Hepatic Proteome Analysis of Fathead Minnow *(Pimephales promelas)* Exposed to Municipal Wastewater in the Bow River

by

Mark George Lubberts

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2019

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Mark Lubberts is the sole author of Chapters 1, 2, and 4, which were written under the supervision of Brendan McConkey.

Chapter 3 of this thesis is part of a larger study on the effects on MWWE on the Bow River. Funding was secured by Dr. M. Vijayan of the University of Calgary and Dr. B. McConkey at the University of Waterloo. Dr. Vijayan and Ms. Lazara-Côté designed the caging study. Ms. Lazaro-Côté performed the caging study and tested the water quality and contaminant levels. Mark Lubberts performed the protein extraction and proteomics works, analyzed the data, and wrote the chapter.

Abstract

Large quantities of contaminants enter Canadian waterways through municipal wastewater effluent (MWWE) each year. MWWE exposure can have a myriad of effects on aquatic organisms, including inducing chronic stress response or altering hormone signaling. A key site for many of these effects is the liver, which controls xenobiotic metabolism, energy regulation, and other functions related to the maintenance of homeostasis. One method for analyzing changes in liver function is shotgun proteomics, which uses mass spectrometry data to identify and quantify proteins in a tissue or sample. The objective of this thesis is to investigate changes in the liver proteome in fathead minnows caged upstream and downstream of wastewater treatment plants (WWTPs) in the Bow River, Alberta.

To effectively analyze mass spectrometry data generated for this project, a fathead minnow specific protein database was constructed from the draft fathead minnow genome and corresponding annotations. The constructed database is compared to the zebrafish reference proteome and UniProt Cyprinidae proteins to determine differences between databases and the effect on protein identification rates. Additionally, several different search engines, including *de novo* and database search engines, are compared against different datasets to determine how mass spectrometry equipment and database accuracy affects protein identification rates. Comparison of the databases showed that a species-specific database provided substantial increases in protein identification rates, with 461 (14.2%) more proteins identified than from the zebrafish reference proteome and Uniprot database. Search tool comparisons revealed that while *de novo* search engines can increase protein identifications within low quality databases, they are outperformed by standard database search engines when an accurate and comprehensive database is available.

Shotgun proteomics was used to analyze changes in liver proteome of fathead minnows caged in five sites along the Bow River around Calgary, Alberta, Canada. Sites were located upstream, downstream, and close to the outflow of wastewater treatment plants (WWTPs) along the river. The constructed database and a TMT-labelled fathead minnow mass spectrometry dataset are used to quantify protein expression changes in the proteome. Differential expression analysis and gene set enrichment analysis are used to compare expression profiles among different sites. 3689 proteins were identified in the fathead

minnow proteome. Differential expression shows large changes in the liver proteomes of fish located near the outflow of the Bonnybrook WWTPs, with a similar, but reduced effect at sites further. Proteins and gene sets in involved in lipid metabolism, oxidative stress, xenobiotic removal were upregulated, while mRNA splicing, and cytoskeleton organization were downregulated. Fish near the outflow also showed changes in cell cycle control and protein modifications compared to fish further downstream.

Identifying changes in the liver proteome of MWWE-exposed organisms is important for identifying potential biomarkers and understanding how exposure affect the health of aquatic organisms. Changes in peroxisomal lipid metabolism and mitochondrial proteins suggest mitochondrial activity for further study of the impacts of MWWE. Additionally, analysis of the impacts of species-specific database on shotgun proteomics is useful for the application of omics to non-model organisms. Comparison of the different database shows that even first draft genome sequences can produced substantially improved protein identification rates over protein databases derived from other organisms.

Acknowledgements

Like all scientific endeavours, this work was not performed in a vacuum, and I owe a great debt to many people for theirs support while I completed it. While I do not have room to thank all of them, a few people deserve specific thanks:

First, I would like to thank Dr. Brendan McConkey, for his advice and support, for providing me with the opportunity to perform this work, and refusing to let me title this work 'What Fish Do: Liver Boogaloo'. I would also like to thank my committee members, Dr. Mark Servos and Dr. Kun Liang, for sharing their advice and knowledge.

