characteristics and Microbe Module Eigengenes (mME)/EigenOTUs were correlated to the hME as detailed in the section above. In each cell of the correlation heatmaps the p-value significance of the correlation is marked with stars. One star (*) represents a p-value below or equal to 0.05, two stars (**) a p-value below or equal to 0.01 and three stars (***) a p-value below or equal to 0.001. P-values are not corrected and should be seen as indications of strength rather than rigorous hypotheses tests.

3.4.5 Part 6 - Detailed exploration of selected modules

GO enrichment analysis

The R package TopGO (Alexa and Rahnenfuhrer 2018) was used to test for gene enrichment.

TopGo creates an object which contains all necessary information needed to do the GO enrichment analysis.

Here the both the classic algorithm + fisher statistic, and weight01 + fisher statistic was used in the TopGO object. For the classic method "Each GO category is tested independently" (Alexa and Rahnenfuhrer 2018). The Weight01 method is a combination of the elim and the weight algorithms, which were both introduced in Alexa, Rahnenführer, and Lengauer (2006).

Gene universe, also called the background set, is the set of all feasible genes i.e. all genes that were used as input in the network and module construction step.

Selection of the genes to be tested for over-representation was not chosen based on direct output membership but instead all genes in the gene universe were correlated to the eigengene of the module selected, creating a fuzzy set. Such a fuzzy set could be used for a Gene Set Enrichment Analysis (GSEA), by looking for enrichment of function in either end of the weighted (correlation) list. Here, however, GO enrichment type analysis was chosen instead as the implementation was easier. By restricting the set of genes to those with a correlation of $>= 0.8$ we get a set of genes for which we can look for over-representation. This strict selection of 0.8 ensures that the genes that are chosen are central to the module, but this threshold is arbitrary, and any threshold could be used given some justification. The set of genes that remained after selection was then used in the GO enrichment analysis.

The GO enrichment results are visualized in a bubble plot, directly inspired by the bubble plot of the *GOplot*-package (Walter, Sánchez-Cabo, and Ricote 2015).

3.5 Comparison to principal component analysis

Results from the PCA went though the same steps as those of weighted network analysis. The base R function *prcomp* was used to calculate principal components for both the transcriptomics and the metagenomics.

3.6 Randomizing data

A simple non-parametric approach was used to investigate the base correlation level, as it is expected that some sources of variance skew the correlation. Using the base R function *sample* with argument 'replace = TRUE' on the OTUs of the OTU-table result in an overall distribution that is identical to the real distribution. The same approach, but a more extensive version was used by Friedman and Alm (2012). On average there should after such a redrawing of abundances be no correlation between OTUs outside of random chance.

3.7 Reproducibility

All analysis was done in R version 3.5.0 "Joy in Playing" (R Core Team 2018). All analysis were run as R-markdown files which are numbered after the order they should be run in. This means that result figures and tables in this thesis will be presented alongside the code that generated them. Some self-made functions are sources in R-scripts to keep the R-markdown files from becoming to large.

The R-markdown files and R-scripts are available at:
`https://gitlab.com/M.Strand/Weighted_Host_Microbiome_Correlation_Network`

Plot are either made in ggplot2 (Wickham 2016) or with base R graphics. Any editing of figures was done in Inkscape version 0.92. General method schema and the eukaryotic gene expression regulation diagrams were created in lucid chart (`www.lucidchart.com`).

Because some of these scripts take a long time to run, and might require more RAM than is available at a standard laptop or stationary computer (depending on the parameters set), R-markdown files were run on the NMBU computational cluster via the SLURM workload manager. For those that want to replicate this, there is a bash-script in appendix .1 on page 78.

Chapter 4

Results

4.1 General overview

This chapter describes the use of our developed pipeline to discover new putative functional interactions between the host and its microbial communities. See an overview of the pipeline on page 25.

As detailed in the pipeline schema, we first construct networks separately for each data type; One for the host gut gene expression and one for the gut microbiome abundance data. We then find network modules and module eigennodes. The module eigennodes are correlated between both data types and to any non-omics variables, as described in section 3.4.3.

The dataset used in this thesis contained samples taken both in freshwater and saltwater. This transition is associated with a massive shift in both host gene expression and microbial community composition, which could overwhelm all other more subtler effects, e.g., the effects of the microbial community impact on host gene expression in a few pathways. To look beyond this dominating fresh-salt transition, we could have considered the data from salt and fresh water as two subsets and analyzed them separately. However, having smaller datasets will possibly also impact our ability to build biologically relevant networks. An alternative to analyzing the salt and freshwater samples separately would be to apply a computational method to remove large effect variables (e.g., freshwater-saltwater transition in this study), enabling us to utilize the full dataset for network construction. In our case these large effect variables can be referred to as "confounding factors", because even though the fresh-saltwater transition is an integral part of our experimental design, they could confound our ability to identify smaller effects such as those of the different feeds.

With this in mind and no prior experience in analyzing this type of data with our pipeline, we decided to try keeping the data as is, and to remove large effect variables and compare their performances. We, therefore, analyzed two versions of the dataset:

- All data (without removal of large effect variables)
- All data with the removal of large effect variables

The removal of confounding variables was done with the PC-correction method from Parsana et al. (2019), because this method pertains directly to co-expression networks.

4.2 All data (without removal of large effect variables)

4.2.1 Network characteristics

The WGCNA authors recommend to always adapt parameters of the WGCNA functions to fit the data in order to obtain more biologically relevant modules than those obtained by default values. In this regard, to better understand the module characteristics, a six-part figure was created to summarize some of the essential aspects of the network construction.

Figure 4.1 and figure 4.2 show the network diagnostics for the host gut gene-expression network and gut microbiome abundance network, respectively. The following text explains the layout for both these figures, as well as other network diagnostic plots presented in later sections.

Part-A: Scale independence measure. Scale-free topology fit index as a function of the soft-thresholding power (β). The plot shows how well the degree distribution at different powers of beta fits (R^2) with a scale-free degree distribution. A model fit of 1 means that the network has a degree distribution that perfectly matches a power-law distribution (scale-free topology). For our purposes, a network can be said to be approximately scale-free if it achieves a model fit of > 0.8 for a reasonable power β. What is and is not a reasonable power is judged in connection with the mean connectivity in part B.
Part-B: Mean connectivity measure. Mean connectivity as a function of the soft-thresholding power β. Mean connectivity should not be too low, or the module detection might not work (Langfelder and Horvath 2014).
Part-C: Module size. The bar plot shows the number of nodes (genes/OTUs) in each module. WGCNA assigns a number and colors to each module. The first module is module 1 (Turquoise); it contains the largest number of nodes. The next largest module is named module 2 (Blue), then 3 (Brown), et cetera. Those nodes that do not fit into any module is left unassigned in the grey/0 'module;' As such, the grey/0 'module' cannot be interpreted as a proper module.
Part-D: ME-correlation density plot. The figure shows the correlation between module eigennodes (MEs). Although each module contains a distinct set of nodes (genes/OTUs), the eigenvector profiles of each module can still have similarities. The important aspect to look for here is that the mean correlation should be approximately zero.
Part-E: ME-correlation heatmap. The figure shows the same information as part C but as a heatmap. The heatmap makes it possible to see which module eigennodes correlates and if there are large groups of module eigennodes that strongly correlate.
Part-F: Within module correlation. The plots show the correlation of the nodes in each module to the eigenvector (density plots). Although WGCNA uses connectivity on TOM networks to define module membership, the intra modular membership can be approximated by correlation to the eigenvector when the measure used to make the network is correlation. This means that the eigennodes can be used as a proxy for working directly with the modules. The density plots are all for individual modules, including the improper module grey/0.

The method outlined in this thesis deals primarily with module eigennodes (eigengenes and eigenOTUs). Because of this, these module eigennodes (MEs) must be acceptable representatives of their respective modules. The correlation of nodes to the respective MEs measures how well the MEs represents the module (part F). The stronger the (positive)

correlation, the better. Also, the MEs must not be strongly correlated to each but instead have a mean correlation of zero (part-D, and -E).

Figures 4.1 and 4.2 show the properties of the networks for host gene expression and microbial abundances, respectively.

The diagnostic plot for the host gene-expression network (Figure 4.1A) shows that the network has approximately scale-free topology at $\beta = 3$ with a relatively high mean connectivity ~ 5600 (Figure 4.1B). The module eigengene diagnostics (Figure 4.1D and 4.1E) show that most eigengenes are not highly correlated . WGCNA detected 15 modules in the uncorrected host gene-expression data. The module sizes (Figure 4.1C) range from 49-9121 genes, while 11251 genes were not assigned to any specific module (grey bar in Figure 4.1C). Figure 4.1F shows how genes in each module correlate to their respective module eigengenes (MEs). The figure shows how the grey unassigned genes do not have any structure as the genes within that module do not correlate with the eigengene hME0. Other module eigengenes behave as expected, where the intramodular correlation is strong and positive.

The microbial co-abundance network properties (Figure 4.2) were distinctly different from the gene-expression network analyses (Figure 4.1). At $\beta = 8$, the network reaches approximate scale-free topology (Figure 4.2A). Mean connectivity is 6.44 at $\beta = 8$ (Figure 4.2B). Compared to the host gene network, the mean is much lower, but the size of the microbiota network is also much smaller. Comparing the networks on a more equal basis, mean connectivity divided by all possible pairwise combinations of nodes, the mean connectivity is still smaller for the gut microbiome network: Host gene expression network $5622.64/\binom{37408}{2} \approx 8.04e^{-4}$, gut microbiome network $6.44/\binom{296}{2} \approx 1.15e^{-4}$.

Only four co-abundance modules were identified, ranging from 13-36 OTUs per module. The unassigned group (grey 'module') was large relative to the other modules and compared to the gene co-expression network, containing 70% of all OTUs (Figure 4.2C). The limited number of modules and the fact that most OTUs falls outside of any module indicates that there is not much structure in the microbiota data compared to the gene expression data. This lack of structure indicates that many bacterial species have variation in abundance across the samples, or that the correlation between OTUs was in groups of less than five since that was the minimum module size set for the microbial abundance data.

The intramodular correlation is high (Figure 4.2D and 4.2E) and we only see two types of relative abundance behavior. Most likely, this reflects the freshwater-saltwater transition. Module eigenOTUs represent their modules well, with a high average mean correlation ~ 0.75 (Figure 4.2F). Surprisingly, the OTU correlations to mME0 are shifted towards low positive correlations, which indicates that there is some similarity among these OTUs, although not strong enough to form proper modules. For all the following analyses, the eigennodes of the grey 'unassigned' groups are omitted (except in section 4.3 – Comparison to Principal Component Analysis).

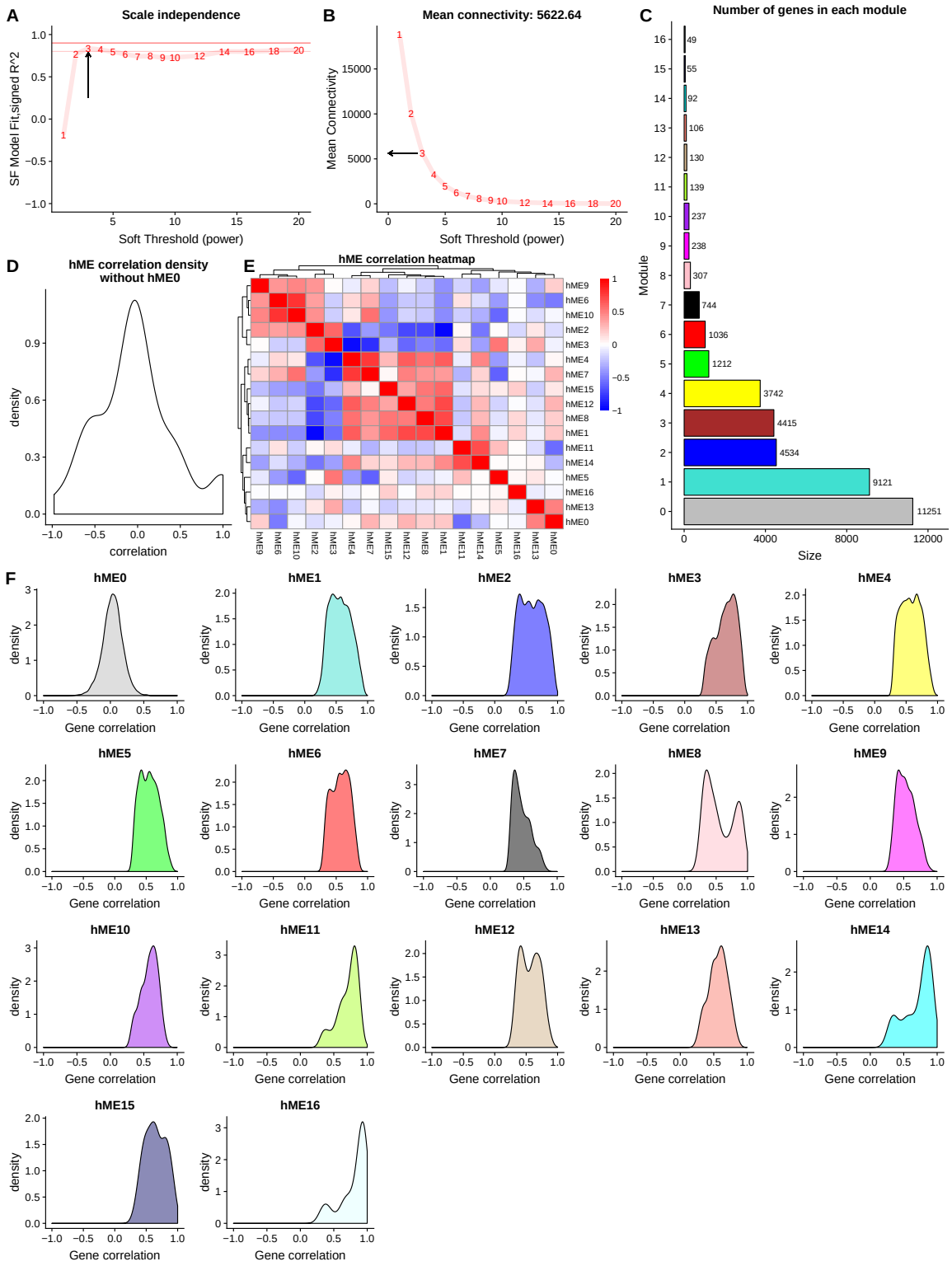

Figure 4.1: **Diagnostic plots for host gene co-expression network – All data. A)** Scale independence. Shows how the degree distribution of networks at different betas fits a scale-free distribution. Arrow shows the beta value chosen for the network. **B)** The mean connectivity of the network. Should monotonically decrease and should not be too low. Mean connectivity as chosen beta value is marked with an arrow, and the numerical value of the means is shown in the title of the plot. **C)** Shows the size of modules i.e. how many genes each module contains. **D)** Density plot of the correlation between host module eigengenes (hMEs), peak of the density should be at zero. **E)** hME correlation heatmap. A strong positive correlation is deep red, and a strong negative correlation is deep blue. A correlation of 0 is white. **F)** Correlations of genes within each module to their respective hME. Because the network is signed all genes within a module should have strong positive correlations i.e. have a peak above 0 and preferably > 0.5. The unassigned genes in grey should have an average correlation of zero.

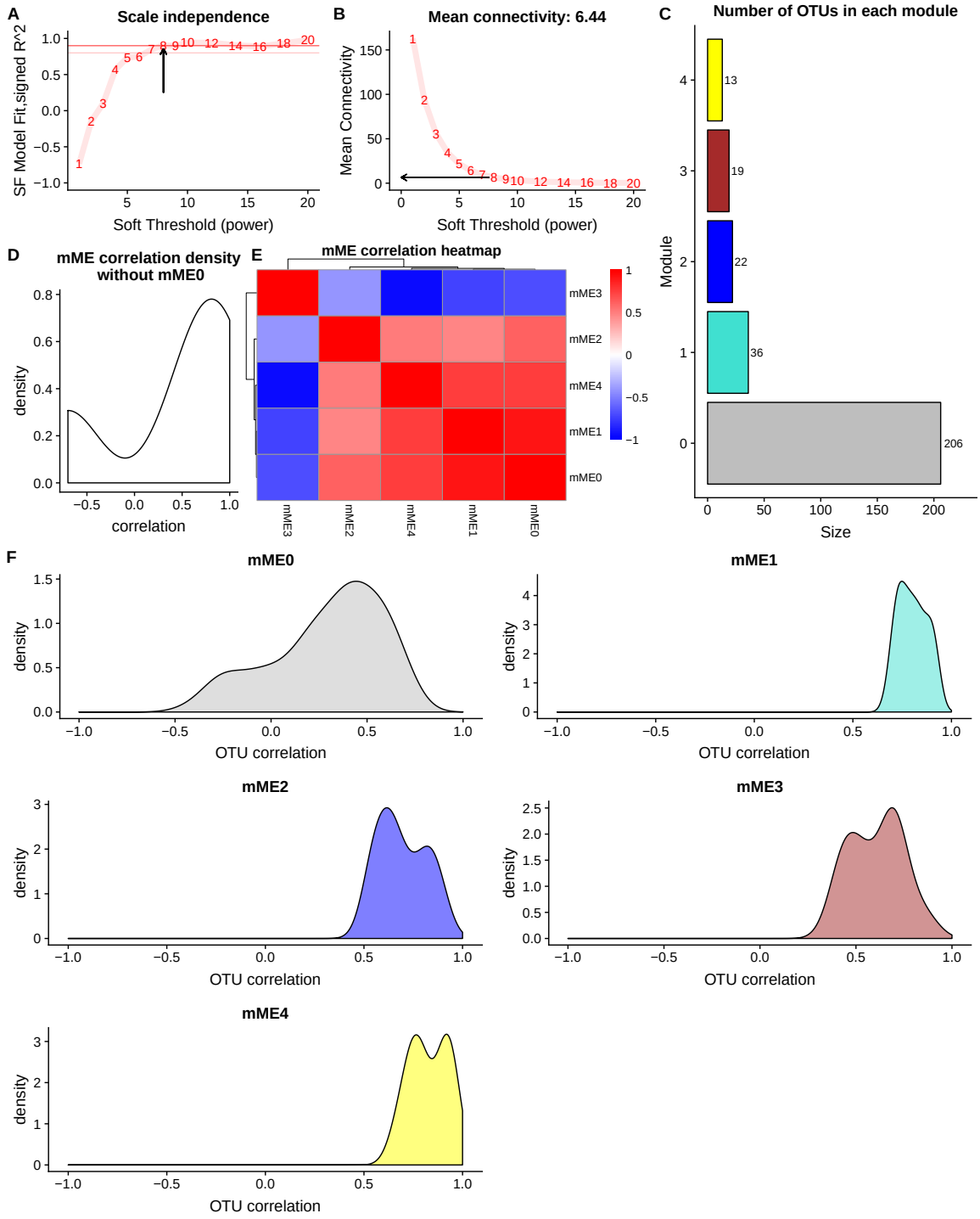

Figure 4.2: **Diagnostic plots for gut microbiome co-abundance network – all data. A)** Scale independence. Shows how the degree distribution of networks at different betas fits a scale-free distribution. Arrow shows the beta value chosen for the network. **B)** The mean connectivity of the network. Should monotonically decrease and should not be too low. Mean connectivity as chosen beta value is marked with an arrow, and the numerical value of the means is shown in the title of the plot. **C)** Shows the size of modules i.e. how many OTUs each module contains. **D)** Density plot of the correlation between microbial module eigenOTUs (mMEs), peak of the density should be at zero. **E)** mME correlation heatmap. A strong positive correlation is deep red, and a strong negative correlation is deep blue. A correlation of 0 is white. **F)** Correlations of OTUs within each module to their respective mME. Because the network is signed all OTUs within a module should have strong positive correlations i.e. have a peak above 0 and preferably > 0.5. The unassigned OTUs in grey should have an average correlation of zero.

4.2.2 Hub nodes

A Hub node is the most connected node in a module. Table 4.1 lists hub genes for all host gene expression modules. Table 4.2 lists hub OTUs for all microbial abundance modules.

Table 4.1: Hubgenes, the most connected gene in each host gene expression module.

Module	Gene id	Gene name	Gene product
1	gene31966:106570826	LOC106570826	MAM domain-containing protein 2-like
2	gene56602:100380491	clrn3	Clarin-3
3	gene12190:106605532	LOC106605532	Protein LYRIC-like
4	gene21074:106560259	LOC106560259	Uncharacterized LOC106560259
5	gene32398:100194994	rir2	Ribonucleoside-diphosphate reductase ...
6	gene4299:106583279	smpdl3b	Sphingomyelin phosphodiesterase ...
7	gene16457:106609569	LOC106609569	Protein BTG1-like
8	gene4583:106585643	LOC106585643	Beta-2-glycoprotein 1-like
9	gene23589:106562741	LOC106562741	Golgin subfamily B member 1-like
10	gene42553:106581405	LOC106581405	arf-GAP with Rho-GAP domain ...
11	gene2952:100196361	rpl12	Ribosomal protein L12
12	gene68501:106590896	LOC106590896	Uncharacterized LOC106590896
13	gene14495:106607713	mdn1	Midasin AAA ATPase 1
14	gene38234:106577118	LOC106577118	Ubiquitin-40S ribosomal protein S27a
15	gene14018:106607375	LOC106607375	Hemoglobin subunit alpha-4
16	gene31978:106570885	LOC106570885	Perforin-1-like

Table 4.2: Hub-OTUs, the most connected OTUs in each microbial abundance module.

Module	OTU	Phylum	Genus
1	OTU_10	Firmicutes	Enterococcaceae
2	OTU_20	Proteobacteria	Vibrionaceae
3	OTU_207	Proteobacteria	Litoricolaceae
4	OTU_30	Proteobacteria	Halomonadaceae

4.2.3 Correlations between host and microbial network modules and to traits

Following network construction, network modules were correlated between data types to identify putative interactions between microbe abundances and host genes. Figure 4.3 shows both correlations between expression eigengenes and experimental variables and correlations between host expression eigengenes and microbial eigenOTUs. As expected, the correlation between host modules and experimental variables are almost exclusively driven by the freshwater-saltwater transition (arrow D). Host module eigengene 2 (hME2) has the strongest positive correlation to the transition (arrow A), while hME1 has the strongest negative correlation to the transition (arrow B). Furthermore, the correlations between microbial modules and host modules show that the same host modules with associations to experiment also has significant associations to microbes (arrow C); All gut microbiota abundance module eigenOTUs (mMEs) respond almost exclusively to the freshwater-saltwater

transition. mME4 has the strongest positive correlation, as can be seen at the intersection of arrow A and arrow C. mME3, left of mME4, has the opposite response. From these two observations, we can indirectly infer that OTUs in module 4 (which correlate strongly to mME4) increase in saltwater and that OTUs in module 3 decrease in saltwater. The variable Day is difficult to separate from the freshwater-saltwater transition. As seen in Figure 4.3 Day responds almost exactly like WaterSW. There is no significant correlation between Feed and hMEs in Figure 4.3. The effect of the fresh-saltwater transition overwhelms/overshadows subtler signals related to feed.