I would like to thank the other members of the McConkey Lab - Janet Lorv, Karsten Rinas, and Monica Gromula - for their continual support, insight, and advice. It was a joy to work with them, and I could not wish for better labmates.

Finally, I would like to thank Nardo Nava Rodriguez. Without your continued support, this thesis could not have been finished.

Dedication

To my father.

The author (center) performing one of his first experiments with his father (left).

Table of Contents

*

List of Figures

List of Tables

Code Listings

Abbreviations

WWTP wastewater treatment plant

MWWE municipal wastewater effluent

PPCP pharmaceutical and personal care product

PPAR peroxisome proliferator-activated receptor

FA fatty acid

FAO fatty acid oxidation

TCA tricarboxylic acid cycle

CPT1 carnitine palmitoyltransferase 1

ACOX1 acyl-CoA oxidase 1

H_2O_2 hydrogen peroxide

CROT carnitine O-octanyltransferase

PPRE peroxisome proliferator response element

CYP cytochrome P450

AHR aryl hydrocarbon receptor

ROS reactive oxygen species

PEPCK phosphenolpyruvate carboxykinase

HPI hypothalamus-pituitary-interrenal

GR glucocorticoid receptor

SOCS suppressors of cytokine signaling

JAKS-STAT Janus kinase-Signal Transducer and Activator of Transcription

FHM fathead minnow

MS mass spectrometry

ESI electrospray ionization

m/z	mass-to-charge ratio
MALDI	matrix assisted laser desorption ionization
LC-MS	liquid chromatography mass spectrometry
TOF	time-of-flight
SILAC	stable isotope labeling with amino acids in cell culture
TMT	Tandem Mass Tags
SETAC	Society for Environmental Toxicology and Chemistry
PSM	peptide-spectrum match
FDR	false discovery rate
ORF	open reading frame
FHMP	Fathead Minnow Predicted Proteome
U90CYP	UniRef90 Cyprinidae Proteome
ZRP	Uniprot Zebrafish Reference Proteome
BUSCO	Benchmarking Universal Single Copy Orthologs
DOGMA	DOmain-based General Measure for transcriptome and proteome quality Assessment
CDA	conserved domain arrangement
PTM	post-translational modification
TPP	Trans-Proteomic Pipeline
VSN	variance stabilized normalization
MDS	multidimensional scaling
GSEA	Gene Set Enrichment Analysis
DE	differentially expressed
GST	glutathione S-tranferase
GSH	reduced glutathione
KEGG	Kyoto Encyclopedia of Genes and Genomes
STRING	Search Tool for Retrieval of Interacting Genes/Proteins

Chapter 1

Introduction

1.1 Municipal Wastewater Effluent

1.1.1 Contaminants in MWWE

Municipal Wastewater Effluent (MWWE) is a key source of contaminants in aquatic environments. More that 86% of Canada's population is served by wastewater treatment plants (WWTPs), which treat an estimated 300–500 L of waste water per capita daily.[1–3] After varying degrees of treatment, that wastewater is then discharged in nearby rivers, lakes, and oceans, and often migrates into groundwater as well.[4] While wastewater treatment methods are generally effective at reducing carbon, nitrogen, and microbial loads, concentration of other contaminants can be unchanged or even increased by different wastewater treatment methods.[5,6] The wide range of chemicals found in WWTP effluent is a result of the vast number of pharmaceutical and personal care products (PPCPs), industrial chemicals, household wastes, and pesticides that enter the wastewater system.[4,7,8] A study of two WWTPs in the United Kingdom investigating 55 different PPCPs; including antibiotics, anti-inflammatory agents, antidepressants, and lipid regulating agents, over a 5 month period found an average load of 6 to 10 kg/d in the influent, effluent, and receiving waters of the plants.[6]