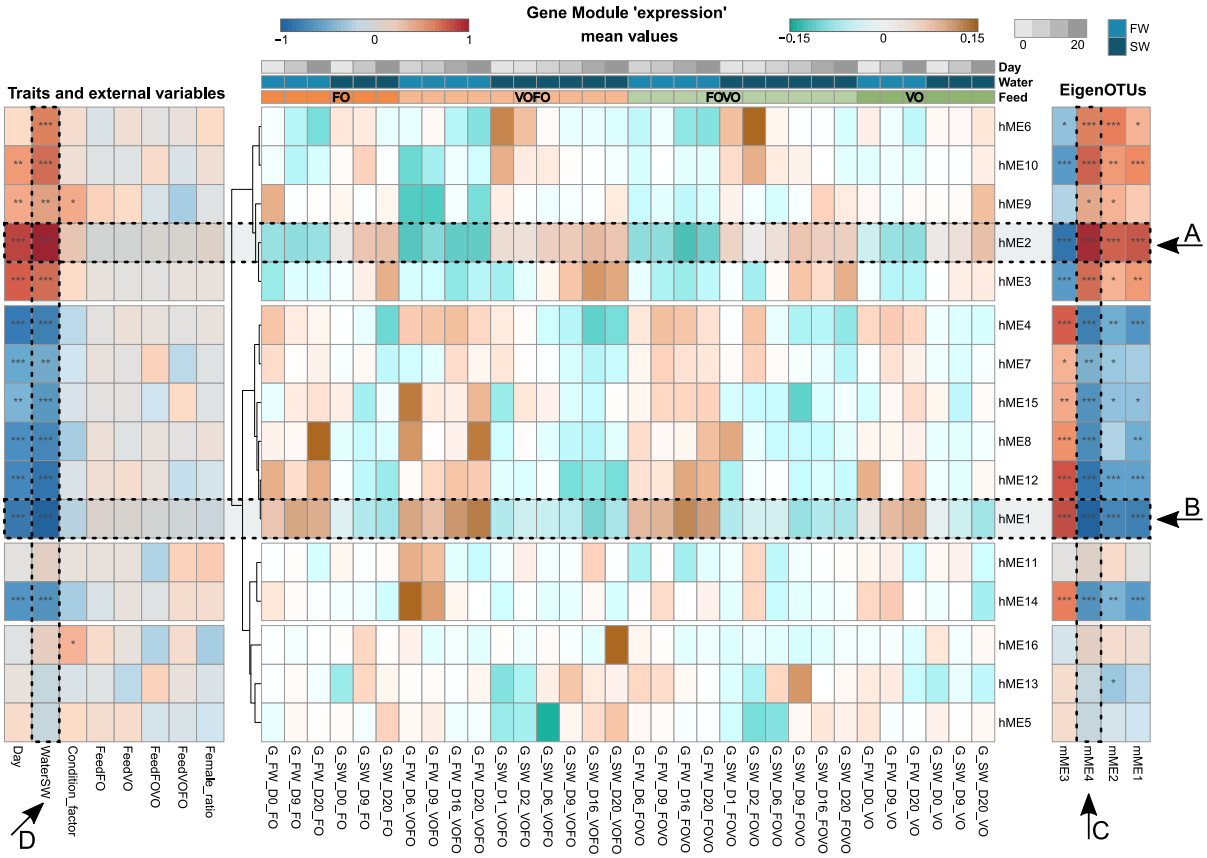

Figure 4.3: A split heatmap that shows the correlation of the host gene expression module eigengenes (hME) with the microbiome abundance module eigenOTUs (mME) and non-omics data such as traits and external variables. The central heatmap shows the mean 'expression' for each hME, each column is a sample group, organized by feed then water then day. The figure should be read horizontally, with each row constituting a new hME, e.g., hME2 is marked with arrow A, and hME1 is marked with arrow B. The leftmost heatmap shows the correlation strength between hMEs and non-omics traits/external variables. The rightmost heatmap shows the correlation strength between each hME and each mME. Every correlation is based on the mean of each sample group i.e. the mean of biological replicates. A strong positive correlation is deep red, and a strong negative correlation is deep blue. A correlation of 0 is white. In each of the correlation-heatmaps the p-value of the correlation is marked with increasing number of stars as it reaches certain thresholds: * $p <= 0.05$, ** $p <= 0.01$, *** $p <= 0.001$. p-values are not corrected for multiple comparisons and should the interpreted as confidence strengths rather than rigorous hypotheses tests.

4.2.4 Selected GO enrichment

Gene ontology (GO) enrichment analysis was done on selected host expression modules to explore the contents of these modules further. Over-representation of a group of genes gives us an indication of what the specific module does. For the GO enrichment, genes were selected based on their correlation to specific module eigengenes. This correlation also ranks the centrality of the genes to the eigengenes' module. The maximum and median correlation in each GO-term is included in the results table.

Table 4.3 shows that host gene expression module 1 is enriched for genes involved in the extracellular matrix organization and disassembly, breaking down of collagen, building up of blood vessels (angiogenesis), endodermal cell differentiation and somatic stem cell division. As can be seen in the max correlation column, several terms (1-4) have the same maximum correlation. This is because they all contain the same gene. GO enrichment works with terms that are subgroups of other terms, and therefore genes can and often will show up in different significant terms.

Table 4.3: **GO enrichment of genes that correlate more than 0.8 to hME1.** Table shows the top 20 most significantly enriched GO terms in descending order of significance starting with the most significant. Observed = the number of genes that are annotated with the given GO term that also correlate ≥ 0.8 to hME1. Expected = the expected (i.e. expected by chance) number of genes that are annotated with the given GO term that also correlate ≥ 0.8 to hME1. P-value is calculated with the weight01 algorithm and fisher statistic. Median corr. is the median correlation to the respective eigengene for genes that correlate ≥ 0.8 to hME1. Max corr. is likewise based on eigengene correlation but shows the maximum correlation of any gene in the GO term i.e. the strongest correlation of any one gene. The highest maximum and median correlation is marked in red.

	GO.ID	Term	Observed	Expected	p-value	Median corr.	Max corr.
1	GO:0030198	Extracellular matrix organization	138	28.38	2.2e-28	0.867	0.962
2	GO:0022617	Extracellular matrix disassembly	47	7.32	6.5e-26	0.89	0.962
3	GO:0030574	Collagen catabolic process	28	3.7	8.5e-16	0.901	0.962
4	GO:0001525	Angiogenesis	109	38.99	1.7e-12	0.845	0.962
5	GO:0030199	Collagen fibril organization	22	3.81	3.5e-11	0.896	0.952
6	GO:0001657	Ureteric bud development	34	8.37	4.3e-11	0.834	0.921
7	GO:0010812	Negative regulation of cell-substrate ad...	30	6.14	4.9e-11	0.858	0.952
8	GO:0001501	Skeletal system development	127	45.81	1.2e-10	0.852	0.952
9	GO:0008285	Negative regulation of cell proliferatio...	120	58.18	2.5e-10	0.844	0.953
10	GO:0007155	Cell adhesion	215	91.39	2.7e-10	0.853	0.96
11	GO:0035987	Endodermal cell differentiation	27	5.28	9.3e-10	0.865	0.962
12	GO:0048103	Somatic stem cell division	18	4.23	1.0e-09	0.841	0.943
13	GO:0071230	Cellular response to amino acid stimulus	21	4.12	1.1e-09	0.89	0.96
14	GO:0021915	Neural tube development	46	18.46	1.2e-09	0.844	0.943
15	GO:0008360	Regulation of cell shape	42	15.36	2.1e-09	0.858	0.915
16	GO:0070208	Protein heterotrimerization	12	1.23	2.2e-09	0.892	0.952
17	GO:0060346	Bone trabecula formation	11	1.02	3.2e-09	0.892	0.952
18	GO:0042476	Odontogenesis	44	11.84	5.1e-09	0.846	0.936
19	GO:0007411	Axon guidance	116	53.74	5.3e-09	0.854	0.962
20	GO:0031960	Response to corticosteroid	31	16.01	1.2e-08	0.851	0.935

Table 4.4 shows that host gene expression module 2 is enriched for genes involved in intracellular protein transport. While this general term is the most significant, the term with the most highly connected gene is phosphatidylethanolamine biosynthesis – i.e. the production of a class of phospholipids. The enrichment analysis also identifies many associated with cholesterol transport and homeostasis.

Table 4.4: **GO enrichment of genes that correlate more than 0.8 to hME2.** Table shows the top 20 most significantly enriched GO terms in descending order of significance starting with the most significant. Observed = the number of genes that are annotated with the given GO term that also correlate ≥ 0.8 to hME2. Expected = the expected (i.e. expected by chance) number of genes that are annotated with the given GO term that also correlate ≥ 0.8 to hME2. P-value is calculated with the weight01 algorithm and fisher statistic. Median corr. is the median correlation to the respective eigengene for genes that correlate ≥ 0.8 to hME2. Max corr. is likewise based on eigengene correlation but shows the maximum correlation of any gene in the GO term i.e. the strongest correlation of any one gene. The highest maximum and median correlation is marked in red.

	GO.ID	Term	Observed	Expected	p-value	Median corr.	Max corr.
1	GO:0006886	Intracellular protein transport	113	68.07	4.3e-21	0.864	0.958
2	GO:0006646	Phosphatidylethanolamine biosynthetic pr...	14	0.95	1.8e-13	0.923	0.971
3	GO:0006890	Retrograde vesicle-mediated transport, G...	19	2.22	6.2e-13	0.840	0.895
4	GO:0006888	ER to Golgi vesicle-mediated transport	46	6.89	5.7e-11	0.855	0.958
5	GO:0006891	Intra-Golgi vesicle-mediated transport	16	1.90	8.1e-11	0.834	0.895
6	GO:0018279	Protein N-linked glycosylation via aspar...	29	6.92	9.5e-11	0.863	0.958
7	GO:0019432	Triglyceride biosynthetic process	22	4.55	9.5e-11	0.879	0.969
8	GO:0032482	Rab protein signal transduction	23	4.50	1.5e-10	0.872	0.954
9	GO:0048205	COPI coating of Golgi vesicle	12	1.00	1.6e-10	0.837	0.895
10	GO:0065005	Protein-lipid complex assembly	18	1.65	2.2e-10	0.885	0.943
11	GO:0032374	Regulation of cholesterol transport	17	2.67	3.1e-10	0.898	0.943
12	GO:0042632	Cholesterol homeostasis	22	4.52	1.1e-09	0.888	0.956
13	GO:0016042	Lipid catabolic process	45	15.61	1.5e-09	0.877	0.956
14	GO:0006654	Phosphatidic acid biosynthetic process	13	1.50	2.3e-09	0.896	0.956
15	GO:0048208	COPII vesicle coating	11	1.02	3.5e-09	0.864	0.958
16	GO:0007030	Golgi organization	31	9.39	1.1e-08	0.858	0.948
17	GO:0036150	Phosphatidylserine acyl-chain remodeling	8	0.55	3.4e-08	0.933	0.956
18	GO:0010873	Positive regulation of cholesterol ester...	6	0.25	4.6e-08	0.900	0.935
19	GO:0006657	CDP-choline pathway	8	0.60	7.5e-08	0.924	0.943
20	GO:0046470	Phosphatidylcholine metabolic process	33	4.30	1.5e-07	0.922	0.971

4.2.5 Microbial relative abundance heatmap

To further elucidate the underlying patterns within each microbe-network module, we performed an exploratory analysis of the relative abundance data by plotting the data directly as a heatmap with the TOM dendrogram (Figure 4.4). Previous work using this exact metagenomics (16S) dataset (and OTU clustering) found four OTUs more likely to be involved in host-microbe interactions based on their persistence across the freshwater-saltwater transition (Rudi et al. 2018) – See table 4.5. Information about these OTUs in the context of co-abundance modules was therefore included in the microbial network analysis. In this analysis, only the most abundant OTUs, those that contributed more than 0.005% to the total abundance, were included. In the heatmap (Figure 4.4), abundance values are relative to the total sample wise abundance of these included 296 OTUs.

Although most OTUs in module 1 (Turquoise) and some OTUs in module 2 (Blue) can be seen to persist across the entire experiment, the phylum of most central OTUs of each module differ (Arrow A and B). Module 1 contains a majority of Firmicutes (Arrow A), while module 2 contains both, Firmicutes and Proteobacteria (Arrow B). The core OTUs found by Rudi et al. (2018) (highlighted in red) are all clustered together and at the center of module 1 (Turquoise) (Arrow A). Just as an aside, a dendrogram sometimes does a terrible

job at showing which OTUs are the most connected. Note that the ordering of a dendrogram can be changed within the constraints set by the dendrogram.

Furthermore, visual inspection of the OTU heatmap demonstrates that more subtle patterns are picked up in the network analyses. In Figure 4.4(Arrows marked C) we can observe the underlying abundance shifts that result in modules 3 (Brown) and 4 (Yellow) being oppositely correlated with transition from fresh- to saltwater in Figure 4.3. Some samples, or sample groups in this case, behave strangely see arrows marked D. Especially the leftmost arrow which is a VOFO transition, saltwater sample group. Here OTUs that are relatively rare in other sample groups have a relatively large relative abundance.

Although it does not belong to any module, one bacterium, genus Corynebacterium (OTU_8, arrow E), appears highly prominent in late stage freshwater samples but is almost gone from saltwater samples. The reason for its lack of a module is related to this unique behavior, which it seems only this bacterium has. As previously seen in Figure 4.3, there is no clear effect of diet on any bacteria. This is also clear from the heatmap in Figure 4.4.

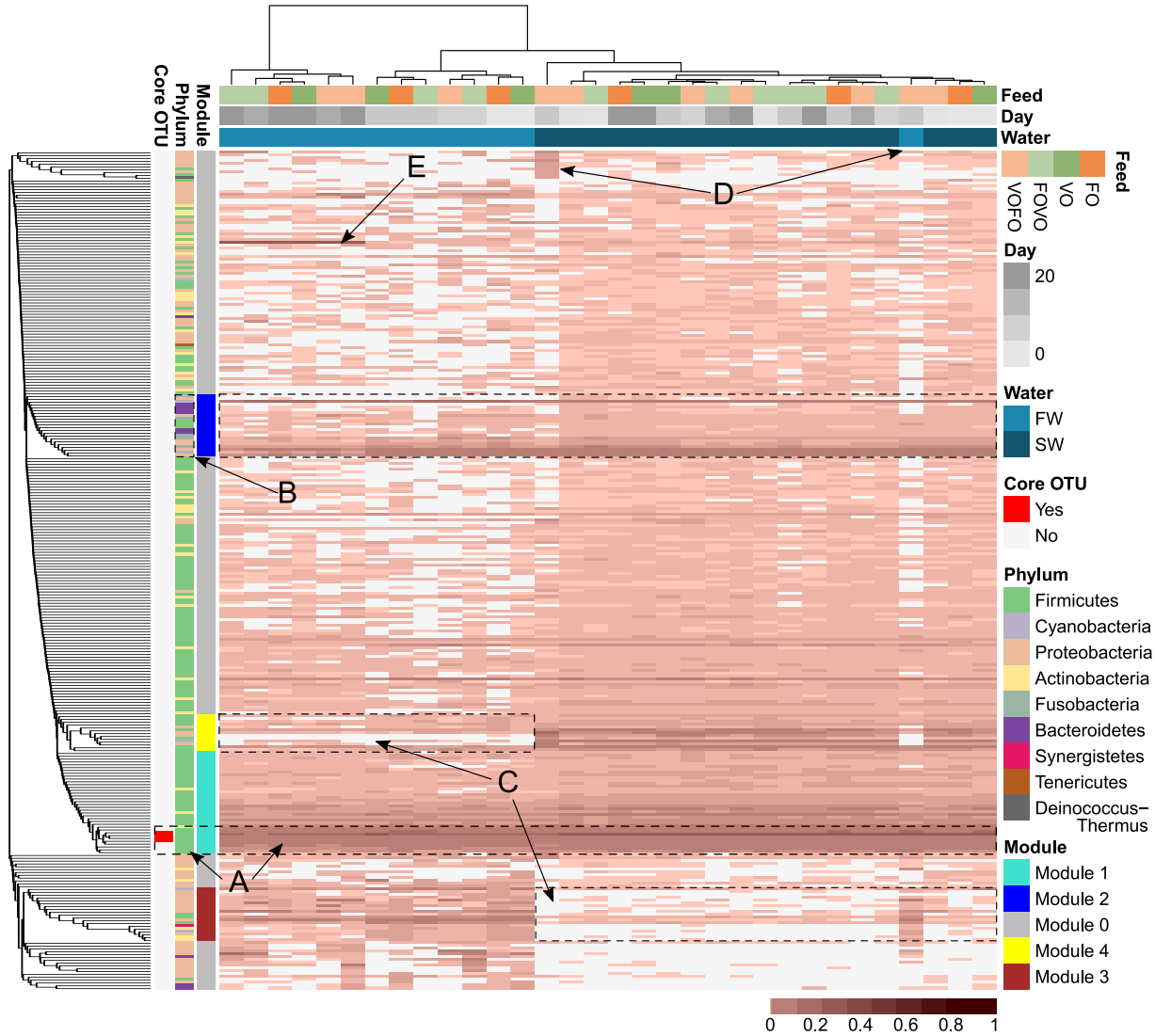

Figure 4.4: **A heatmap of the relative abundances – All data.** OTUs are rows and samples are columns (mean values of replicates). The color scheme has been adjusted so that zero abundance is light grey, while any abundance >0 is colored with interval which emphasises the low end of the abundance. There are three row-annotation-bars: The first shows in red the four core OTUs found in Rudi et al. (2018), the second shows the determined phyla of the OTU, the third shows module membership. The three column-annotation-bars are identical to column-bars in figure 4.3, but organized with ward's method.

4.2.6 Taxonomy for selected microbial abundance modules

From the work done in Rudi et al. (2018) we know of four interaction candidate OTUs. Their taxonomy can be seen in table 4.5. Tables 4.6 and 4.7 lists all OTUs in microbial abundance module 1 and 2 respectfully. As can be seen by the ordering of the tables, and the location of the (m); Here both the most connected OTU and the OTU that correlates strongest to the eigenOTU is the same.

Table 4.5: The four putative core bacteria. (OTU 5 here is OTU 6 in Rudi et al. (2018))

	Phylum	Class	Order	Family	Genus
OTU_1	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
OTU_2	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
OTU_5	Firmicutes	Clostridia	Clostridiales		Peptoniphilus
OTU_10	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Vagococcus

Table 4.6: OTUs in microbial abundance module 1. (hub) means that the OTU is the hub OTU having the highest connection strength in the module. All OTUs are ordered after correlation to the module eigenOTU. Correlation means the correlation of the OTU to the module eigenOTU.

	OTU	Correlation	Phylum	Class	Order	Family	Genus
1	OTU_10 (hub)	0.958	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Vagococcus
2	OTU_2	0.935	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
3	OTU_1	0.933	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
4	OTU_14	0.933	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
5	OTU_11	0.925	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
6	OTU_5	0.908	Firmicutes	Clostridia	Clostridiales		Peptoniphilus
7	OTU_13	0.905	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
8	OTU_15	0.872	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Denitrobacterium
9	OTU_19	0.863	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Carnobacterium
10	OTU_24	0.852	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
11	OTU_21	0.845	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Robinsoniella
12	OTU_6	0.830	Firmicutes	Clostridia	Clostridiales		Gallicola
13	OTU_35	0.829	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Atopobium
14	OTU_71	0.823	Firmicutes	Clostridia	Clostridiales		Mogibacterium
15	OTU_22	0.810	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium
16	OTU_27	0.799	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella
17	OTU_429	0.794	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Vagococcus
18	OTU_65	0.787	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium
19	OTU_1045	0.780	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Carnobacterium
20	OTU_1684	0.773	Firmicutes	Clostridia	Clostridiales		Peptoniphilus
21	OTU_77	0.768	Firmicutes	Bacilli	Bacillales	Bacillaceae	Lysinibacillus
22	OTU_46	0.761	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Trichococcus
23	OTU_45	0.755	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
24	OTU_115	0.740	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
25	OTU_674	0.737	Firmicutes	Clostridia	Clostridiales		Anaerovorax
26	OTU_76	0.736	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	Saccharopolyspora
27	OTU_85	0.735	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella
28	OTU_64	0.725	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	Peptococcus
29	OTU_62	0.714	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
30	OTU_108	0.714	Firmicutes	Clostridia	Clostridiales		Anaerovorax
31	OTU_91	0.695	Firmicutes	Bacilli	Bacillales	Planococcaceae	Kurthia
32	OTU_63	0.677	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella
33	OTU_116	0.664	Firmicutes	Bacilli	Bacillales	Planococcaceae	Sporosarcina
34	OTU_40	0.651	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
35	OTU_102	0.644	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Pseudoramibacter
36	OTU_61	0.632	Firmicutes	Clostridia	Clostridiales		Anaerovorax

4.2.7 Graph model of microbiome network

Both dendrograms and tables do not represent the relationships between OTUs in a very spatial way. Figure 4.5 is a more traditional representation of the topological overlap mea-

Table 4.7: OTUs in microbial abundance module 2. (hub) means that the OTU is the hub OTU having the highest connection strength in the module. All OTUs are ordered after correlation to the module eigenOTU. Correlation means the correlation of the OTU to the module eigenOTU.