Besides PPCPs, a wide range of other contaminants are present in MWWE, including heavy metals, persistent organic pollutants, pesticides, perfluoroalkyl acids, artificial sweeteners, industrial waste, and chlorination and disinfectant byproducts.[4,7,8] These contaminants, along with PPCPs, often remain in detectable concentrations in MWWE, even after treatment, and are detectable in downstream surface waters.[2–4,7,9,10] Removal efficiencies vary drastically due to interacting effects of contaminant type and treatment methods on the rate of degradation and sorption. Additionally, influent characteristics and

treatment efficiencies change over time, both on a daily and seasonal scale.[4,6,8] Observation of more than 20 PPCPs in various wastewater treatment plant along the Bow River in Calgary, Alberta found detectable concentrations in effluent and surface waters,[11] and many of those contaminants were later detected kilometers downstream of the nearest WWTP.[12]

Thus, MWWE is the primary source of hundreds of different anthropogenic chemical compounds found in aquatic environments at various concentrations.[1] Despite trends towards increased treatment, it remains an important source of contamination of aquatic environments.[2] Understanding the impact of MWWE on aquatic environments will be an important factor in improving wastewater treatment methods.

1.1.2 Treatment of MWWE

Wastewater treatment is generally classified into three different categories, depending on the method of processing and type of waste targeted for removal. The first and most common level of treatment, primary treatment, uses sedimentation processes to remove the majority of suspended solids and reduce biological oxygen demand.[1,5] Primary treatment is generally found to be minimally effective at reducing PPCP contaminant levels, as they remain in the aqueous phase rather than settling into the sediment, though some highly adsorbed chemicals may be removed.[5]

Secondary treatment uses methods such as activated sludge or trickling filter beds to remove organic material with aerobic or anaerobic microorganisms.[1,6] Secondary treatment is where the bulk of easily-degraded compounds, such as acetaminophen, are removed, either by biodegradation or abiotic chemical processes.[4,5] Compounds which do not biodegrade easily, such as carbamazepine, will pass through secondary treatment largely unchanged, or in some cases may even be regenerated by deconjugation of metabolites present in the wastewater.[4,6,13]

Tertiary treatments are used for a variety of purposes, such as targeted removal of contaminants, or disinfection of effluent by ultraviolet light, chlorination, or other methods. The impact of tertiary treatment varies with different treatment methods and the type of contaminants present.[1,2,13] While wastewater treatment methods have improved over the last two decades, MWWE still has significant impact of the health of downstream environments, especially with the variety of emerging contaminants like PPCPs being introduced into municipal wastewater.[2,14]

1.1.3 Impacts of MWWE

The properties of PPCPs makes them inherently concerning as a contaminant. Pharmaceuticals are a primary methods of patient care in modern medicine; commonly used pharmaceuticals are specifically designed to be chemically stable and active in low doses, as they need to persist in the body long enough

to have effect in the face of physiological elimination mechanisms.[6,13] For example, previous studies of antidepressants such as venlafaxine have shown that over 30% of the treatment dose may be excreted directly or as a biologically active metabolite.[3] The lipid-lowering drug bezafibrate is excreted as more than 50% active compound, and for pain management drugs, such as codeine or acetaminophen, more than 70% of the dose may be excreted unchanged.[6] The large volume of PPCPs can result in up to 3 kg/d of PPCPs entering the receiving waters of a single plant serving 110,000 people, as is the case in the previously mentioned study in the UK.[6] Generally these compounds will enter aquatic environments in large quantities. Even if these contaminants are not toxic, or are found below toxic concentrations, exposure to PPCPs can have chronic and sub-lethal effects, causing the dynamics of receiving water ecosystems to also be significantly altered.[1,2] Additionally, contaminants can have complicated interactive effects - multiple chemicals may cause similar or overlapping disruptions in organisms, compounding effects beyond what would be observed a similar level of an individual contaminant.[2,3,9,15] For example, exposure to a combination of carbamazepine and clofibric acid has been shown to be more dangerous than either alone[13], and gemfibrozil, in addition to altering lipid metabolism, is an CYP450 inhibitor that may reduce metabolism of other drugs, increasing the effects of exposure.[16] Other factors, such as total dissolved oxygen and nutrient levels, may also play a role in determining the toxicity and impact of particular contaminants.[1,2] However, toxicity and thresholds of effect are usually determined for individual compounds[13], and studies of sub-lethal effects have generally focused on single compounds as well, limiting their use for understanding the effects of complex mixtures.[12]

Studies of both general MWWE exposure and particular contaminants have identified a wide range of impacts on aquatic organisms. Some contaminants, such as the pesticide dichlorvos, are of concern due to their ability to 'directly' damage the exposed organism, inducing injury or death.[17] Other contaminants, pharmaceuticals in particular, are intended to alter biological pathways in humans. These pathways that are often conserved in fish, leading to changes in behavior, metabolism, or other functions.[18] Thus, while exposure at environmentally relevant concentrations may not be toxic *per se*, PPCPs can impact a variety of pathways that may reduce the fitness of a population.