	OTU	Correlation	Phylum	Class	Order	Family	Genus
1	OTU_20 (hub)	0.893	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Photobacterium
2	OTU_7	0.874	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Tropicibacter
3	OTU_4	0.84	Cyanobacteria		Oscillatoriales		Planktothrix
4	OTU_82	0.788	Proteobacteria	Gammaproteobacteria	Alteromonadales	Moritellaceae	Moritella
5	OTU_141	0.785	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
6	OTU_37	0.77	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Psychrilyobacter
7	OTU_104	0.76	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
8	OTU_1386	0.752	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Photobacterium
9	OTU_25	0.747	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Cetobacterium
10	OTU_154	0.712	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
11	OTU_94	0.688	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
12	OTU_132	0.662	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Bulleidia
13	OTU_214	0.646	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
14	OTU_169	0.643	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
15	OTU_136	0.638	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Oribacterium
16	OTU_461	0.616	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Myroides
17	OTU_98	0.611	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas
18	OTU_498	0.583	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	Sphingobacterium
19	OTU_384	0.581	Cyanobacteria		Oscillatoriales		Planktothrix
20	OTU_641	0.576	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
21	OTU_170	0.56	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	Phascolarctobacterium
22	OTU_3	0.517	Cyanobacteria		Oscillatoriales		Planktothrix

sure that the dendrogram ordering the relative abundances in figure 4.4 was based on. Because this is a weighted network, the lines between OTUs represent the strength of the topological overlap measure (TOM),

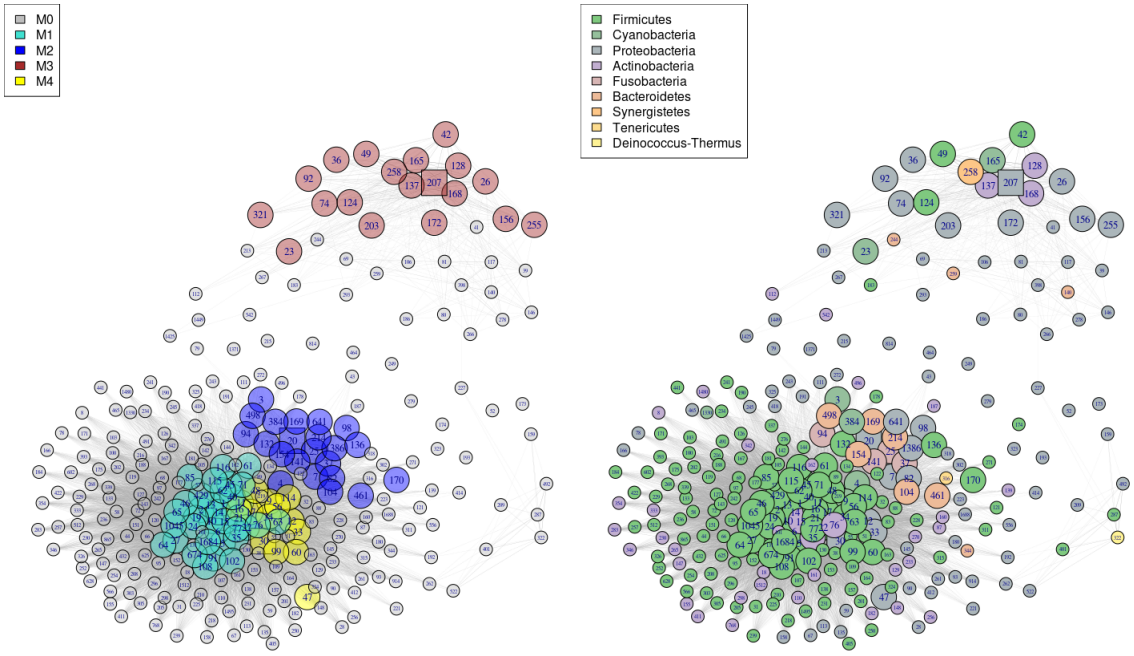

Figure 4.5: Two representations of the same OTU network. The left most graph shows module membership, the rightmost graph shows taxonomic classification of the phylum level. Each circle represents an OTU, a larger circle means it is part of a module. Squares are hub OTUs. Lines between OTUs are drawn based on the strength of the connection (TOM) – The more pronounced the line the stronger the connection. Layout with the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold 1991), implemented in the R-package igraph (Csárdi and Nepusz n.d.).

4.3 Comparison to principal component analysis (PCA)

WGCNA uses singular value decomposition to find module eigengenes (MEs) (or eigenOTUs) for each module. A ME is the first principal component (1st PC) of the expression data for each module. In a PCA analysis a common approach to summarize the information content of PCs is to calculate the proportion of variance that each PC captures. This naturally leads to the question of how well the MEs explain the variance within a module and how this compares to a 'whole data' PCA. Figure 4.6 and Figure 4.7 shows how a principal component analysis relates to the MEs produced by WGCNA. Figure 4.6A shows that PC1 of the 16S rRNA gene amplicon relative abundance PCA explains 27.6% while PC2 explains 6.9%. PC1 and PC2 together separate fresh and saltwater samples. Saltwater samples are clustered more tightly than freshwater samples. Figure 4.6B shows that PC1 of the host gene expression PCA explains 26.1% of the variation in the host salmon gene expression. PC2 explains 7.1%. Also, here, PC1 and PC2 explains enough variation to separate fresh from saltwater samples. We observed the concentration of FO and later stage VOFO feed samples in the bottom right corner of the biplot. It also shows how the feed switches have started to look like their unchanged counterparts i.e. VOFO starts to look like FO. Figure 4.6C shows that the module eigengenes are different to the whole data PCs. For one, the module eigengenes can have similar expression profiles even when they are based on different sets of genes. Most of the strong correlations are present in the first 3 PCs. The genes and OTUs within their respective modules can be ranked within the module similarly to the loading's in a PCA. This centrality measure carries with it the interpretation that the nodes have a central placement in the module and is important for how that module behaves expression-wise. In Figure 4.6C we see that the first "whole data" principal component (PC) corresponds as expected to all the module eigenOTUs ('OTUs section'), the strongest correlation, however, is with mME0 which is eigenOTUs of the unassigned 'module'. In the 'Host genes: PCA vs modules' section, we see that hME1 and hME2 corresponds strongest to the first "whole data" PC. hME6 and hME10 to PC2, and the unassigned 'module' hME0 has a strong positive correlation to PC3.

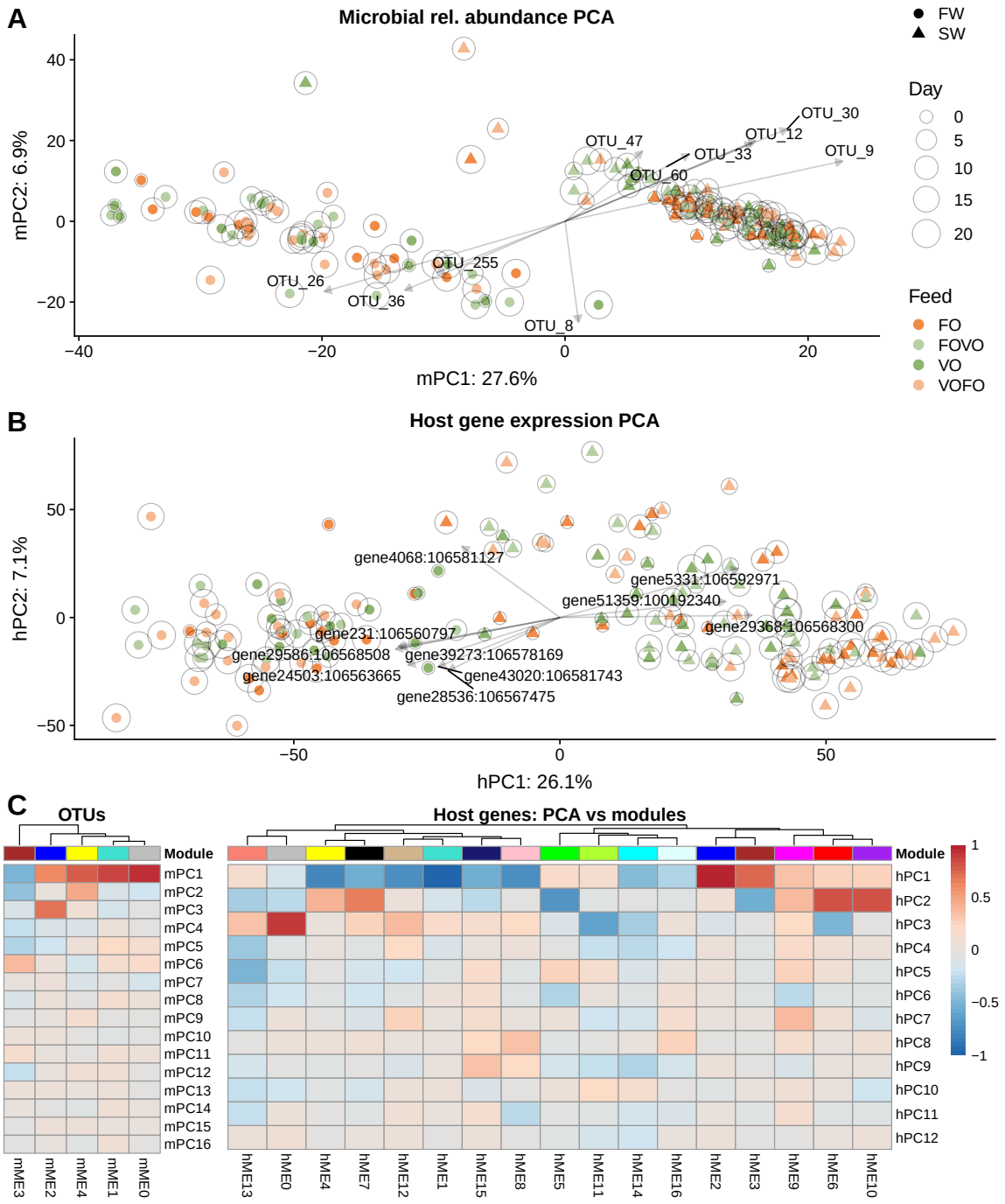

Figure 4.6: Comparison of PCA and Module eigenvectors. **A** and **B** shows biplots. X-axis is PC1 and y-axis is PC2. Points are the response of samples to the new PC1 and PC2 coordinate system, were the shapes signify whether the sample is from fresh or saltwater. Arrows are loadings i.e. how the original variables influence PCs, only the ten most influential variables are chosen based on the Euclidean distance they span between the two axes. **C** shows the correlation between "all data" PCs and the Modules eigengenes/eigenOTUs we found earlier. Deep blue indicates a strong negative correlation, while deep red indicates a strong positive correlation.

Figure 4.7 shows that most of the variance is captured by PC1, but also that dividing the data into subsets of genes and subsets of OTUs found by network module analysis of each dataset increases the explained variance for PC1. As can be seen in Figure 4.7A and 4.7B, both eigenOTUs and eigengenes, respectively, are better representatives of the variance in these subsets than PC1 is for the whole data (all OTUs, all genes). The figure also shows that the other PCs contain very little information compared to the first PC.

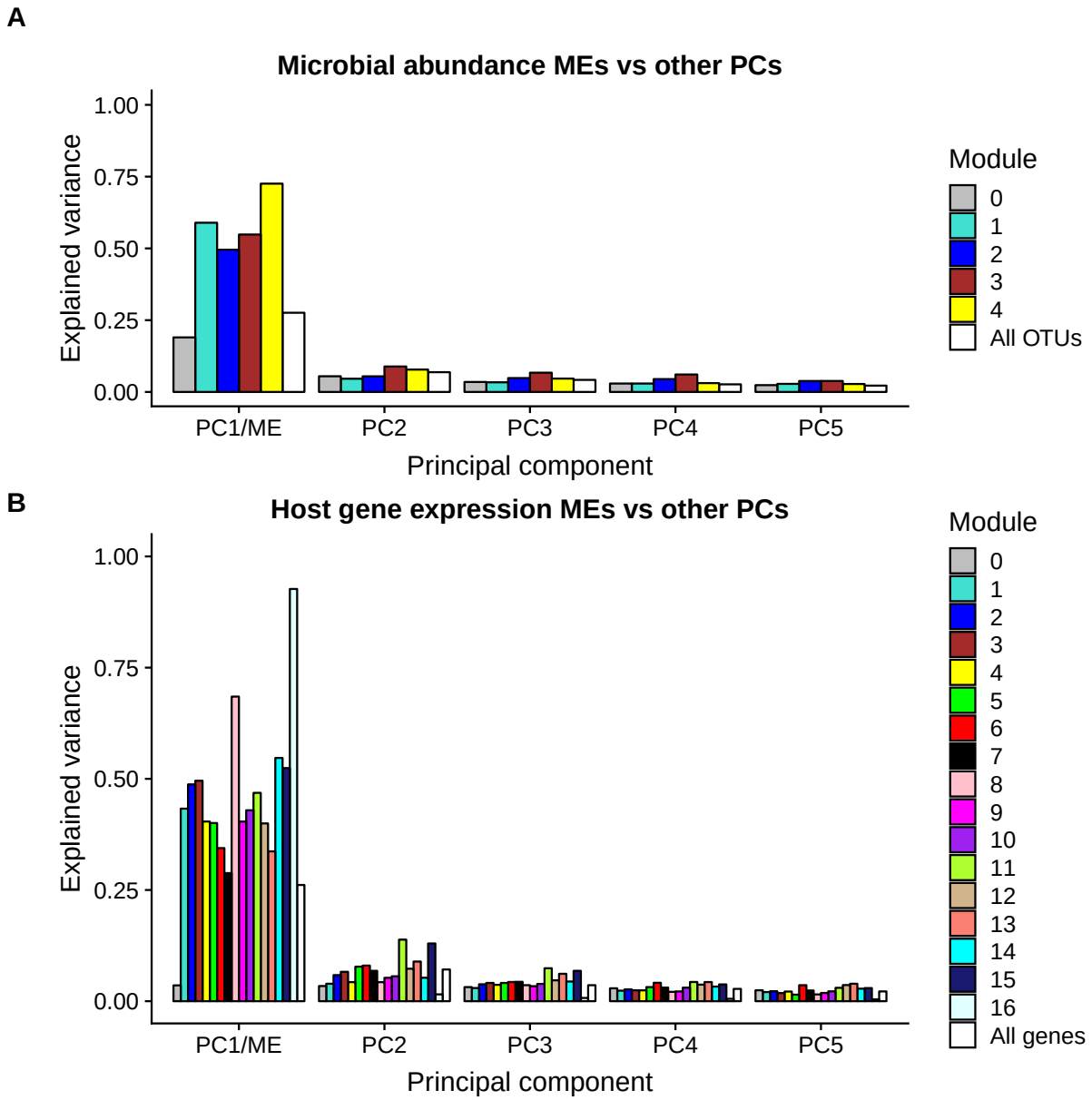

Figure 4.7: PCA of OTU subsets in A, and genes subsets in B. The first principal component (PC1) corresponds to the module eigengene (ME) hence the PC1/ME label. Y-axis of both figures show the proportion of explained variance. For all modules, and for the whole data, principal components 1-5 are shown. In WGCNA only the first PC, the eigengene, is used as a representative, all other PCs are disregarded.

4.4 All data with the removal of large effect variables

4.4.1 Network characteristics

As expected, using all data with the network correlation approach simply resulted in the detection of the large and biologically un-surprising effects of transition between fresh- and saltwater. Hence, we next used a method to remove the most extreme variance structures in the data, to potentially reveal more subtle correlations between host gene expression and microbial community shifts. And, potentially allowing us to identify associations between fish that gets diets with slightly different lipid content (FO vs VO) or fish that had been switching diets (VO->FO, FO->VO). The results from the network constructions can be seen in Figure 4.7 and Figure 4.8.

The PC-corrected host gene expression network has approximately scale free topology

(>0.9) at beta = 18 (Figure 4.7A), which was drastically different from the uncorrected data at beta = 3 (Figure 4.1). Interestingly, for the gene expression data, the removal of the major principal components (16 PCs) reduced the module sizes dramatically from 49-9121 genes down to 30-206 genes. Most genes are now in the 'unassigned' module (36805 genes). However, this is not surprising since we already showed that most genes change expression between salt and fresh water, and removing this effect thus leaves most genes with an uninformative expression profile across these samples. Unsurprisingly the mean connectivity is very low (0.52), but the connectivity within each module is high.

For the microbial abundance data, the removal of major principal components (5 PCs) unexpectedly did not change the number of modules the same way it did for the gene expression data. For the microbial abundance data, the number of modules instead increased from 4 to 7. Module sizes now range from 5 to 22 OTUs (Figure 4.8C) compared to 13 to 36 OTUs (Figure 4.2C). In the initial analyses (Figure 4.2D), most MEs were extremely highly correlated. The application of the PC-correction method removed the high intramodular correlation improving the modules which now represent more distinct abundance-behavior (Figure 4.8D). Correlation to mME0 has an expected peak at zero, which is what you would want from the mME0, while all genes in other modules have a positive correlation to their respective mMEs. The correlation is however lower than it was in the uncorrected data, averaging around 0.5.

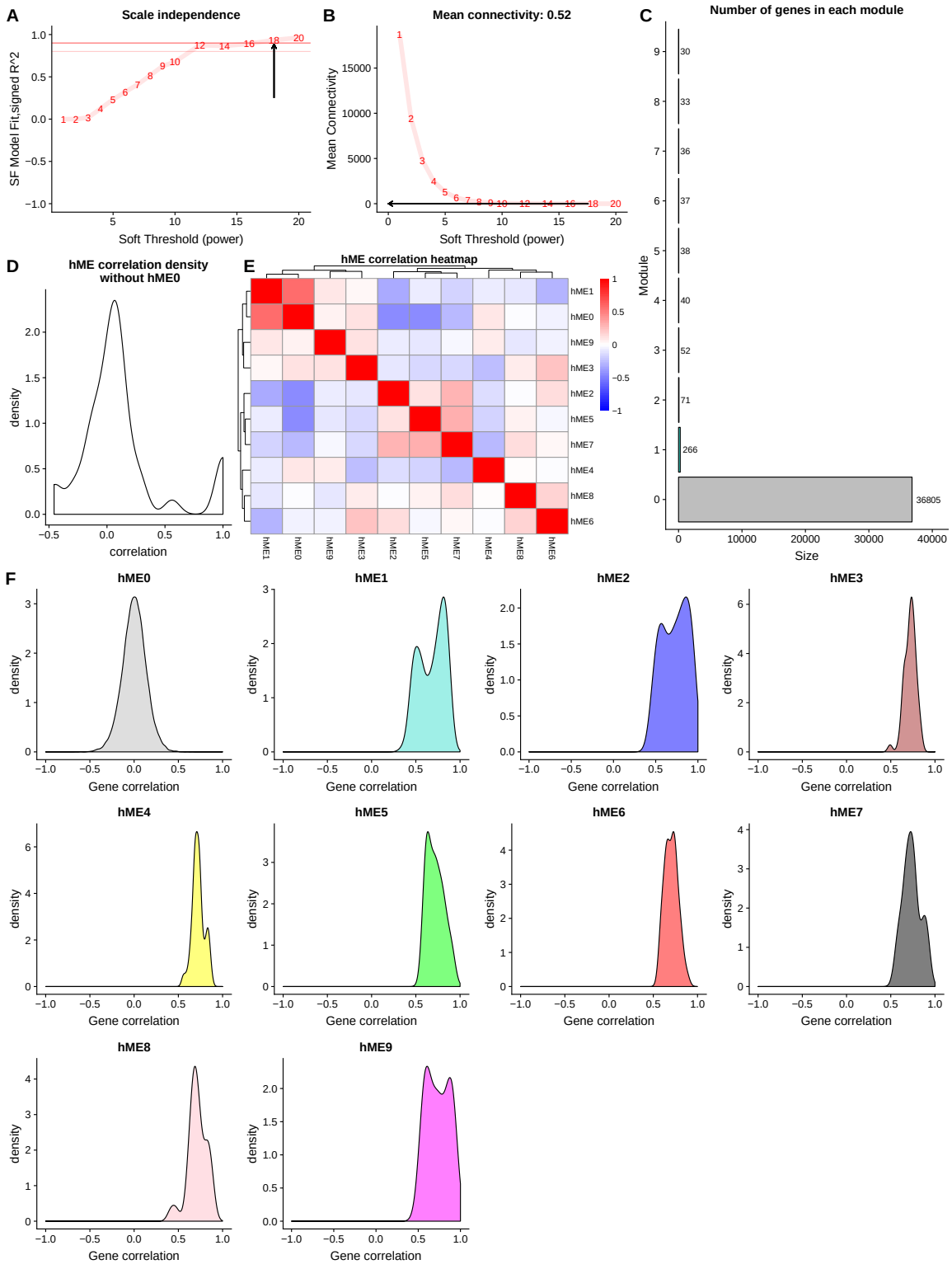

Figure 4.8: **Diagnostic plots for the host gene co-expression network – PC-correction. A)** Scale independence. Shows how the degree distribution of networks at different betas fits a scale-free distribution. Arrow shows the beta value chosen for the network. **B)** The mean connectivity of the network. Should monotonically decrease and should not be too low. Mean connectivity as chosen beta value is marked with an arrow, and the numerical value of the means is shown in the title of the plot. **C)** Shows the size of modules i.e. how many genes each module contains. **D)** Density plot of the correlation between host module eigengenes (hMEs), peak of the density should be at zero. **E)** hME correlation heatmap. A strong positive correlation is deep red, and a strong negative correlation is deep blue. A correlation of 0 is white. **F)** Correlations of genes within each module to their respective hME. Because the network is signed all genes within a module should have strong positive correlations i.e. have a peak above 0 and preferably above 0.5. The unassigned genes in grey should have an average correlation of zero.

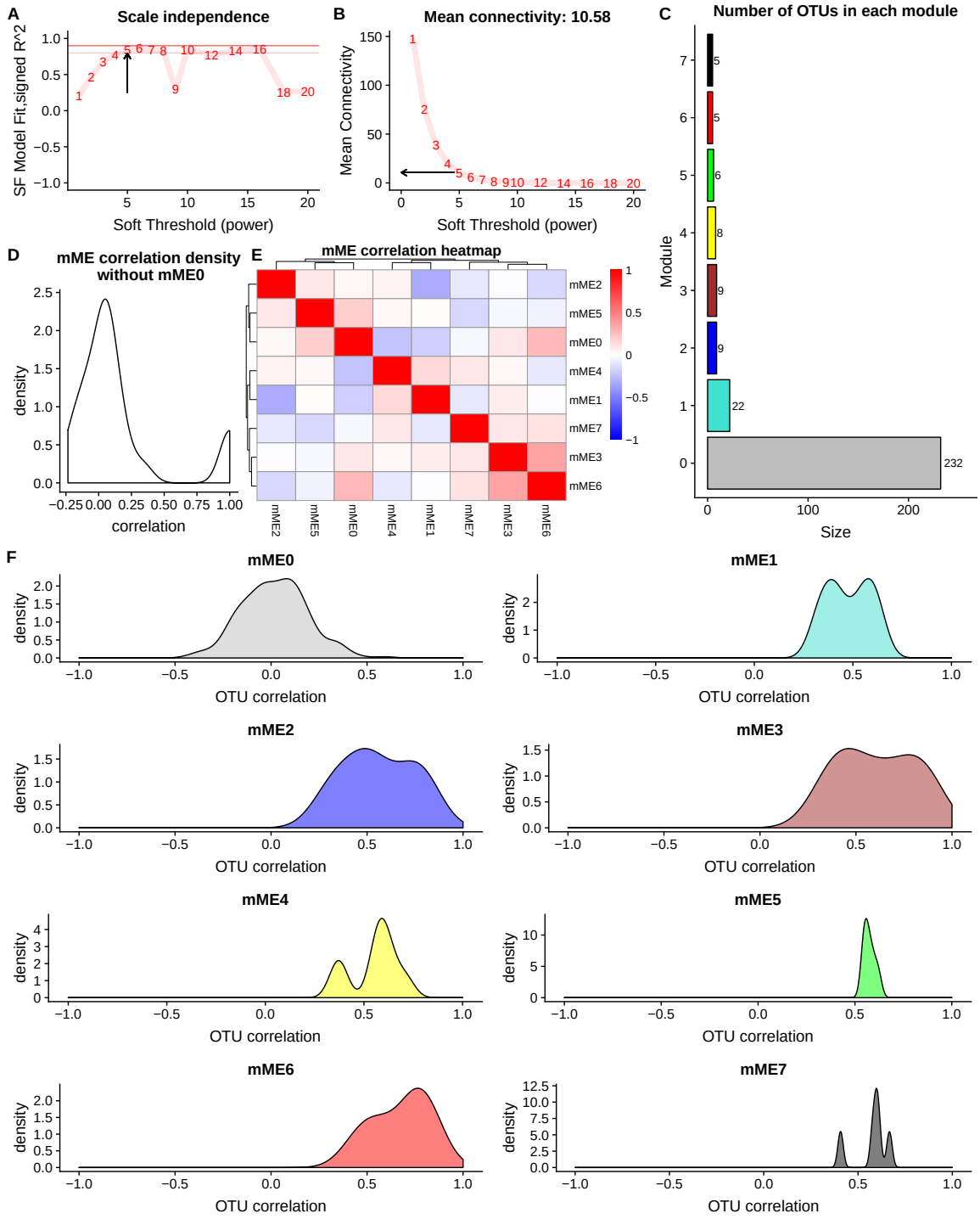

Figure 4.9: **Diagnostic plots for the microbial co-abundance network – PC-correction. A)** Scale independence. Shows how the degree distribution of networks at different betas fits a scale-free distribution. Arrow shows the beta value chosen for the network. **B)** The mean connectivity of the network. Should monotonically decrease and should not be too low. Mean connectivity as chosen beta value is marked with an arrow, and the numerical value of the means is shown in the title of the plot. **C)** Shows the size of modules i.e. how many OTUs each module contains. **D)** Density plot of the correlation between microbial module eigenOTUs (mMEs), peak of the density should be at zero. **E)** mME correlation heatmap. A strong positive correlation is deep red, and a strong negative correlation is deep blue. A correlation of 0 is white. **F)** Correlations of OTUs within each module to their respective mME. Because the network is signed all OTUs within a module should have strong positive correlations i.e. have a peak above 0 and preferably above 0.5. The unassigned OTUs in grey should have an average correlation of zero.

4.4.2 Correlations between host and microbial network modules and to traits

After correcting for PCs containing information from the dominating freshwater-saltwater transition, the correlations between network modules are as predicted much less dramatic (Figure 4.10). The analysis reveals significant associations between different feed types (VO: arrow A and FO: arrow B) and the gene expression network module hME2, hME7 and hME8 (Figure 4.10). Moreover, two microbial abundance modules, mME2 (arrow C) and mME6 (arrow D), correlate with two gene expression modules hME3 and hME9.

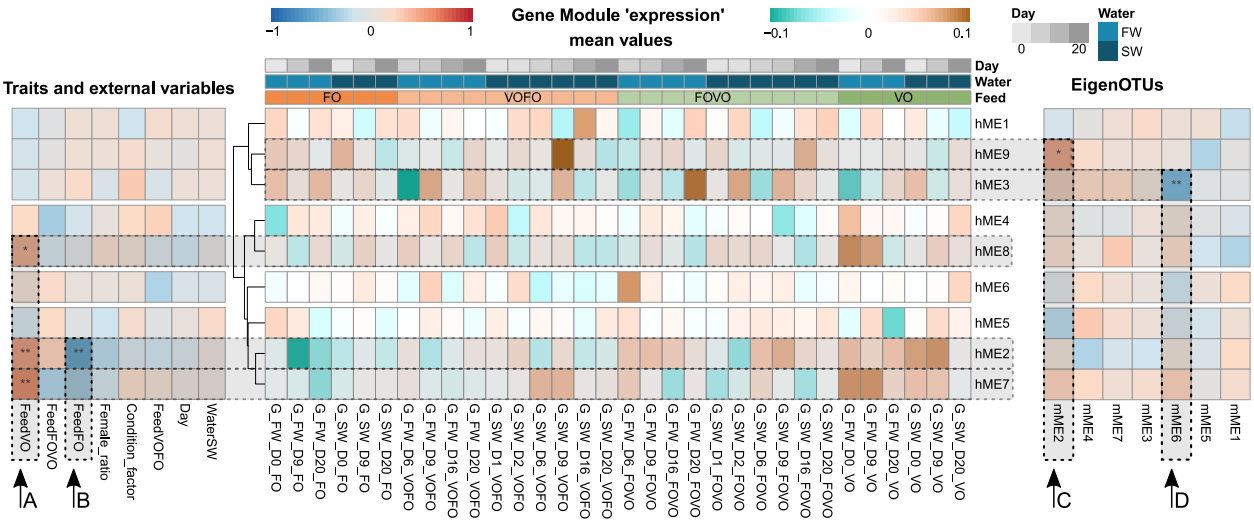

Figure 4.10: Split heatmap of PC-corrected data. The split heatmap shows the correlation of the host gene expression module eigengenes (hME) with the microbiome abundance module eigenOTUs (mME) and non-omics data such as traits and external variables. The central heatmap shows the mean 'expression' for each hME, each column is a sample group, organized by feed then water then day. The figure should be read horizontally, with each row constituting a new hME. The leftmost heatmap shows the correlation strength between hMEs and non-omics traits/external variables. The rightmost heatmap shows the correlation strength between each hME and each mME. Every correlation is based on the mean of biological replicates. A strong positive correlation is deep red, and a strong negative correlation is deep blue. A correlation of 0 is white. In each of the correlation-heatmaps the p-value of the correlation is marked with increasing number of stars as it reaches certain thresholds: * p <= 0.05, ** p <= 0.01, *** p <= 0.001. P-values are not corrected for multiple comparisons and should the interpreted as confidence strengths rather than rigorous hypotheses tests. Letters A-D mark interesting columns.

In each module there is a number of genes, and each gene can be ranked according to how connected it is. The most connected gene of each module is called the hub gene, which is presented below in Table 4.8.

Table 4.8: shows the gene in each module with the highest connectivity.

Module	Gene id	Gene name	Gene product
1	gene41808:106580649	LOC106580649	60S ribosomal protein L35a
2	gene49988:106588568	LOC106588568	lanosterol 14-alpha demethylase
3	gene28947:106567942	LOC106567942	cytoskeleton-associated protein 2-like
4	gene31337:106570212	LOC106570212	alpha-N-acetylgalactosaminide alpha-2,6-sialyltransferase 1-like
5	gene9067:106602547	glb1	galactosidase, beta 1
6	gene50894:106589386	LOC106589386	interferon-induced protein with tetratricopeptide repeats 5-like
7	gene4092:106580834	LOC106580834	protein S100-A1-like
8	gene35477:106574241	LOC106574241	desmin-like
9	gene28231:106567109	LOC106567109	G0/G1 switch protein 2-like

Because there were so few modules after PC-correction

The most connected gene in host expression module 2 is LOC106588568. This gene has, as is the case for many genes in the Atlantic salmon genome, been computationally classified and assigned membership to a defined protein family (UniProt 2015; The UniProt Consortium 2019), what UniProt calls "Protein inferred from homology". LOC106588568 most resembles and is believed to be an ortholog of Lanosterol 14-alpha demethylase which in humans (Uniprot:Q16850) and zebrafish (Uniprot:Q1JPY5) is part of steroid biosynthesis, more specifically the first step of synthesizing zymosterol from lanosterol. Zymosterol is an intermediate in cholesterol biosynthesis.

For host expression module 7 the most connected gene is LOC106580834. This gene is also inferred by homology to be Protein S100-A1-like which in humans "probably acts as a CA2+ signal transducer." It has its highest expression level in the heart. The hub gene of host expression module 8 is LOC106574241, inferred to be Desmin-like, a muscle specific type III intermediate-sized filament which is essential for proper muscular structure and function (Uniprot:P17661). The highest expression level is in the esophagogastric junction muscularis propria.

Host expression module 3 has LOC106567942 as its most central gene. Inferred by homology: Cytoskeleton-associated protein 2-like. This protein has its highest expression level in the testis and its expression is dependent on the cell-cycle with strong expression at the metaphase to telophase(Uniprot:Q8IYA6).

Presentation of host expression module 9 is dropped because the correlation is later shown to be caused by only one sample.

To further look into the correlations identified in Figure 4.10 we visualized the response of selected hMEs between different feeds. Since Feed is a nominal categorical variable and the hMEs are continuous the most natural visualization is a boxplot/stripchart combination.

In Figure 4.11A, we observed step like increases from FO to VO where VOFO and FOVO are intermediates. This step-like response is not observed in hME7 (Figure 4.11B) or hME8 (Figure 4.11C), but both these hMEs, however, also have higher expression in VO than the other feed categories. Using replicates instead of averages does not change the interpretation.

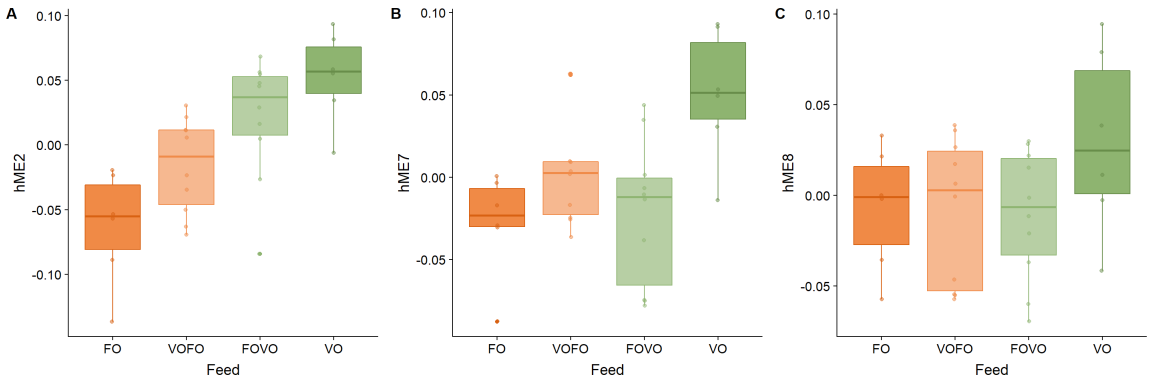

Figure 4.11: Boxplot of hME2, hME7 and hME8 response to different feeds. Y-axis is the 'expression' of the hME. X-axis divides this response into the various feed groups.

As was seen in table 4.8 above, each module has, in addition to an eigengene, a way to rank genes within the module. The gene with the highest connectivity is called the hub gene. Here the hub genes of modules 2, 7, and 8, were selected and their expression plotted as boxplots for the different types of feed. As expected, the response of the hub gene in modules 2 and 7 show the same expression behavior as their respective module eigengenes. The hub gene of module 8, however, is quite different from its eigengene, as seen in figure 4.12C the expression of this gene is highly influenced by two VO samples. The

mean expression between different feeds are not as large as the one observed in the mean expression of the module eigengene in figure 4.11C.

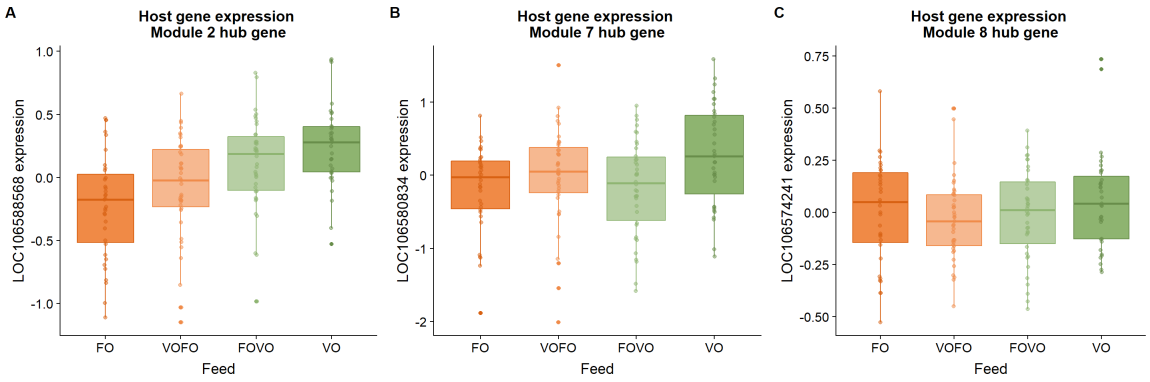

Figure 4.12: shows the expression response of specific hub genes for host expression modules 2, 7 and 8 respectfully. Because of the PC-correction approach the expression levels have been centered around zero. The figures also plot the biological replicates in contrast to figure 10 that plots only the sample group averages, i.e., the average of biological replicates.

In Figure 4.13A it is clear that only one sample group is responsible for the significant correlation. This correlation can therefore be discarded. In Figure 4.13B the correlation seems more robust and is not dependent on one sample group.

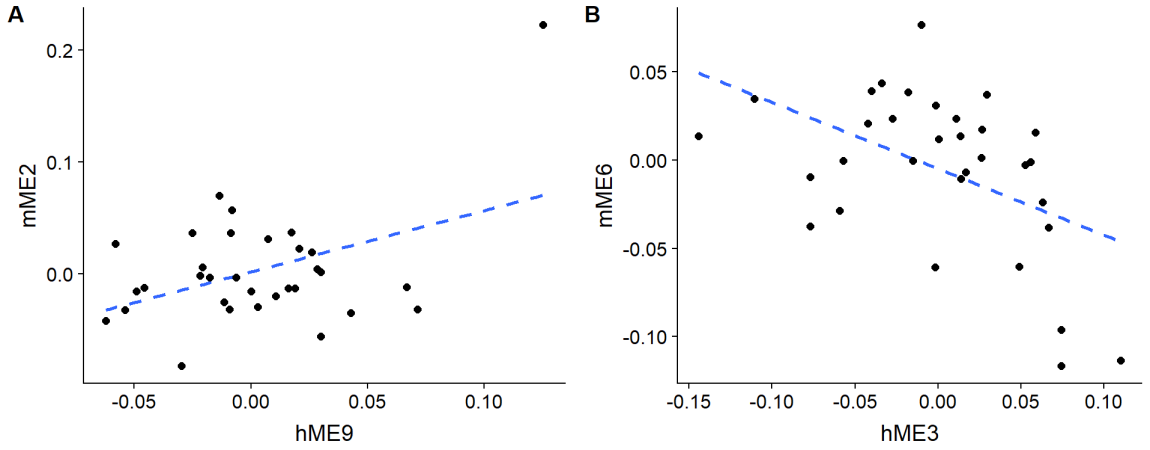

Figure 4.13: Two scatter plots with regression lines. Microbial abundance modules on the y-axis, host gene expression modules on the x-axis.

4.4.3 Selected GO enrichment

To investigate the link between expression modules and function in the host, we investigated the GO term enrichment in hME2, hME3, and hME7. hME2 has a strong positive correlation to the VO feed (Figure 4.10 and 4.11A) which is rich in vegetable oil (and low in LC-PUFA) and a strong negative correlation to FO feed (rich in LC-PUFA) (Figure 4.10 and 4.11A). In accordance with this, GO enrichment (only genes in module with correlation of >0.8 to the module eigengene is included) showed that the functions of these genes are linked to lipid and fatty acid metabolism (Table 4.9).

In genes that correlate (>0.5) to hME3, the module that corresponds to the microbial abundance module 6, there is enrichment for mitosis genes (Table 4.10). The most correlated gene is a gene involved in cell division and centrosome cycle. While the GO term with the highest median correlation is involved in control of the mitotic spindle.

Table 4.9: **GO enrichment of genes that correlate more than 0.8 to hME2.** Table shows the top 20 most significantly enriched GO terms in descending order of significance starting with the most significant. Observed = the number of genes that are annotated with the given GO term that also correlate ≥ 0.8 to hME2. Expected = the expected (i.e. expected by chance) number of genes that are annotated with the given GO term that also correlate ≥ 0.8 to hME2. P-value is calculated with the weight01 algorithm and fisher statistic. Median corr. is the median correlation to the respective eigengene for genes that correlate ≥ 0.8 to hME2. Max corr. is likewise based on eigengene correlation but shows the maximum correlation of any gene in the GO term i.e. the strongest correlation of any one gene. The highest maximum and median correlation is marked in red.

	GO.ID	Term	Observed	Expected	p-value	Median corr.	Max corr.
1	GO:0006695	Cholesterol biosynthetic process	21	0.09	1.0e-30	0.906	0.962
2	GO:0016129	Phytosteroid biosynthetic process	11	0.02	9.3e-16	0.906	0.948
3	GO:0006696	Ergosterol biosynthetic process	6	0.02	3.9e-15	0.926	0.948
4	GO:0045540	Regulation of cholesterol biosynthetic p...	8	0.05	2.6e-13	0.911	0.962
5	GO:0045338	Farnesyl diphosphate metabolic process	4	0	3.7e-12	0.915	0.934
6	GO:0055114	Oxidation-reduction process	13	2.15	3.2e-10	0.906	0.962
7	GO:0019287	Isopentenyl diphosphate biosynthetic pro...	3	0.01	1.7e-08	0.863	0.94
8	GO:0016114	Terpenoid biosynthetic process	4	0.05	1.3e-07	0.925	0.934
9	GO:0008299	Isoprenoid biosynthetic process	10	0.09	2.0e-06	0.915	0.954
10	GO:0006489	Dolichyl diphosphate biosynthetic proces...	2	0.01	1.9e-05	0.901	0.94
11	GO:0071398	Cellular response to fatty acid	4	0.23	7.3e-05	0.929	0.954
12	GO:0061051	Positive regulation of cell growth invol...	2	0.02	1.1e-04	0.929	0.934
13	GO:0009753	Response to jasmonic acid	2	0.02	1.6e-04	0.939	0.948
14	GO:0033574	Response to testosterone	4	0.15	1.9e-04	0.929	0.954
15	GO:0045542	Positive regulation of cholesterol biosy...	2	0.02	2.5e-04	0.929	0.934
16	GO:0015936	Coenzyme A metabolic process	2	0.03	3.2e-04	0.861	0.869
17	GO:0070723	Response to cholesterol	4	0.09	7.0e-04	0.929	0.954
18	GO:0044182	Filamentous growth of a population of un...	2	0.04	8.6e-04	0.939	0.948
19	GO:0046490	Isopentenyl diphosphate metabolic proces...	4	0.01	8.7e-04	0.88	0.94
20	GO:0071397	Cellular response to cholesterol	2	0.04	9.0e-04	0.901	0.954

Table 4.10: **GO enrichment of genes that correlate more than (>0.5 because they were so few) to hME3.** Table shows the top 20 most significantly enriched GO terms in descending order of significance starting with the most significant. Observed = the number of genes that are annotated with the given GO term that also correlate ≥ 0.5 to hME3. Expected = the expected (i.e. expected by chance) number of genes that are annotated with the given GO term that also correlate ≥ 0.5 to hME3. P-value is calculated with the weight01 algorithm and fisher statistic. Median corr. is the median correlation to the respective eigengene for genes that correlate ≥ 0.5 to hME3. Max corr. is likewise based on eigengene correlation but shows the maximum correlation of any gene in the GO term i.e. the strongest correlation of any one gene. The highest maximum and median correlation is marked in red.

	GO.ID	Term	Observed	Expected	p-value	Median corr.	Max corr.
1	GO:0000236	Mitotic prometaphase	18	0.69	8.7e-21	0.661	0.799
2	GO:0000088	Mitotic prophase	13	0.48	1.8e-15	0.549	0.799
3	GO:0000087	Mitotic M phase	29	1.7	1.1e-13	0.646	0.799
4	GO:0000281	Mitotic cytokinesis	15	0.72	1.5e-12	0.582	0.753
5	GO:0051301	Cell division	43	4.98	6.0e-12	0.605	0.809
6	GO:0007077	Mitotic nuclear envelope disassembly	9	0.28	9.5e-12	0.584	0.799
7	GO:0090307	Mitotic spindle assembly	15	0.58	8.8e-11	0.639	0.809
8	GO:0007147	Female meiosis II	5	0.03	1.1e-10	0.646	0.737
9	GO:0043987	Histone H3-S10 phosphorylation	5	0.04	3.0e-10	0.639	0.735
10	GO:0007098	Centrosome cycle	12	0.81	4.6e-10	0.624	0.809
11	GO:0051255	Spindle midzone assembly	10	0.14	9.0e-10	0.581	0.753
12	GO:0000086	G2/M transition of mitotic cell cycle	19	2.03	1.0e-09	0.64	0.799
13	GO:0007080	Mitotic metaphase plate congression	9	0.27	2.3e-09	0.702	0.76
14	GO:0007094	Mitotic spindle assembly checkpoint	9	0.35	5.2e-09	0.717	0.749
15	GO:0034501	Protein localization to kinetochore	6	0.14	6.6e-09	0.658	0.799
16	GO:0032467	Positive regulation of cytokinesis	11	0.44	7.3e-09	0.605	0.79
17	GO:0001556	Oocyte maturation	8	0.37	7.9e-09	0.684	0.737
18	GO:0000090	Mitotic anaphase	11	1.13	1.7e-08	0.639	0.76
19	GO:0007264	Small GTPase mediated signal transductio...	24	7.15	3.7e-08	0.648	0.799
20	GO:0000022	Mitotic spindle elongation	11	0.45	3.9e-08	0.601	0.753

From figure 4.10 we see that the module eigenOTUs of microbe abundance module 6 is negatively correlated to the eigengene of host gene expression module 3. From figure 4.13B we see that this correlation looks promising.