For example, exposure to estrogenic compounds such as 17α-ethinylestradiol commonly found in MWWE induces expression of the egg yolk protein vitellogenin and formation of oocytes in testis, and can alter population sex ratios.[14,18,19] A study using trout transferred from upstream controls sites to a MWWE-exposed environment showed altered spawning patterns and increased vitellogenin in male fish.[20] These changes impact the ability of individuals to reproduce, reducing the overall population over time.

Exposure to PPCPs can also alter fish behavior and responses to environmental stressors. Experiments with selective serotonin reuptake inhibitors has shown reduced territorial aggression, prey capture, and feeding behavior in exposed fish.[18] Caging studies in MWWE outflows has also shown changes in cortisol,

the primary hormone involved in stress response and maintenance of homeostasis. Impacted fish are have reduced ability to feed themselves, affecting growth and long term survival.

MWWE can also alter normal biological function within an organism. Many contaminant are known to interact with the peroxisome proliferator-activated receptor (PPAR) transcription factors[10,21], which control fatty acid metabolism in many different tissues including heart, liver, muscle, and fatty tissue.[22] These can result in changes in glycogen and lipid content in the organism, as well as alteration of enzymes and pathway governing glucose and fatty acid metabolism.[9,20]

These types of changes in the exposed organism can impact fitness and reproductive success and can reduce population levels of affected population. Thus, understanding the impacts of complex and variable MWWE mixtures on aquatic environments requires analysis of a broad range of potential pathways that may be impacted. Ideally, better understanding the impacts of MWWE would allow for identification of various biomarkers that identify sub-lethal effect of exposure.

1.2 Liver Function in Teleosts

The diverse potential effects of MWWE allow for a variety of approaches for studying the effects of exposure. One tissue with a large number of potentially impacted functions in fish is the liver. Many contaminants are hydrophobic and accumulate in lipids and fatty tissues,[10] and the liver accounts for 10-20% or more of total body lipids in fish.[21] The liver also regulates a large number of key pathways for xenobiotic metabolism, lipid metabolism, and hormone signaling in teleosts. Contaminants that disrupt or alter liver function are common in MWWE,[18] and can have significant impact on organism health and fitness.[9]

1.2.1 Lipid Metabolism and Energy Regulation

The liver is a key site of lipid metabolism and gluconeogenesis. Unlike mammals, fish utilize lipids and proteins as energy source better than carbohydrates.[23] β-oxidation, or fatty acid oxidation (FAO) is a process that occurs in both peroxisomes and mitochondria, during which fatty acids are transformed in acyl-CoA and successively shortened by 2 carbons atoms, producing acetyl-CoA and a shorter acyl-CoA.[24] The acetyl-CoA derived from this process enters the tricarboxylic acid cycle (TCA) in the mitochondria for production of ATP and the regeneration of NADH and FADH2 through the oxidative phosphorylation pathway. Alternatively, if excess acetyl-CoA is available, TCA intermediates will be exported from the mitochondria for the generation of glucose through gluconeogenesis.[10]

In mitochondria, β-oxidation of long, medium, and short chain fatty acid (FA) is initiated by acyl-CoA

dehydrogenase (ACAD), and the electrons produced are used in oxidative phosphorylation for energy generation.[24] Mitochondrial FAO is primarily limited to dietary FAs, and transport across the mitochondrial membrane is a rate limiting step.[10] While short and medium chain FA can enter the mitochondria unmodified, long chain FAs must be transported as acylcarnitine intermediates through the carnitine shuttle. carnitine palmitoyltransferase 1 (CPT1) regulates mitochondrial β-oxidation by controlling the conversion to acylcarnitine at the mitochondrial membrane.[10]