In table 4.11 we can see the five OTUs in the module, three are Cyanobacteria, of which two are from the genus *Planktothrix*. In addition to the three *cyanobacteria* there are two *proteobacteria*.

Table 4.11: OTUs in microbial abundance module 6. (hub) means that the OTU is the hub OTU having the highest connection strength in the module. All OTUs are ordered after correlation to the module eigenOTU. Correlation means the correlation of the OTU to the module eigenOTU.

	OTU	Correlation	Phylum	Class	Order	Family	Genus
1	OTU_7 (hub)	0.816	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Tropicibacter
2	OTU_3	0.801	Cyanobacteria		Oscillatoriales		Planktothrix
3	OTU_4	0.725	Cyanobacteria		Oscillatoriales		Planktothrix
4	OTU_20	0.678	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Photobacterium
5	OTU_9	0.283	Cyanobacteria	Oscillatoriophycideae	Chroococcales	Xenococcaceae	Chroococcidiopsis

The eigengene of host gene expression module 7 correlated with feed. As such it would be interesting to see what genes what terms that this module was enriched in. Unfortunately, even with a relaxed inclusion criterion, enriched GO terms for module 7 only had one gene. Therefore, the GO enrichment was dropped for this module.

Figure 4.14 shows a visual representation of the microbial abundance modules found. Compared to figure 4.5, there is much less structure in the graph.

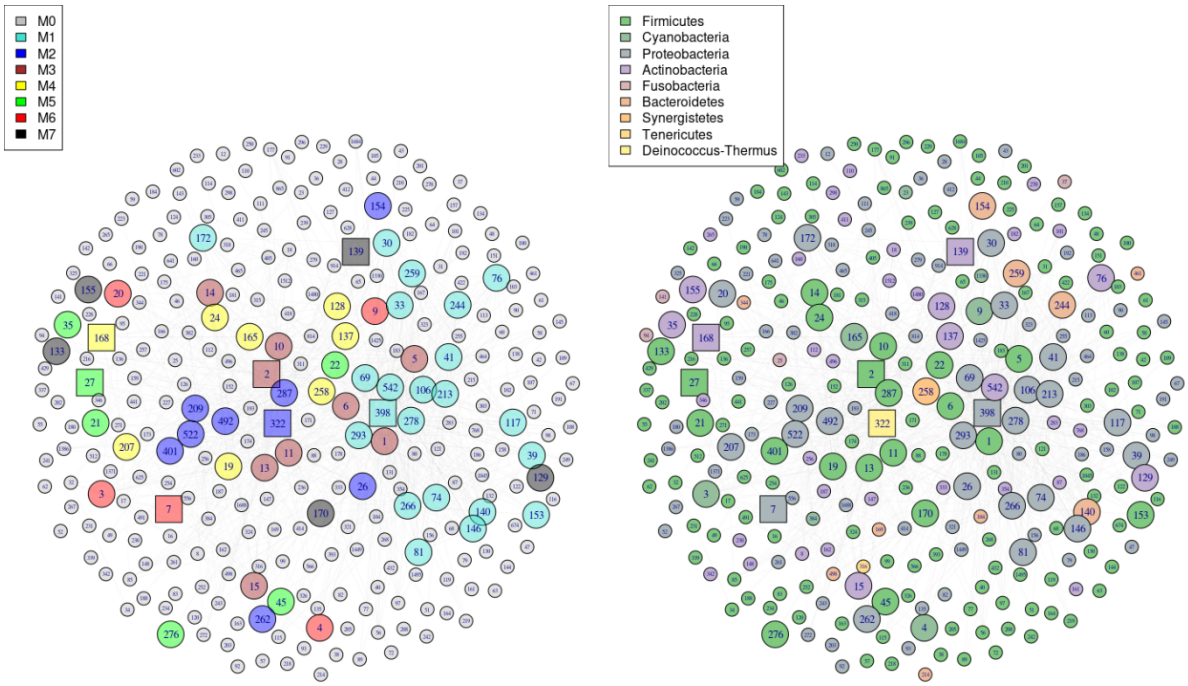

Figure 4.14: Two representations of the same OTU network. The left most graph shows module membership, the rightmost graph shows taxonomic classification of the phylum level. Each circle represents an OTU, a larger circle means it is part of a module. Squares are hub OTUs. Lines between OTUs are drawn based on the strength of the connection (TOM) – The more pronounced the line the stronger the connection. Layout with the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold 1991), implemented in the R-package igraph (Csárdi and Nepusz n.d.).

4.5 Comparison to principal component analysis (PCA)

A comparison between PCA and module eigennodes was also done for data that had been PC-corrected.

Figure 4.15 shows the biplots of both the microbial abundance **A** data and the gene expression data **B** after PC-correction has been applied. Compared to figure 4.6 we can see that very little is explained by each principal component. Interestingly enough, there seems to be a little separation between feeds where VO and FOVO tends towards the top right corner, and FO and VOFO tends towards the left bottom corner. Although there is no clean separation between them.

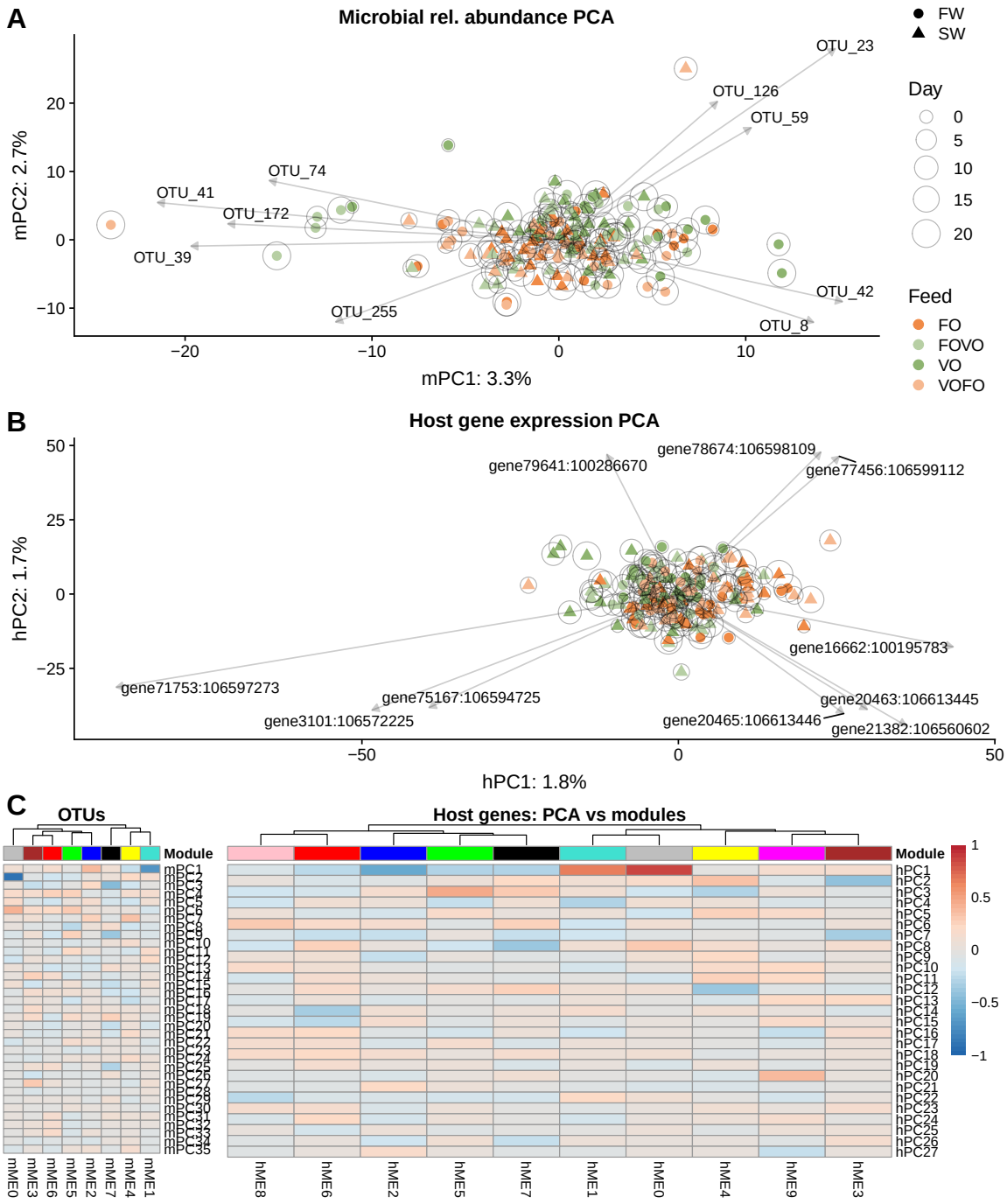

Figure 4.15: Comparison of PCA and Module eigenvectors. **A** and **B** shows biplots. X-axis is PC1 and y-axis is PC2. Points are the response of samples to the new PC1 and PC2 coordinate system, were the shapes signify whether the sample is from fresh or saltwater. Arrows are loadings i.e. how the original variables influence PCs, only the ten most influential variables are chosen based on the Euclidean distance they span between the two axes. **C** shows the correlation between "all data" PCs and the Modules eigengenes/eigenOTUs we found earlier. Deep blue indicates a strong negative correlation, while deep red indicates a strong positive correlation.

After applying the PC-correction we want to see what impact it has had on the explanatory power of the module eigennodes. Figure 4.16 shows how the eigengenes of the gene expression data has great levels of explained variance, but that the eigenOTUs no longer are good representations of the abundances in the module.

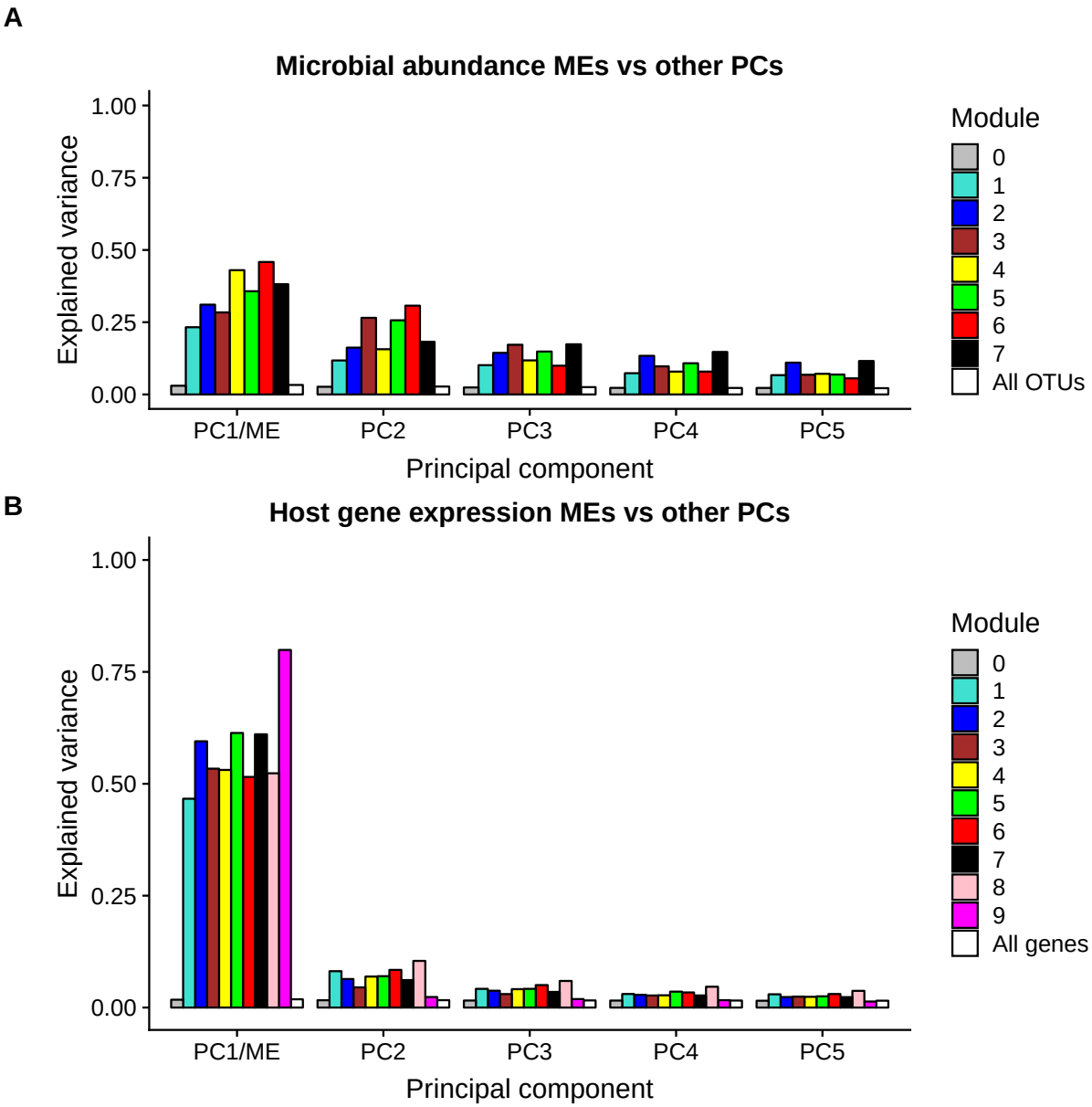

Figure 4.16: PCA of OTU subsets in A, and genes subsets in B. The first principal component (PC1) corresponds to the module eigengene (ME) hence the PC1/ME label. Y-axis of both figures show the proportion of explained variance. For all modules, and for the whole data, principal components 1-5 are shown. In WGCNA only the first PC, the eigennode, is used as a representative, all other PCs are disregarded.

Chapter 5

Discussion

5.1 Concept

In this thesis, we have developed a pipeline that uses the WGCNA package as a network-based dimensionality reduction approach to analyze host-microbiome multi-omics data. The resulting eigennodes can be used to relate groups of genes and groups of microbes to each other and other non-omics data variables. The result is a holistic view of the entirety of the data.

Beyond that, network statistics, such as the weighted degree, lets us rank members of modules according to their centrality.

5.2 Steps made to adapt WGCNA to 16S data

One of the characteristics that define 16S data is sparsity. In contrast to RNAseq data, where most genes are expressed throughout the experiment, most taxonomic variables appear only sporadically (Paulson et al. 2013). In an attempt to limit the number of zero values in the data, the microbial abundance data was filtered heavily. As the structure of the relationships between OTUs is affected by this, this is not an ideal approach. For proteomics and metabolomics data where the sparsity is also a problem, imputation has been used to correct for missing values (Pei, L. Chen, and W. Zhang 2017). However, when the percentage increases above a certain threshold, the imputation methods no longer work (Pei, L. Chen, and W. Zhang 2017). For OTU tables, the number of zero values can be even higher. It is also difficult to know whether zero-values are a result of data loss, or if it reflects actual absence. No imputation was therefore attempted.

Two parameter changes were necessary to use WGCNA on 16S data. The first is the reduction of the minimum module size from 30 to 5. This reduction of the minimum module size is justified by the low number of variables that are present in the microbial abundance data.

For the network construction and module detection of the 16S data, the use of partitioning around medoids (PAM) drastically changed the eigenOTU values. For modules where the size is limited, the PAM step is therefore not recommended. For module detection in the host gene expression network, there was no effect of turning PAM on or off.

5.3 Analysis (without removal of large effect variables)

Firstly, network and module creation was successful for both the host gene expression data and the microbial relative abundance data. Although the eigenOTUs displayed a very

strong inter-correlation. We will often merge such strong correlated modules. However, in this case, modules were kept as they were because there were so few of them. And because the freshwater-saltwater transition was known to be the cause.

Some of the modules found in the gene expression network were large (Figure 4.1C) – Containing several thousand genes. The large size of the modules is a direct result of the high mean connectivity (Figure 4.1B). The presence of one or more highly influential external variables can explain the high connectivity. For example, the transition from freshwater to saltwater is one such influential variable that results in many genes changing expression in the same direction and thus becoming co-expressed (Figure 4.3). This results in two large modules with two eigengenes: one with a negative relationship to the environmental transition (hME1 – host gene expression module eigengene 1), and one with a strong positive relationship to the transition (hME2).

Although it is not easy to separate Day as a variable from Water, there are some subtle differences. However, we will not go into more detail about these differences.

Using gene ontology enrichment, we found that host gene expression module 1 (hM1), for which hME1 is the eigengene, is over-represented in genes we can unilaterally relate to some aspect of growth and development. These development genes include genes for extracellular matrix organization, cell attachment, genes involved in cell differentiation, their shape and control of their growth rate, genes involved in the development of a skeletal system, and the formation of new blood vessels (angiogenesis) among many others (see Table 4.3). It no surprise then that this module, or rather its eigengene, has such a strong negative correlation to the transition to saltwater. After all, such genes are needed most in freshwater when the salmon is a juvenile.

During smoltification and transition to the saltwater life stage, the Atlantic salmon undergoes dramatic shifts in its lipid metabolism (Gillard et al. 2018). While the most significant GO-term for genes central to hM2 was intercellular protein transport, the GO-term with the highest correlating gene was GO:0006646 – phosphatidylethanolamine biosynthetic process. Phosphatidylethanolamine is the second most abundant phospholipid in the eukaryotic cell. In mammals, it is essential for a variety of cellular processes (Calzada, Onguka, and Claypool 2016). Other terms related to lipid metabolism were also found within the top 20 most enriched terms.

In freshwater, salmon has an endogenous production of LC-PUFA from other shorter dietary lipids. In saltwater, the salmon does not need such endogenous production as the lipids are readily available in its diet. There the salmon has instead adapted to take advantage of this increased LC-PUFA. While the endogenous production of LC-PUFA happens in the liver, the uptake of lipids must necessarily occur in the gut (Gillard et al. 2018). Therefore, an increase in the expression of genes involved in both transport and lipid influx is reasonable.

The GO enrichments in these two modules related to the fresh/saltwater transition are consistent with some of the changes we would expect to see between freshwater and saltwater.

All module eigenOTUs have a strong correlation to hME1 and hME2, but the strongest correlation is with mME4 (microbial module eigenOTU 4). From the relative abundance heatmap, figure 4.4, it is clear why; mM4 has central OTUs that appear only in saltwater (although some OTUs in the module have low abundances in freshwater). On the opposite side of the spectrum are OTUs in mM3, which are present in freshwater but disappears in saltwater. Both modules corroborate the correlations in figure 4.3, but are of little interest as far as host-microbe interactions go. Of more interest to us are the modules of mME1 and mME2.

The module mM2 has the greatest diversity of taxa among microbial abundance modules, as seen in figure 4.5. The module mM2 includes among its ranks, several *Cyanobacteria*

of genus *Planktothrix*, a photosynthetic organism. The hub OTU of this module, OTU_20 *photobacterium*, is known for both its ubiquity in marine environments, and in the intestinal contents of marine animals (*Photobacterium - an overview | ScienceDirect Topics* n.d.). The genus includes some bacteria that form symbiotic relationships with some luminous fish(*Photobacterium - an overview | ScienceDirect Topics* n.d.). This symbiosis, however, is relegated to specialized luminous organs (*Photobacterium - an overview | ScienceDirect Topics* n.d.) and not the fish gut. OTU_7 is also a *Proteobacteria* but in the genus *Tropicibacter*. The genus *Tropicibacter*, or genera in close association, degrades organic acids and are found in seawater all over the world from Indonesia (Harwati et al. 2009) and Japan (Iwaki, Nishimura, and Hasegawa 2012) to Mediterranean Spain (Lucena et al. 2012). The ubiquity of these bacteria, along with the presence of *Cyanobacteria* points to that this module might contain bacteria from the salmon's feed. Unfortunately, this study did not take samples of the feed for comparison, and therefore we cannot say for certain that they are.

By far the largest of the microbial abundance modules, mM1 has OTUs that persist across the entire sample space (see figure 4.4, arrow A), crossing the freshwater saltwater divide. The central OTUs in this module (table 4.6) are also consistent with the finds in Rudi et al. (2018) (table 4.5).

The hub OTU in this module is OTU_10 (genus *Vagococcus*). The species is related to another species that can utilize mucus in the host (Rudi et al. 2018). This ability can determine how readily a bacterial species can colonize the mucosal surface (Tailford et al. 2015).

So why not just use prevalence and abundance to find interesting OTUs? What would be the benefit of using network-based methods?

In a 2019 study, Layeghifard, Li, et al. (2019) compared three different approaches for identifying significant OTUs related to the prediction of patient health in patients suffering from cystic fibrosis. They found that network interconnectedness, identification of hub OTUs, was better at predicting the progression of the disease than taxa found by either high abundance or prevalence.

Agler et al. (2016) theorizes that the host controls the microbiota through hub microbes. These microbes then influence other microbes, though microbe-microbe interactions 'down the line'.Network analysis then gives a way of ranking these OTUs in a way that eases the discovery of potential symbionts to the host. Species in microbial communities are dependent on each other and should not be regarded as independent entities (Mallick et al. 2017)

As far as the correlations that the mME1 and mME2 have to the environmental shift, this is most likely because of the increase in abundance that these bacteria have in saltwater (Rudi et al. 2018). Because the transition from freshwater to saltwater dominates the data, it is difficult to find correlations to host gene expression modules that are the result of interaction instead of the influence of external variables. We have seen that mM2 and especially mM1 contain interesting OTUs. Still, beyond that, the total dominance of the freshwater-saltwater transition means that it is not possible with this pipeline to detect more subtle interactions.

Therefore, as stated previously in the results section, a method for the removal of large effect variables was used. This PC-correction had a massive impact on the creation of modules, both for the host gene expression data and for the microbial abundance data.

5.4 Comparison to PCA

Because a module eigennode in a network is the first principal component of the node profiles in that module, in this case, the gene expression or relative abundance, it was of interest to better understand the relationship between a principal component analysis and the module eigennodes. In the context of a PCA, the only difference between a weighted network analysis and a PCA is the subdivision of the omics variables, i.e. the WGCNA approach uses network clustering to subdivide genes/OTUs into modules and then performs PCA on each module while a normal PCA analysis performs a PCA on all variables.

For the analysis of the 16S data, there is hardly any difference in time usage between a weighted network analysis and a principal component analysis. For the gene expression data (RNAseq), however, tens of thousands of genes increase the computational resources needed and time required dramatically between a simple PCA and WGCNA. Because a weighted network analysis must calculate both a co-expression network ($O(n^2)$) and a TOM matrix ($O(n^3)$), a PCA has the great advantage of being much quicker (Time usage in section 2.8.4). What then are the benefits of subdividing the data into smaller groups based on their correlation i.e., finding modules?

As previously mentioned in the introduction, Kokou et al. (2018b), found a link between host genetics of cold-resistant fish and its gut microbiota. In their study, they analyzed the gut microbiota composition using 16S and the liver gene expression of the host response using RNAseq.

After finding the temperature to be a significant factor in shaping microbial composition (diversity and richness) using linear mixed model analysis, they used Canonical correlation Analysis on the top 5 principal components of the microbiome and the transcriptome. This analysis showed that 80% of the variance in the microbiome could be explained by the variance in the transcriptomics response.

For identifying influential OTUs, however, Kokou et al. (2018b) did not use the principal component analysis. Instead, they used Indicator Species Analysis, a way of finding species that are characteristic of a group of samples. The reason is simple: loadings, to what degree the original variables influence a principal component, are difficult to interpret biologically.

In figure 4.7 we can see that even in data, which should favor the use of a global method like PCA, because the saltwater transition dominates the entire data, the module eigennodes explain more of the variance than the principal components. This variance explanation illustrates both that the eigengene is a good representation of the expression/abundance in a module, and that subdividing the data can increase the explanatory power of the analysis.

The statement above is true under certain circumstances but with ambiguous interpretation. Because a PC is a PCA is a linear combination of all variables in the dataset, and the MEs are only linear combinations of a subset of the variables, they are not directly comparable. Using all components for both PCA and eigennodes, both will explain 100% of the variation. Using all variables misses the point of a PCA. If we instead limit the number of components, multiply each percentage of explained variance with the number of genes/OTUs in the module. For the PCA, the amount will be constant. Then MEs always have a slight lead. This total variance score includes the variance from the grey module. If genes/OTUs from the present in the grey module is removed from the dataset before PCA, the PCA will end up explaining more of the variance. That the PCA benefits from losing these genes/OTUs illustrate that the genes present in the grey module limit the explained variance of the PCA. Grey genes/OTUs contribute only noise to the data.

The above comparison considers the total explained variance of all components. However, this does not reflect the intended usage of MEs. Analysis in WGCNA only uses the first principal component as the module eigennode. WGCNA also disregards the ME0. If

we focus only on the MEs, without ME0, and multiply these with the number of nodes (genes/OTUs) in each module, the sum becomes a representative number for the variance explained of these MEs. If all principal components in the PCA get multiplied by every the number of gene/OTUs, the PCA will obviously "explain" more "variance." A fairer comparison will be one that divides the number of nodes equally between PCA components. With this approach, MEs "explain more variance." Again, the comparison may be a little unbalanced. We can, however, argue that a large percentage of known variance in a limited number of variables is easier to understand and decipher than the linear combination of all variables in the dataset.

A practical example of the difference between PCA and MEs can be seen in figure 4.6A. The figure is a biplot of the microbial abundance, and shows how one OTU is especially influential for the second principal component. OTU_8 *Corynebacterium*. The same OTU was marked in figure 4.4 arrow E. While this OTU will undoubtedly help to differentiate samples based on late-stage freshwater, the fact that it appears to do so alone indicates that it is of little importance to the community structure as a whole. And since it does not form any module, it "falls under the radar" in the network analysis.

Because of this difference in focus, genes, and OTUs that are deemed important in the context of an eigennode will not be considered relevant in a PCA. They, therefore, may not share any overlap in OTUs or genes. For the PCA, an OTU that is stable will contribute very little to the overall variance of the data, because, on a global scale, these OTUs have the least variation. In contrast, an OTU that appears only briefly like OTU_8 will, by its very sporadic nature, contribute more to the overall variance. However, since this OTU is not correlated to other OTUs, it will not form a module and will hence not be considered in the WGCNA approach. PCA gives more a global overview. If we are interested in the specific genes and microbes that are involved then network analysis offers greater insight.

5.5 Analysis – Removal of large effect variables

PC-correction (Parsana et al. 2019) was applied to both the gene expression data and to the microbial abundance data. The PC-correction approach estimates the number of principal components to remove using a permutation based approach from Buja and Eyuboglu (1992) called parallel analysis (PA). PA was originally intended for finding how many components to keep in a PCA, but the PC-correction approach instead uses the estimate to know how many components to remove. This means that the PC-correction approach will remove broad correlations, but should correlations that are caused by the influence between genes (Parsana et al. 2019).

One question is if this is valid for the the microbial abundance data or not, it was certainly not an explicit intent by Parsana et al. (2019). This might depend upon whether the reason for the observed modularity in the microbial abundance data is caused by response to external stimuli, or if it is caused by microbe-microbe interactions. It could also be the case that the PC-correction removes variance were one microbe influences the whole community.

It is also of interest to see if this way of removing confounding variables/large effect variables is useful for the pipeline, or if it just ends up removing the basis for why this approach works. Or goal is only indirectly linked to inference of the networks, more critical is the correlation of eigengenes to eigenOTUs and to external variables.

For the PC-corrected data network and module creation was successful for both the host gene expression data (Figure 4.8) and the microbial relative abundance data (Figure 4.9).

As expected the module size of the gene expression network decreased dramatically. which also unsurprisingly makes the mean connectivity very small. Just as an aside: One remedy for the small module size can be to choose a smaller β-value in the network construc-

tion. The PC-correction had less impact on the module sizes in the microbial abundance network 4.9C. The correction even made the mean connectivity of the microbial abundance network increase (Figure 4.9B). The intermodular correlation as seen in figure 4.9D, have been reduced. But the intramodular correlations have deceased (Figure 4.9F). The modules sizes in the microbial abundance network seems less affected by the PC-correction.

We will in the following paragraphs refer to the modules of the PC-corrected network with the prefix pcc-, and the uncorrected network modules as all-.

Eigengene pcc-hME9 has a positive correlation to eigenOTU pcc-mME2, but this correlation is only influenced by only one sample (Figure 4.13A). Disregarding this "significant" correlation there was only one strong negative correlation left. For any symbiotic relation the relation would be mutual and therefore the correlation should be positive. This could leave us to conclude that there are in fact no interactions between the Atlantic salmon host and the gut microbiota. Is this likely?

It is useful to recap some of what we know about the microbiome in fish and in the Atlantic salmon. Host-associated microorganisms everywhere in nature (Sommer and Bäckhed 2013). We know that some fish are able to retain microbes, and that some fish might have substantial number of symbionts (Sullam et al. 2012). We also know that feed has some influence on the establishment of gut microbiota in the Atlantic salmon (Jin et al. 2019). That different diets with low or high LC-PUFA content feed in freshwater influences the microbial composition in saltwater.

There has been some attempts at elucidating what the core microbiota of the farmed salmon gut might be, estimates of around 19-20 (Gajardo et al. 2016; Dehler, Secombes, and Martin 2017) have been found by some. But as discussed in Rudi et al. (2018), the farmed Atlantic salmon might have a potentially low number of host-associated bacterial species but that this is only evident when the fish is followed trough the environmental shift from freshwater to saltwater. Finding no interaction between the Atlantic salmon host and its gut microbiota is therefore quite significant.

Besides the lack of any strong positive correlations between host and microbiota there is a strong negative correlation between eigengene pcc-hME3 and eigenOTU pcc-mME6 (Figure 4.13B). Almost every OTU found in pcc-mM6 except OTU_9 (Table 4.11) was present in all-mM2 (Table 4.7).

After PC-correction, the data should no longer have anything to do with the transition to salt water, or day, or any other external influence, which significantly affects large numbers of variables in the data.

Genes in pcc-hME2 have a strong positive correlation to vegetable feed (VO) and a strong negative correlation to fish oil (FO) (Figure 4.10. The response can be seen to increase in a step-wise manner (Figure 4.11A), the hub gene of this module shares the same response (Figure 4.12A) although slightly less clear (Note that data point in Figure 4.12 are not the averages but all replicates.)

The module is most enriched in different genes that are involved in the synthesis of cholesterol (21/90). The most connected gene of the module, the hub gene, is *lanosterol 14-alpha demethylase* (LOC106588568). This enzyme is the first step in synthesizing zymosterol (an intermediate in cholesterol biosynthesis) from lanosterol by removing a C-14α-methyl group, part of a highly conserved metabolic pathway (Lepesheva and Waterman 2007). Cholesterol is essential in cell membranes and a precursor in the synthesis of bile acids and steroid hormones (Cerqueira et al. 2016).The success of the PC-correction method in revealing genes involved in the synthesis of cholesterol shows that the eigengenes can still be related to external variables.

While the lack of a sharp drop after the ME in the scree plot of the PC-corrected microbial abundance modules is purely circumstantial evidence figure 4.16 seems to show

that PC-correction was not beneficial for network construction and module detection in the microbial abundance data. The TOM matrix graph of the PC-corrected microbial abundance data further emphasizes this point (Figure 4.14), as the modules seem chaotic.

In the uncorrected data, the network construction and module detection gave several convincing candidate OTUs that corresponded with Rudi et al. (2018). This correspondence suggests to us that it is the PC-correction approach does not work for microbial abundance data.

It is undoubtedly true that the characteristics of the two datasets are wildly different. For one, the structure of the host gene expression data is much more complex than the community structure of the microbial community. The PC-correction method has some assumptions that the microbial abundance data might not fulfill.

One of the fundamental assumptions made in the PC-correction method is that the network to be inferred has scale-free topology (Parsana et al. 2019). The microbial network did achieve scale-free topology both before (Figure 4.2) and after the PC-correction of the data (Figure 4.9), which would exclude this as the reason for the method not working.

5.5.1 What are we measuring?

The use of the 16S gene to detail the complex microbial communities is challenging. The technology is defined by its many problems (Brooks et al. 2015; Martinson et al. 2019).

For us, the correlation between OTUs can obscure the actual underlying system. OTUs are imperfect representations of real microbes. While there exist methods that promise finer resolution of the 16S gene (Callahan et al. 2016), the use of the 16S gene is in it self limiting.

An improvement could be the use of metatranscriptomics Aylward et al. (2015). Treating the entire microbial community as one transcriptome mapping to orthologous genes Aylward et al. (2015). While the increase in complexity and data size is dramatic, WGCNA is already made for handling large data.

Eukaryotic gene regulation is complex. Only 40% of the protein level explained by mRNA (Vogel and Marcotte 2012). For this thesis, we summarized genes, potentially obscuring gene isoforms that have different functions. Overlapping transcription factors also means that we cannot elucidate every behavior of the gene regulatory network from gene expression data (Kel et al. 2006).

Going even further, true interactions between the host and its microbiome happens primarily on the protein level. (Guven-Maiorov, Tsai, and Nussinov 2017). Pathogens, for example, use molecular mimicry to hijack host pathways (Guven-Maiorov, Tsai, and Nussinov 2016).

Including reference sequences, spike-ins, in future studies, both for RNA-seq and 16S-seq, will reduce the complexities of data-specific normalization approaches. An RNA spike-in is a quantity of known sequences that can be identified as separate from the data that is being analyzed but still shares similar characteristics e.g., GC content, length (K. Chen et al. 2016). There has also been some work developing a synthetic spike-in for 16S data (Tourlousse et al. 2016). Spike-ins can give ground truth to the analysis, allowing absolute quantification and improving reproducibility (Tourlousse et al. 2016).

5.5.2 Proportionality

There has been a discussion about the use of correlation measures on data that, in principle, should be thought of as compositional (Erb and Notredame 2016; Lovell et al. 2015; Thomas P. Quinn et al. 2017).

Counts from sequencers are always proportions of reads, which means that we do not get absolute counts but compositions. This composition means that what looks like a reduction in one gene's expression might instead be the increase of expression from other genes. The same applies equally to the amplicon sequencing of 16S rRNA or other marker genes. This fact calls into question the validity of using correlation coefficients such as Pearson or the biweight midcorrelation on such data.

Avoiding compositions means that normalization methods like transcript per million (TPM) should not be used for RNAseq data. Normalization techniques such as 'trimmed mean of M-values' (TMM) try to re-scale each sample with the assumption that most genes are not differentially expressed (Robinson and Oshlack 2010). If the assumption holds, expression values can now be used as if they are absolute values. For now, the pipeline uses this normalization.

For the 16S data in this dataset we can clearly see in figure 4.4 that stability is the exception and not the rule. This excludes the use of methods like TMM.

Many papers dealing with correlation of OTUs recommend using transformations such as the centered log-ratio CLR or the additive log-ratio (ALR) transform (Thomas P Quinn et al. 2018; Gloor et al. 2017). Both these transformations requires a reference but differ in how it is found. CLR uses the geometric mean as a reference which means that CLR also implicitly assumes that most variables remain unchanged (Erb and Notredame 2016). Again, this is clearly not the case for this dataset. ALR depends on an unchanged reference (Erb and Notredame 2016). In an exploratory analysis, this would be difficult to find. Although a solution for this is using a spike-in as mentioned above.

The log-ratio variance (VLR) from compositional data analysis would make a good measure of association were it not for the fact that it is not comparable between pairs of variables (Thomas P. Quinn et al. 2017). This makes it unsuitable for correlation network analysis.

There have been proposed at least three different measures related to proportionality as substitute for correlation (Thomas P. Quinn et al. 2017). These coefficients offers a potential to avoid spurious associations that happens when correlation is applied to relative data (Thomas P. Quinn et al. 2017).

ϕ, ρ_p and ϕ_s, all are implemented in the R-package propr, see figure 5.1. Of the three, the measure ρ_p seems to have the most proper properties of the proportion measures in propr. It is worth noting that there is still no consensus on the analysis of relative abundance data (Thomas P. Quinn et al. 2017).

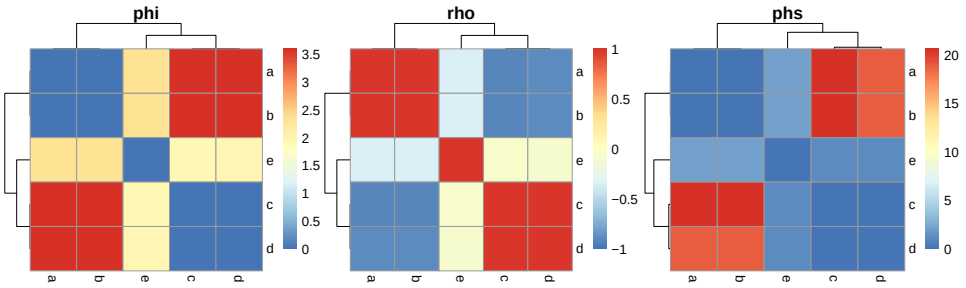

Figure 5.1: Characteristics of three proportionality measures detailed in Thomas P. Quinn et al. (2017). From the left: phi = ϕ, rho = ρ_p and phs = ϕ_s.

Initial tests using ρ_p as a measure gave uncharacteristic behavior in the mean connectivity plot – Whereas the mean connectivity is expected to monotonically decrease with the increase of β (Horvath 2011, p.82), the mean connectivity with applied ρ_p instead gave an unexpected rise then drop (Figure .4) in the appendix. Further work on the applicability of

proportionality measures for WGCNA should be done before any conclusive statements are made.

5.5.3 Experimental design

The evaluation of the pipeline in this thesis was limited to an experiment that was not designed to answer the question of specific host-microbiome interaction. The main interest of the study was to elucidate the regulation of lipid metabolism genes during freshwater and saltwater stages by perturbing the feeds of the fish (Gillard et al. 2018). Because the microbial data came as an afterthought, there was no analysis done of the microbial composition of the feed.

While the computational removal of large effect variables worked to reveal genes involved in response to different feeds, it did not reveal any positive correlations with any OTU module. There can be several reasons for this finding, one of which is that the removal of 5 principal components was too harsh.

Controlling for large effect variables computationally like we have done in this thesis should be the last resort. If it can be controlled for in the experimental design, then it should be addressed there.

It could be that 16S data is not well suited for discovering such interactions. 16S data has a lot of unfavorable characteristics, such as ill-defined species resolution and high degree of sparsity. However, there are studies where 16S was used to find such interactions.

These studies however are usually done on much simpler systems with less heterogeneous data. Such studies usually use mice that are raised germ-free, or mice that have been treated with antibiotics to become germ-free (Kennedy, King, and Baldridge 2018).

It should be stated that where the Transkingdom approach has been successful is with more suited experimental design, for example the use of knockout mice (Greer et al. 2016). For a dataset as complicated and dominated as this, it is difficult to see how potential interactions could make much impact on the data as a whole.

One way to side-step the problems created by the use of 16S data, is to expand the experiment to use whole-community transcriptomics. One obvious drawback is the increased cost and complexity this would add to the gathering of data. However, when correctly pre-processed, this data could go directly into the pipeline with only slight adjustments to the pipeline parameters. WGCNA can deal with such large datasets.

In their 2015 paper, Aylward et al. (2015) used WGCNA to analyze the day-night photosynthetic cycle of different ocean communities of bacteria. To study the transcriptional regulation of the entire community, Aylward et al. (2015) mapped reads to ortholog protein clusters, then analyzed the resulting data with WGCNA.

It is also crucial to keep in mind that the gut environment that the microbes inhabit is very complex and can differ significantly by location (Donaldson, Lee, and Mazmanian 2015). Even on the local level, inter-fold regions of the intestine can have very different microbes (Donaldson, Lee, and Mazmanian 2015). Keeping track of these locations will undoubtedly increase the complexity of the sampling procedures. But, if we are to gain a complete picture of host-microbial interaction, we cannot ignore such details.

5.5.4 Simulation

The only rigorous way to evaluate which settings and methods that give better results are to know the real classifications and relations within the data. Users can know of such data, or they can try to simulate it.

This thesis and pipeline would benefit significantly from a simulated dataset where the covariance structure is known. For this thesis, sufficiently realistic simulations were not achievable within the allotted time.

There is a need for general simulators that has proper documentation and has been benchmarked and validated through peer-review (Escalona, Rocha, and Posada 2016). For this pipeline and others like it, the simulation would need to create covariance structures that have has realistic network behavior. What this exactly entails is beyond the scope of this thesis.

WGCNA has some simulation capability, it is, however, somewhat limited, and independent simulation packages would be preferable as in-package-simulations are expected to be over-optimized to recreate structures and characteristics that the method is sensitive for in the data.

There exist quite a few simulations packages for simulating DNA sequences (Escalona, Rocha, and Posada 2016). Unfortunately, most simulators for next-generation sequencing data are simply intended for testing OTU clustering methods, differential abundance/expression, or for testing quantification pipelines. For this pipeline, however, emphasis should be put on controlling the covariance structure of the data.

Bibliography

Agler, Matthew T. et al. (Jan. 2016). "Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation". In: *PLoS Biology* 14.1. ISSN: 15457885. DOI: 10.1371/journal.pbio.1002352.

Alberts, Bruce et al. (2015). *Molecular Biology of the cell*. 6th. Garland Science. ISBN: 978-0-81-53-4464-3.

Alexa, Adrian and Jorg Rahnenfuhrer (2018). *topGO: Enrichment Analysis for Gene Ontology*.

Alexa, Adrian, Jörg Rahnenführer, and Thomas Lengauer (July 2006). "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure". In: *Bioinformatics* 22.13, pp. 1600–1607. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl140.

Alne, Henriette et al. (Jan. 2011). "Reduced growth, condition factor and body energy levels in Atlantic salmon Salmo salar L. during their first spring in the sea". In: *Aquaculture Research* 42.2, pp. 248–259. ISSN: 1355557X. DOI: 10.1111/j.1365-2109.2010.02618.x.

Ashburner, Michael et al. (May 2000). *Gene ontology: Tool for the unification of biology*. DOI: 10.1038/75556.

Aylward, Frank O et al. (Apr. 2015). "Microbial community transcriptional networks are conserved in three domains at ocean basin scales". In: *Proceedings of the National Academy of Sciences* 112.17, pp. 5443–5448. ISSN: 0027-8424. DOI: 10.1073/pnas.1502883112.

Ballouz, S., W. Verleyen, and J. Gillis (July 2015). "Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers". In: *Bioinformatics* 31.13, pp. 2123–2130. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv118.

Barabási, Albert László and Zoltán N. Oltvai (2004). "Network biology: Understanding the cell's functional organization". In: *Nature Reviews Genetics* 5.2, pp. 101–113. ISSN: 14710056. DOI: 10.1038/nrg1272.

Barabási, Albert-László and Réka Albert (Oct. 1999). "Emergence of Scaling in Random Networks". In: *Science* 286.5439, pp. 509–512. ISSN: 0036-8075. DOI: 10.1126/science.286.5439.509.

Barnham, Charles and Alan Baxter (1998). *Condition Factor, K, for Salmonid Fish*. Tech. rep. State of Victoria, Department of Primary Industries, pp. 1–3.

Black, Paul E. (2005). *Greedy Algorithm*.

Bordenstein, Seth R. and Kevin R. Theis (Aug. 2015). "Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes". In: *PLOS Biology* 13.8. Ed. by Matthew K. Waldor, e1002226. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1002226.

Broido, Anna D. and Aaron Clauset (Dec. 2019). "Scale-free networks are rare". In: *Nature Communications* 10.1, p. 1017. ISSN: 20411723. DOI: 10.1038/s41467-019-08746-5.

Brooks, J Paul et al. (Dec. 2015). "The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies Ecological and evolutionary microbiology". In: *BMC Microbiology* 15.1, p. 66. ISSN: 14712180. DOI: 10.1186/s12866-015-0351-6.

Buja, Andreas and Nermin Eyuboglu (Oct. 1992). "Remarks on Parallel Analysis". In: *Multivariate Behavioral Research* 27.4, pp. 509–540. ISSN: 0027-3171. DOI: 10.1207/s15327906mbr2704{_}2.

Callahan, Benjamin J. et al. (2016). "DADA2: High-resolution sample inference from Illumina amplicon data". In: *Nature Methods* 13.7, pp. 581–583. ISSN: 15487105. DOI: 10.1038/nmeth.3869.

Calzada, Elizabeth, Ouma Onguka, and Steven M. Claypool (2016). "Phosphatidylethanolamine Metabolism in Health and Disease". In: *International Review of Cell and Molecular Biology*. Vol. 321. Elsevier Inc., pp. 29–88. ISBN: 9780128047071. DOI: 10.1016/bs.ircmb.2015.10.001.

Cao, Xi Hang, Ivan Stojkovic, and Zoran Obradovic (Sept. 2016). "A robust data scaling algorithm to improve classification accuracies in biomedical data". In: *BMC Bioinformatics* 17.1, p. 359. ISSN: 1471-2105. DOI: 10.1186/S12859-016-1236-X.

Cerqueira, Nuno M. F. S. A. et al. (Oct. 2016). "Cholesterol Biosynthesis: A Mechanistic Overview". In: *Biochemistry* 55.39, pp. 5483–5506. ISSN: 0006-2960. DOI: 10.1021/acs.biochem.6b00342.

Chen, Kaifu et al. (Mar. 2016). "The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses". In: *Molecular and Cellular Biology* 36.5, pp. 662–667. ISSN: 0270-7306. DOI: 10.1128/mcb.00970-14.

Csárdi, Gábor and Tamás Nepusz (n.d.). *The igraph software package for complex network research*. Tech. rep.

Dehler, Carola E., Christopher J. Secombes, and Samuel A.M. Martin (Jan. 2017). "Environmental and physiological factors shape the gut microbiota of Atlantic salmon parr (Salmo salar L.)" In: *Aquaculture* 467, pp. 149–157. ISSN: 00448486. DOI: 10.1016/j.aquaculture.2016.07.017.

Desbonnet, L et al. (Feb. 2014). "Microbiota is essential for social development in the mouse". In: *Molecular Psychiatry* 19.2, pp. 146–148. ISSN: 1359-4184. DOI: 10.1038/mp.2013.65.

DiLeo, Matthew V. et al. (Oct. 2011). "Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome". In: *PLoS ONE* 6.10. Ed. by Peter Csermely, e26683. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0026683.

Dominguez-Bello, Maria Gloria et al. (June 2019). "Role of the microbiome in human development." In: *Gut* 68.6, pp. 1108–1114. ISSN: 1468-3288. DOI: 10.1136/gutjnl-2018-317503.

Donaldson, Gregory P., S. Melanie Lee, and Sarkis K. Mazmanian (Dec. 2015). *Gut biogeography of the bacterial microbiota*. DOI: 10.1038/nrmicro3552.

Dong, Jun and Steve Horvath (Dec. 2007). "Understanding network concepts in modules". In: *BMC Systems Biology* 1.1, p. 24. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-24.

Duran-Pinedo, Ana E. et al. (Dec. 2011). "Correlation Network Analysis Applied to Complex Biofilm Communities". In: *PLoS ONE* 6.12. Ed. by Jack Anthony Gilbert, e28438. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0028438.

Egerton, Sian et al. (May 2018). *The gut microbiota of marine fish*. DOI: 10.3389/fmicb.2018.00873.

Eisen, Michael B. et al. (1998). "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.22, pp. 12930–12933.

Erb, Ionas and Cedric Notredame (June 2016). "How should we measure proportionality on relative gene expression data?" In: *Theory in biosciences = Theorie in den Biowissenschaften* 135.1-2, pp. 21–36. ISSN: 1611-7530. DOI: 10.1007/s12064-015-0220-8.

Escalona, Merly, Sara Rocha, and David Posada (Aug. 2016). "A comparison of tools for the simulation of genomic next-generation sequencing data". In: *Nature Reviews Genetics* 17.8, pp. 459–469. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.57.

Faust, Karoline and Jeroen Raes (Aug. 2012). "Microbial interactions: from networks to models". In: *Nature Reviews Microbiology* 10.8, pp. 538–550. ISSN: 1740-1526. DOI: 10.1038/nrmicro2832.

Freeman, Linton C. (Jan. 1978). "Centrality in social networks conceptual clarification". In: *Social Networks* 1.3, pp. 215–239. ISSN: 03788733. DOI: 10.1016/0378-8733(78)90021-7.

Friedman, Jonathan and Eric J. Alm (Sept. 2012). "Inferring Correlation Networks from Genomic Survey Data". In: *PLoS Computational Biology* 8.9. Ed. by Christian von Mering, e1002687. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002687.

Fruchterman, Thomas M. J. and Edward M. Reingold (Nov. 1991). "Graph drawing by force-directed placement". In: *Software: Practice and Experience* 21.11, pp. 1129–1164. ISSN: 00380644. DOI: 10.1002/spe.4380211102.

Gaby, John Christian and Daniel H. Buckley (July 2012). "A comprehensive evaluation of PCR primers to amplify the nifH gene of nitrogenase". In: *PLoS ONE* 7.7. Ed. by Jose Luis Balcazar, e42149. ISSN: 19326203. DOI: 10.1371/journal.pone.0042149.

Gajardo, Karina et al. (Aug. 2016). "A high-resolution map of the gut microbiota in Atlantic salmon (Salmo salar): A basis for comparative gut microbial research". In: *Scientific Reports* 6. ISSN: 20452322. DOI: 10.1038/srep30893.

Geng, Haifeng et al. (July 2016). "Changes in the Structure of the Microbial Community Associated with Nannochloropsis salina following Treatments with Antibiotics and Bioactive Compounds". In: *Frontiers in Microbiology* 7. ISSN: 1664-302X. DOI: 10.3389/fmicb.2016.01155.

Gillard, Gareth et al. (2018). "Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon". In: *Molecular Ecology* 27.5, pp. 1200–1213. ISSN: 1365294X. DOI: 10.1111/mec.14533.

Gloor, Gregory B et al. (2017). *Microbiome datasets are compositional: And this is not optional*. DOI: 10.3389/fmicb.2017.02224.

Greer, Renee L. et al. (Dec. 2016). "Akkermansia muciniphila mediates negative effects of IFNγ on glucose metabolism". In: *Nature Communications* 7.1, p. 13329. ISSN: 2041-1723. DOI: 10.1038/ncomms13329.

Guven-Maiorov, Emine, Chung Jung Tsai, and Ruth Nussinov (Oct. 2016). *Pathogen mimicry of host protein-protein interfaces modulates immunity*. DOI: 10.1016/j.semcdb.2016.06.004.

— (Oct. 2017). *Structural host-microbiota interaction networks*. Ed. by Richard A. Bonneau. DOI: 10.1371/journal.pcbi.1005579.

Hammer, Tobin J, Jon G Sanders, and Noah Fierer (May 2019). "Not all animals need a microbiome". In: *FEMS Microbiology Letters* 366.10. ISSN: 1574-6968. DOI: 10.1093/femsle/fnz117.

Harwati, Theresia Umi et al. (2009). "Tropicibacter naphthalenivorans gen. nov., sp. nov., a polycyclic aromatic hydrocarbon-degrading bacterium isolated from Semarang Port in Indonesia". In: *International Journal of Systematic and Evolutionary Microbiology* 59.2, pp. 392–396. ISSN: 14665026. DOI: 10.1099/ijs.0.65821-0.

Hawrylycz, Michael J. et al. (Sept. 2012). "An anatomically comprehensive atlas of the adult human brain transcriptome". In: *Nature* 489.7416, pp. 391–399. ISSN: 00280836. DOI: 10.1038/nature11405.

Horvath, Steve (2011). *Weighted Network Analysis*. New York, NY: Springer New York, pp. 77–78. ISBN: 978-1-4419-8818-8. DOI: 10.1007/978-1-4419-8819-5.

Horvath, Steve and Jun Dong (Aug. 2008). "Geometric Interpretation of Gene Coexpression Network Analysis". In: *PLoS Computational Biology* 4.8. Ed. by Satoru Miyano, e1000117. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000117.

Huang, Yong et al. (July 2017). "Microbes Are Associated with Host Innate Immune Response in Idiopathic Pulmonary Fibrosis". In: *American Journal of Respiratory and Critical Care Medicine* 196.2, pp. 208–219. ISSN: 1073-449X. DOI: 10.1164/rccm.201607-1525OC.

Illumina (n.d.). *16S Metagenomic Sequencing Library Preparation*.

Iwaki, Hiroaki, Ayaka Nishimura, and Yoshie Hasegawa (Apr. 2012). "Tropicibacter phthalicus sp. nov., a phthalate-degrading bacterium from seawater". In: *Current Microbiology* 64.4, pp. 392–396. ISSN: 03438651. DOI: 10.1007/s00284-012-0085-8.

IWGSC (Aug. 2018). "Shifting the limits in wheat research and breeding using a fully annotated reference genome". In: *Science* 361.6403, eaar7191. ISSN: 0036-8075. DOI: 10.1126/science.aar7191.

Jin, Y et al. (Jan. 2019). "Atlantic salmon raised with diets low in long-chain polyunsaturated n-3 fatty acids in freshwater have a Mycoplasma-dominated gut microbiota at sea". In: *Aquaculture Environment Interactions* 11, pp. 31–39. ISSN: 1869-215X. DOI: 10.3354/aei00297.

Kel, Alexdander et al. (Sept. 2006). "Beyond microarrays: Finding key transcription factors controlling signal transduction pathways". In: *BMC Bioinformatics* 7.S2, S13. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-S2-S13.

Kennedy, Elizabeth A, Katherine Y King, and Megan T Baldridge (2018). "Mouse Microbiota Models: Comparing Germ-Free Mice and Antibiotics Treatment as Tools for Modifying Gut Bacteria." In: *Frontiers in physiology* 9, p. 1534. ISSN: 1664-042X. DOI: 10.3389/fphys.2018.01534.

Kokou, Fotini et al. (Nov. 2018a). "Host genetic selection for cold tolerance shapes microbiome composition and modulates its response to temperature". In: *eLife* 7, e36398. ISSN: 2050084X. DOI: 10.7554/eLife.36398.

— (Nov. 2018b). "Host genetic selection for cold tolerance shapes microbiome composition and modulates its response to temperature". In: *eLife* 7. ISSN: 2050-084X. DOI: 10.7554/eLife.36398.

Langfelder, Peter (2018). *Signed or unsigned: which network type is preferable?*

Langfelder, Peter and Steve Horvath (Dec. 2007). "Eigengene networks for studying the relationships between co-expression modules". In: *BMC Systems Biology* 1.1, p. 54. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-54.

— (Dec. 2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1, p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559.

— (2014). *Tutorial for the WGCNA package for R*. Tech. rep.

Langfelder, Peter, Bin Zhang, and Steve Horvath (Mar. 2008). "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R". In: *Bioinformatics* 24.5, pp. 719–720. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btm563.

Layeghifard, Mehdi, David M. Hwang, and David S. Guttman (Mar. 2017). "Disentangling Interactions in the Microbiome: A Network Perspective". In: *Trends in Microbiology* 25.3, pp. 217–228. ISSN: 0966842X. DOI: 10.1016/j.tim.2016.11.008.

Layeghifard, Mehdi, Hannah Li, et al. (Dec. 2019). "Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations". In: *npj Biofilms and Microbiomes* 5.1, p. 4. ISSN: 20555008. DOI: 10.1038/s41522-018-0077-y.

Leek, Jeffrey T et al. (Mar. 2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6, pp. 882–883. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/bts034.

Lepesheva, Galina I. and Michael R. Waterman (Mar. 2007). *Sterol 14α-demethylase cytochrome P450 (CYP51), a P450 in all biological kingdoms*. DOI: 10.1016/j.bbagen.2006.07.018.

Levin, Oscar (n.d.). *Discrete Mathematics: An Open Introduction*.

Lokesh, Jep et al. (Apr. 2019). "Succession of embryonic and the intestinal bacterial communities of Atlantic salmon (Salmo salar) reveals stage-specific microbial signatures". In: *MicrobiologyOpen* 8.4, e00672. ISSN: 20458827. DOI: 10.1002/mbo3.672.

Lovell, David et al. (Mar. 2015). "Proportionality: A Valid Alternative to Correlation for Relative Data". In: *PLoS Computational Biology* 11.3. Ed. by Roland L. Dunbrack Jr., e1004075. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004075.

Lucena, Teresa et al. (Apr. 2012). "Tropicibacter multivorans sp. nov., an aerobic alphaproteobacterium isolated from surface seawater". In: *International Journal of Systematic and Evolutionary Microbiology* 62.4, pp. 844–848. ISSN: 14665026. DOI: 10.1099/ijs.0.030973-0.

Mallick, Himel et al. (2017). "Experimental design and quantitative analysis of microbial community multiomics". In: *Genome Biology* 18.1, pp. 1–16. ISSN: 1474760X. DOI: 10.1186/s13059-017-1359-z.

Martinson, Jonathan N.V. et al. (Sept. 2019). "Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults". In: *ISME Journal* 13.9, pp. 2306–2318. ISSN: 17517370. DOI: 10.1038/s41396-019-0435-7.

McHardy, Ian H. et al. (June 2013). "Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships". In: *Microbiome* 1.1. ISSN: 20492618. DOI: 10.1186/2049-2618-1-17.

Molyneaux, Philip L. et al. (June 2017). "Host-microbial interactions in idiopathic pulmonary fibrosis". In: *American Journal of Respiratory and Critical Care Medicine* 195.12, pp. 1640–1650. ISSN: 15354970. DOI: 10.1164/rccm.201607-1408OC.

Moran, Nancy A. and Daniel B. Sloan (Dec. 2015). "The Hologenome Concept: Helpful or Hollow?" In: *PLOS Biology* 13.12, e1002311. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1002311.

Mukaka M.M. (2012). "Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research". In: *Malawi Medical Journal* 24.3, pp. 69–71.

Oldham, Michael C. et al. (Nov. 2008). "Functional organization of the transcriptome in human brain". In: *Nature Neuroscience* 11.11, pp. 1271–1282. ISSN: 10976256. DOI: 10.1038/nn.2207.

Parsana, Princy et al. (Dec. 2019). "Addressing confounding artifacts in reconstruction of gene co-expression networks". In: *Genome Biology* 20.1, p. 94. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1700-9.

Patro, Rob et al. (2017). "Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference". In: *Nature Methods* 14.4, pp. 417–419. ISSN: 15487105. DOI: 10.1038/nmeth.4197.

Paulson, Joseph N et al. (2013). "Differential abundance analysis for microbial marker-gene surveys". In: *Nature Methods* 10.12, pp. 1200–1202. ISSN: 15487091. DOI: 10.1038/nmeth.2658.

Pei, G., L. Chen, and W. Zhang (2017). "WGCNA Application to Proteomic and Metabolomic Data Analysis". In: *Methods in Enzymology* 585.December 2016, pp. 135–158. ISSN: 15577988. DOI: 10.1016/bs.mie.2016.09.016.

Pereira, Mariana Buongermino et al. (2018). "Comparison of normalization methods for the analysis of metagenomic gene abundance data". In: *BMC Genomics* 19.1, p. 274. ISSN: 14712164. DOI: 10.1186/s12864-018-4637-6.

Photobacterium - an overview | ScienceDirect Topics (n.d.).

Plessis, L. du, N. Skunca, and C. Dessimoz (Nov. 2011). "The what, where, how and why of gene ontology–a primer for bioinformaticians". In: *Briefings in Bioinformatics* 12.6, pp. 723–735. ISSN: 1467-5463. DOI: 10.1093/bib/bbr002.

Quinn, Thomas P. et al. (Dec. 2017). "propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis". In: *Scientific Reports* 7.1, p. 16252. ISSN: 2045-2322. DOI: 10.1038/s41598-017-16520-0.

Quinn, Thomas P et al. (2018). "A field guide for the compositional analysis of any-omics data". In: *bioRxiv*, p. 484766. DOI: 10.1101/484766.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ravasz, E et al. (Aug. 2002). "Hierarchical Organization of Modularity in Metabolic Networks". In: *Science* 297.5586, pp. 1551–1555. ISSN: 00368075. DOI: 10.1126/science.1073374.

Robinson, Mark D and Alicia Oshlack (Mar. 2010). "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.3, R25. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-3-r25.

Rodrigues, Richard R, Natalia Shulzhenko, and Andrey Morgun (2018). "Transkingdom Networks: A Systems Biology Approach to Identify Causal Members of Host–Microbiota Interactions". In: *Methods in Molecular Biology*. Vol. 1849, pp. 227–242. DOI: 10.1007/978-1-4939-8728-3{_}15.

Rudi, Knut et al. (2018). "Stable core gut microbiota across the freshwater-to-saltwater transition for farmed Atlantic salmon". In: *Applied and Environmental Microbiology* 84.2, pp. 1–9. ISSN: 10985336. DOI: 10.1128/AEM.01974-17.

Rudman, Seth M et al. (Oct. 2019). "Microbiome composition shapes rapid genomic adaptation of Drosophila melanogaster." In: *Proceedings of the National Academy of Sciences of the United States of America* 116.40, pp. 20025–20032. ISSN: 1091-6490. DOI: 10.1073/pnas.1907787116.

Shendure, Jay (2008). "The beginning of the end for microarrays?" In: *Nature Methods* 5.7, pp. 585–587.

Shlens, Jonathon (Apr. 2014). "A Tutorial on Principal Component Analysis". In:

Sommer, Felix and Fredrik Bäckhed (Apr. 2013). "The gut microbiota — masters of host development and physiology". In: *Nature Reviews Microbiology* 11.4, pp. 227–238. ISSN: 1740-1526. DOI: 10.1038/nrmicro2974.

Soneson, Charlotte, Michael I Love, and Mark D Robinson (2016). "Open Peer Review Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]". In: DOI: 10.12688/f1000research.7563.1.

Sullam, Karen E. et al. (July 2012). "Environmental and ecological factors that shape the gut bacterial communities of fish: A meta-analysis". In: *Molecular Ecology* 21.13, pp. 3363–3378. ISSN: 09621083. DOI: 10.1111/j.1365-294X.2012.05552.x.

Tailford, Louise E. et al. (2015). "Mucin glycan foraging in the human gut microbiome". In: *Frontiers in Genetics* 5.FEB. ISSN: 16648021. DOI: 10.3389/fgene.2015.00081.

The Gene Ontology Consortium (Jan. 2019). "The Gene Ontology Resource: 20 years and still GOing strong". In: *Nucleic Acids Research* 47.D1, pp. D330–D338. ISSN: 0305-1048. DOI: 10.1093/nar/gky1055.

The UniProt Consortium (Jan. 2019). "UniProt: a worldwide hub of protein knowledge". In: *Nucleic Acids Research* 47.Database issue, pp. D506–D515. ISSN: 0305-1048. DOI: 10.1093/nar/gky1049.

Tourlousse, Dieter M. et al. (Dec. 2016). "Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing". In: *Nucleic Acids Research*, gkw984. ISSN: 0305-1048. DOI: 10.1093/nar/gkw984.

UniProt (2015). *Protein existence*.

Venturelli, Ophelia S et al. (June 2018). "Deciphering microbial interactions in synthetic human gut microbiome communities." In: *Molecular systems biology* 14.6, e8157. ISSN: 1744-4292. DOI: 10.15252/MSB.20178157.

Vogel, Christine and Edward M. Marcotte (2012). "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses". In: *Nature Reviews Genetics* 13.4, pp. 227–232. ISSN: 14710056. DOI: 10.1038/nrg3185.

Voineagu, Irina et al. (2011). "Transcriptomic analysis of autistic brain reveals convergent molecular pathology". In: *Nature* 474.7351, pp. 380–386. ISSN: 00280836. DOI: 10.1038/nature10110.

Wagner, Günter P., Mihaela Pavlicev, and James M. Cheverud (Dec. 2007). "The road to modularity". In: *Nature Reviews Genetics* 8.12, pp. 921–931. ISSN: 1471-0056. DOI: 10.1038/nrg2267.

Walter, Wencke, Fátima Sánchez-Cabo, and Mercedes Ricote (Sept. 2015). "GOplot: an R package for visually combining expression data with functional analysis". In: *Bioinformatics* 31.17, pp. 2912–2914. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv300.

Wang, Zhong, Mark Gerstein, and Michael Snyder (Jan. 2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484.

Weisburg, W G et al. (Jan. 1991). "16S ribosomal DNA amplification for phylogenetic study". In: *Journal of Bacteriology* 173.2, pp. 697–703. ISSN: 00219193. DOI: 10.1128/jb.173.2.697-703.1991.

Wickham, Hadley (2016). *ggplot2 : elegant graphics for data analysis*, p. 212. ISBN: 9780387981406.

Woo, P.C.Y. et al. (Oct. 2008). "Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories". In: *Clinical Microbiology and Infection* 14.10, pp. 908–934. ISSN: 1198743X. DOI: 10.1111/j.1469-0691.2008.02070.x.

Xue, Jia et al. (Feb. 2014). "Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation". In: *Immunity* 40.2, pp. 274–288. ISSN: 10747613. DOI: 10.1016/j.immuni.2014.01.006.

Yip, Andy M and Steve Horvath (Dec. 2007). "Gene network interconnectedness and the generalized topological overlap measure". In: *BMC Bioinformatics* 8.1, p. 22. ISSN: 14712105. DOI: 10.1186/1471-2105-8-22.

Zhang, Bin et al. (2005). "A General Framework for Weighted Gene Co-Expression Network Analysis". In: *Statistical Applications in Genetics and Molecular Biology* 4.1.

Zilber-Rosenberg, Ilana and Eugene Rosenberg (Aug. 2008). "Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution". In: *FEMS Microbiology Reviews* 32.5, pp. 723–735. ISSN: 1574-6976. DOI: 10.1111/j.1574-6976.2008.00123.x.

Appendices

.1 Bash script

```
#!/bin/bash
#SBATCH --ntasks=1                # Number of CPUs
#SBATCH --nodes=1                 # Number of nodes
#SBATCH --job-name="name_of_job" # Name for the job
#SBATCH --partition=verysmallmem # verysamellmem < 10 GB RAM, smallmem 10-150GB
#SBATCH --mem=4G                  # Expected memory usage


module purge
module load anaconda3 # Has pandoc installed
module load R/3.5.0 # Correct R version
module list

Rscript -e "Sys.setenv(RSTUDIO_PANDOC='/usr/lib/rstudio-server/bin/pandoc'); \
rmarkdown::render('rna_makeModules_template.Rmd', \
output_file='rna_makeModules_template.html')"
```

.2 Outlier detection

.3 Lineplots

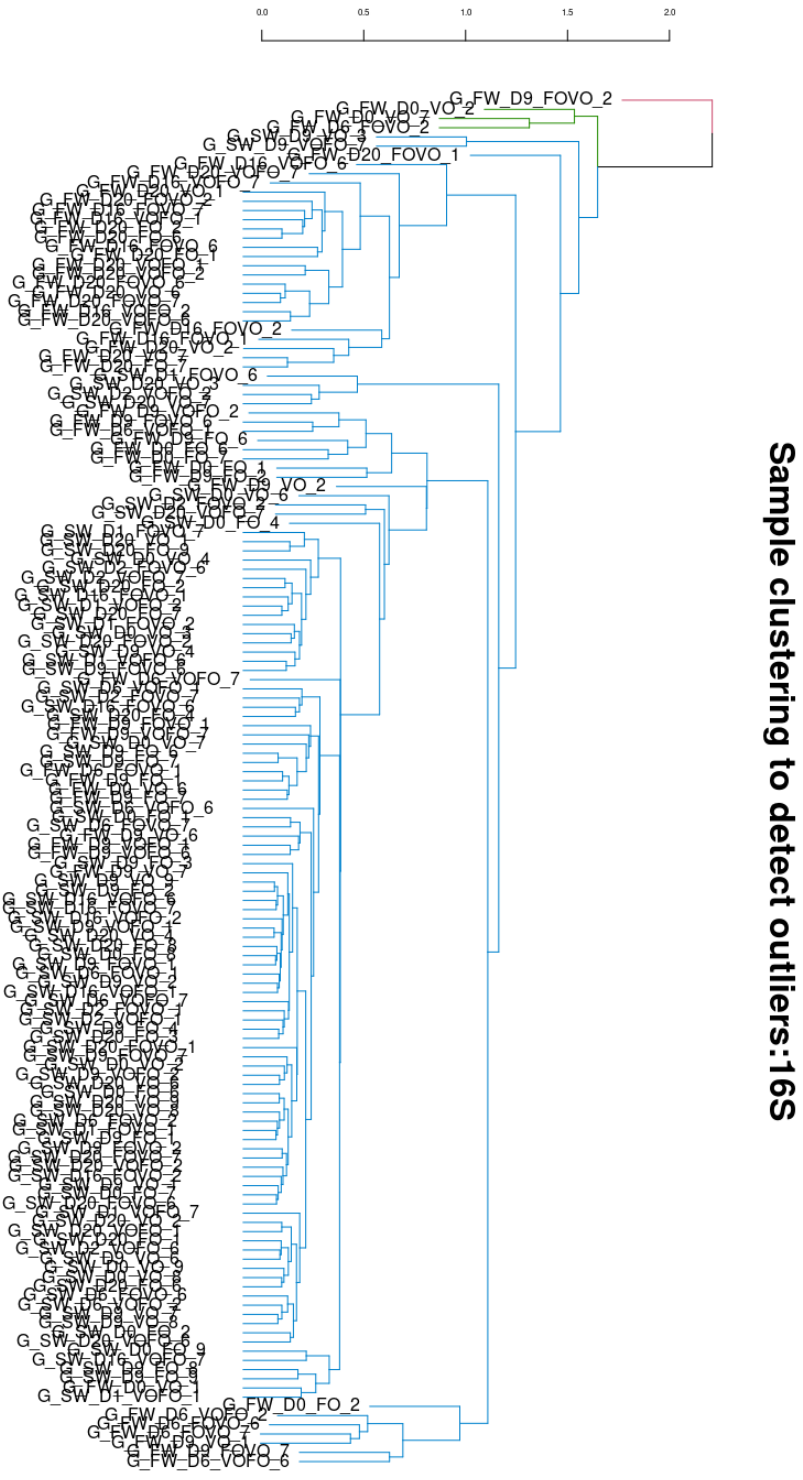

Figure .2: Sample clustering of the 16S metagenomics data. Notice that one sample, G_FW_D9_FOVO_2, is different from all other 16S samples.

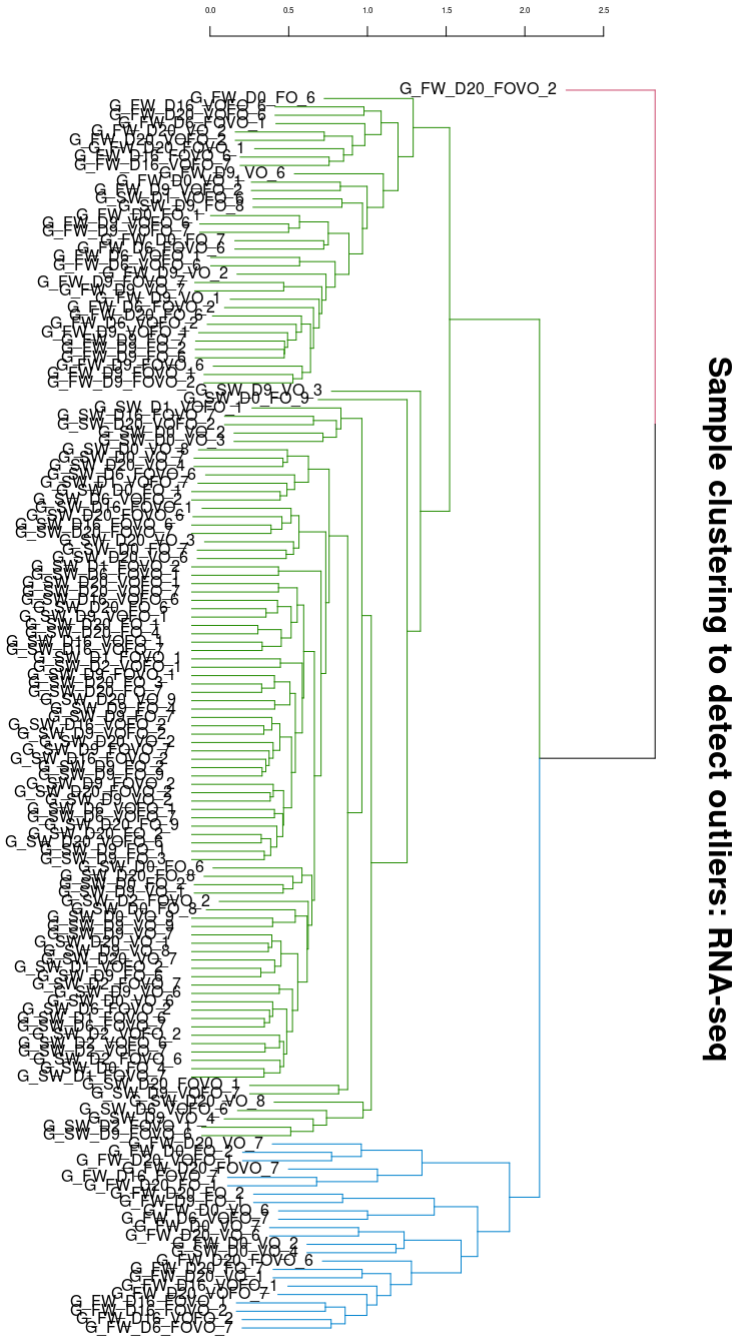

Figure .3: Sample clustering of the RNA-seq transcriptomics data. Notice that one sample, G_FW_D20_FOVO_2, is different from all other 16S samples.

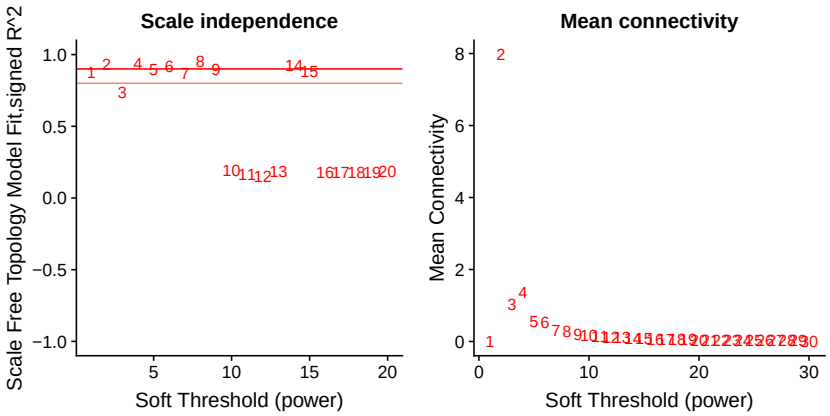

Figure .4: Scale free independence and mean connectivity plot. Use of ρ_p as a association measure on saltwater samples.

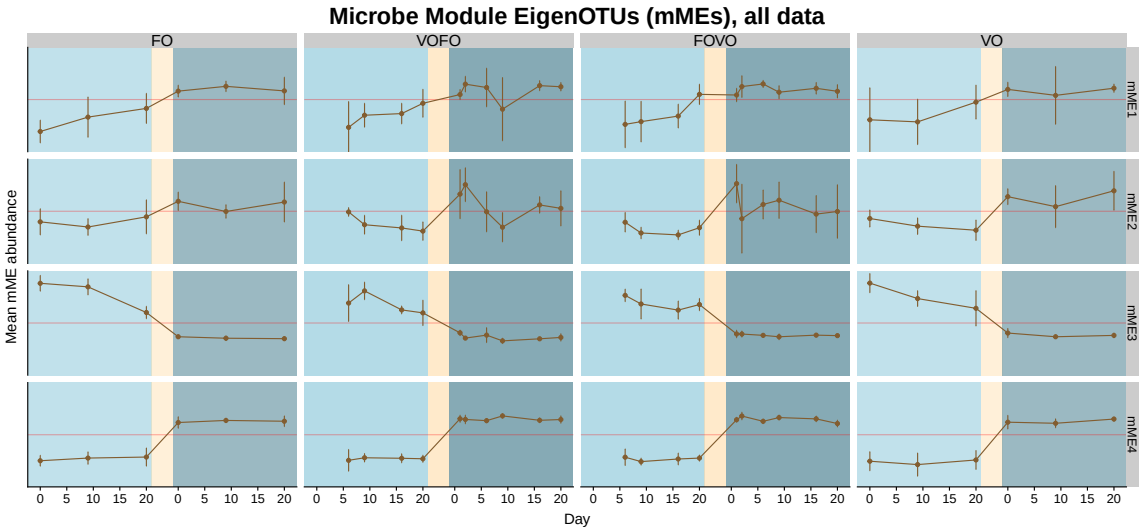

Figure .5

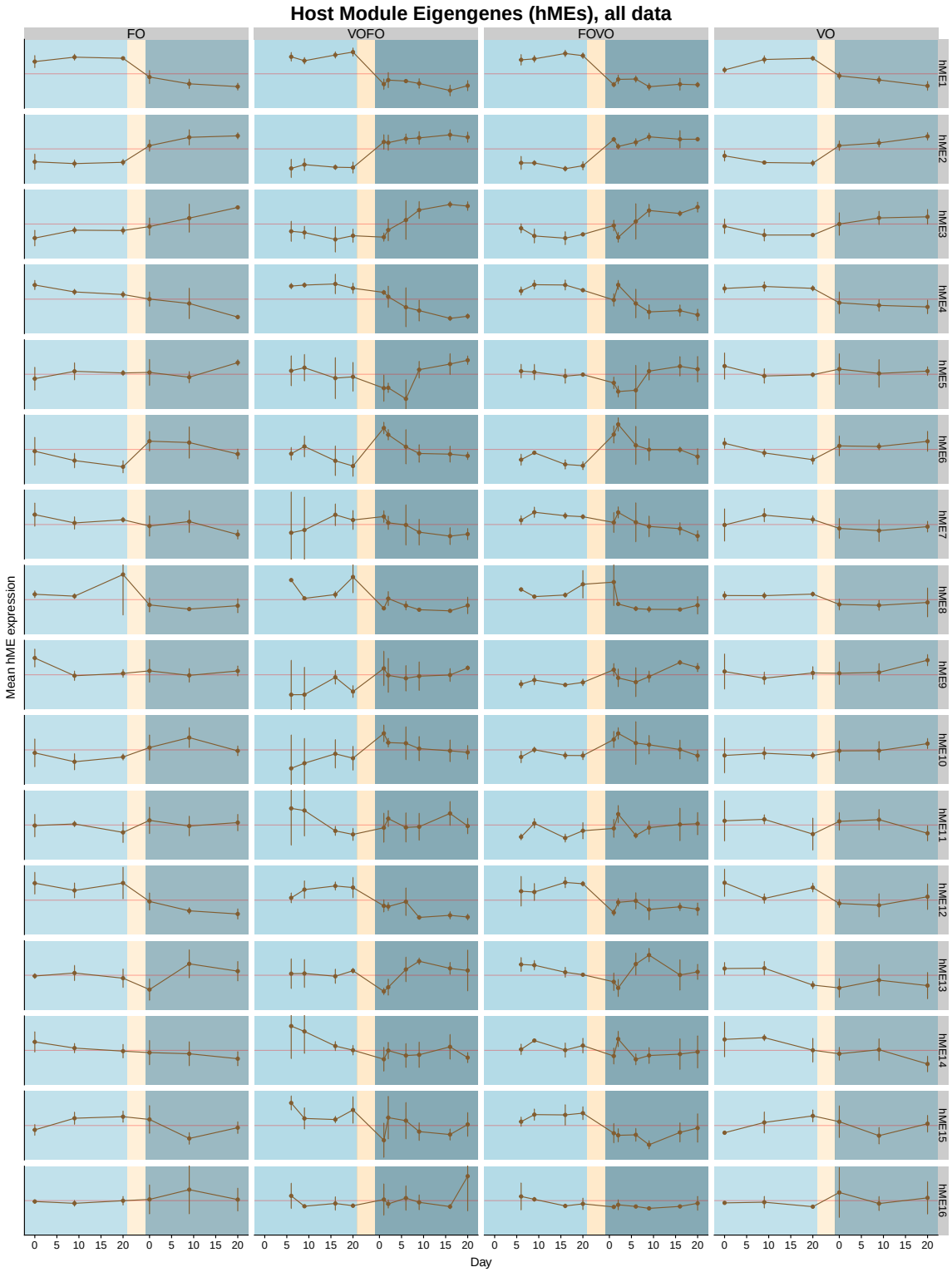
Host Module Eigengenes (hMEs), all data
FO
VOFO
FOVO
VO
hME1
hME2
hME3
hME4
hME5
hME6
hME7
hME8
hME9
hME10
hME11
hME12
hME13
hME14
hME15
hME16
Mean hME expression
0 5 10 15 20
Day

Figure .6

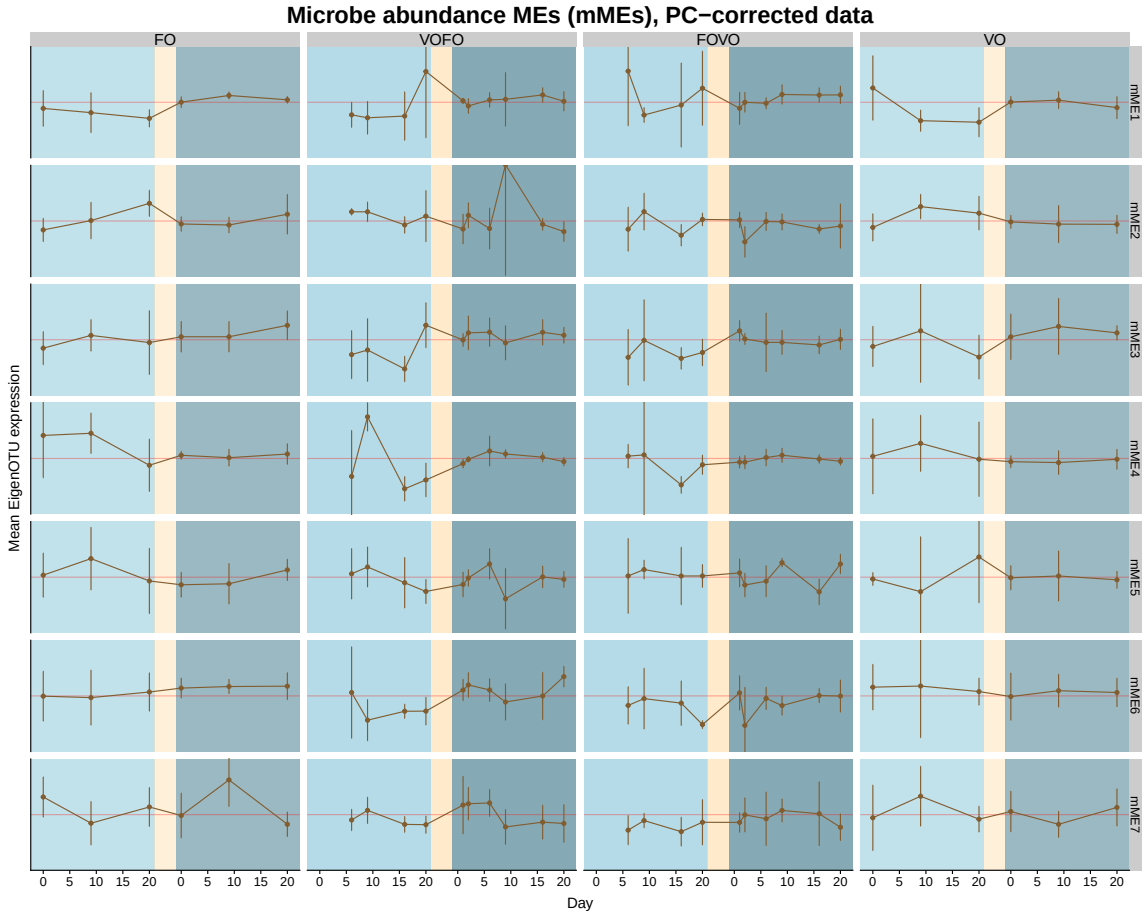
Microbe abundance MEs (mMEs), PC−corrected data
FO
VOFO
FOVO
VO
mME1
mME2
mME3
mME4
mME5
mME6
mME7
Mean EigenOTU expression
0 5 10 15 20
Day

Figure .7

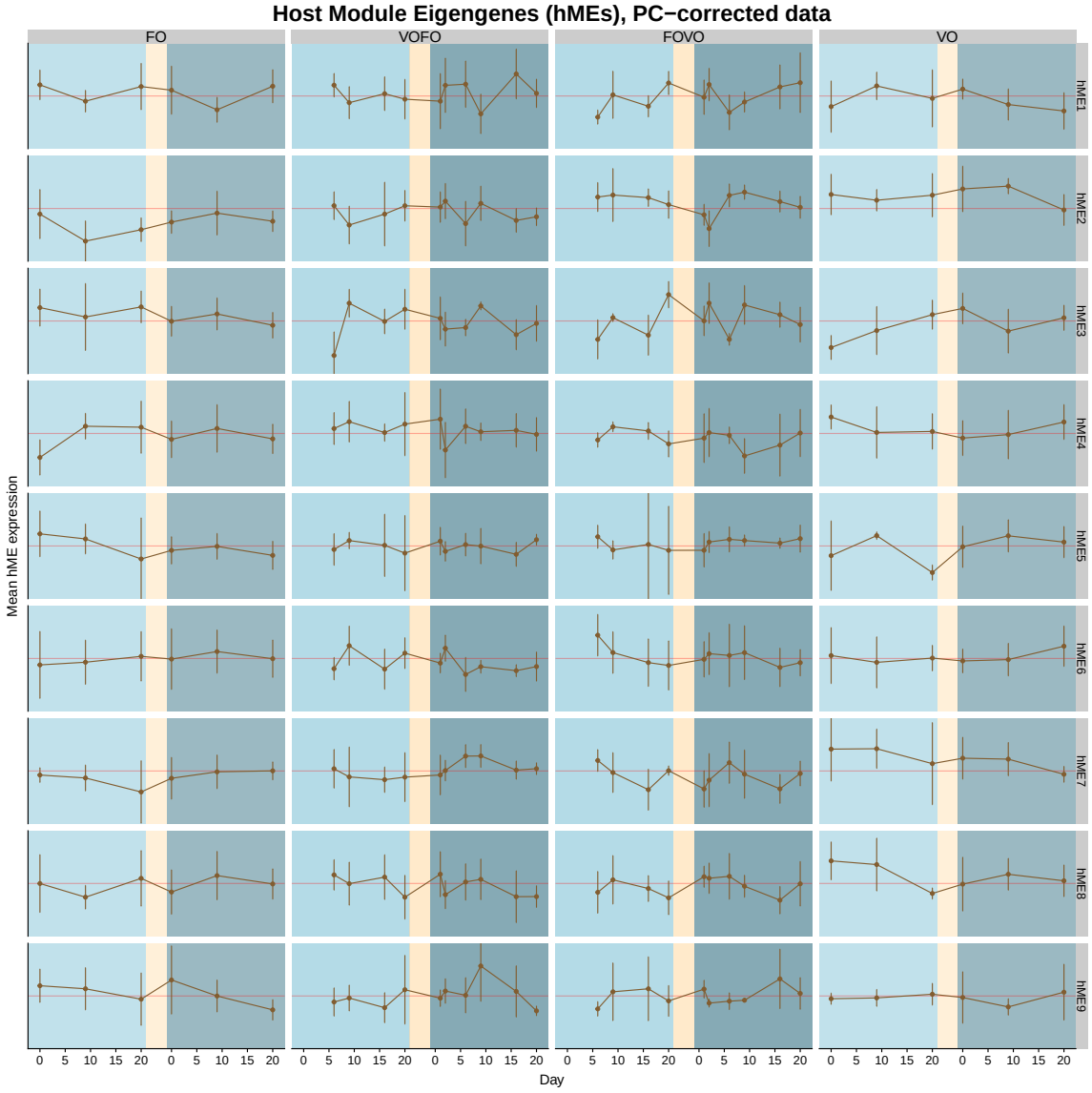
Host Module Eigengenes (hMEs), PC−corrected data
FO
VOFO
FOVO
VO
hME1
hME2
hME3
hME4
hME5
hME6
hME7
hME8
hME9
Mean hME expression
0 5 10 15 20
Day

Figure .8