β-oxidation in the peroxisome is primarily limited to long and very-long chain FAs. While similar to mitochondrial β-oxidation, electrons are not used in oxidative phosphorylation, and the initiating enzyme, acyl-CoA oxidase 1 (ACOX1) instead produces H_2O_2 as a by-product during each round of oxidation.[24] β-oxidation in peroxisomes does not proceed to complete shortening of the chain into acetyl-CoA as it does in the mitochondria. Instead, carnitine O-octanyltransferase (CROT) converts shortened acyl-CoAs to acylcarnitine intermediates for transfer to the mitochondria for complete β-oxidation. Peroxisomes also contain modified and complimentary enzymes that allow for the oxidation of a wide range of other compounds containing hydroxyl, methyl, and other large modifications, such as steroids or branched FAs and FA derived signaling molecules like prostanoids.[24,25] Thus FAO in peroxisomes plays a role in the maintenance of signaling and regulatory FAs.

A variety of regulatory factors for FA metabolism exist. Acetyl-CoA in the cytosol, in addition to being incorporated into the TCA or used in gluconeogenesis, may be carboxylated by fatty acide synthase into the building block of fatty acid, malonyl-CoA. Malonyl-CoA also inhibits CPT1 activity, preventing entry of long-chain FAs into the mitochondria for β-oxidation and favoring the synthesis and extension of FAs for energy storage and export.

PPARs, occur in vertebrates as a family of three transcription factors that are key regulators of peroxisomal β-oxidation. PPARs function as nuclear receptors that dimerize with the retinoid X receptor (RXR) and together bind the peroxisome proliferator response element (PPRE). PPARs are activated by a variety of FAs, FA metabolites, and eicoseinoids, and regulate the expression of genes involved in FA metabolism, glucose utilization, lipoprotein metabolism, lipogenesis, as well as inflammation and a variety of other effects. In particular PPARα is highly expressed in liver and controls many FA metabolism genes. Activation of PPARα increases the size and number of peroxisomes in liver tissue, as well as the rate of β-oxidation.

1.2.2 Xenobiotic Transformation

Cytochrome P450s (CYPs) are a superfamily of heme-containing enzymes predominantly expressed in the liver and catalyze oxidoreductase reactions, usually acting as a mono-oxygenase to incorporate hydroxyl

groups into their substrate.[26] CYPs play a role in metabolism of a large number of endogenous and exogenous compounds and are important for the biotransformation of xenobiotics into water-soluble forms for excretion.[27]

CYPs are broadly distributed in all eukaryotes, and 19 families of CYP genes have been identified in vertebrates.[28] CYP genes in families 5-51 are considered to be involved in the metabolism of endogenous compounds, having higher substrate specificity and greater phylogenetically stability than families 1-4. They have generally consistent distribution, frequency, and function across Vertebrata, most being part of the metabolism of steroids, sterols, and signaling molecules.[29] CYP51, for example, is found is all biological kingdoms, where it is part of the sterol biosynthetic pathway.[30]

CYP families 1-4 play a prominent in role in xenobiotic response and metabolism.[28] These families are much more diverse than genes in families 5-51, and contain the majority of teleost-specific subfamilies.[28,29] Substrate specificity is generally broader and the lack of similarity across organisms means the function of these CYPs is less well understood compared to more conserved CYP families.[31]

The CYP1 family has been well studied as a biomarker, as it is known to respond to a wide variety of contaminants including dioxins and polyaromatic hydrocarbons which activate the aryl hydrocarbon receptor (AHR). A primary role of the CYP1 family is to oxidize carcinogenic compounds, increasing the rate of excretion.[26]

CYP family 2 is the most diverse, with multiple lineage specific subfamilies and considerable functional divergence across species.[29] While the function of CYP2 isozymes are poorly understood, many play a role in metabolism of both endogenous and exogenous substrates.[29] CYP2K, for example, metabolizes aflatoxin B_1 in both zebrafish and rainbow trout, but metabolizes lauric acid only in rainbow trout.[28] Many CYP2 genes are located at conserved loci and have undergone diversification and expansion in different organisms. The human CYP2J2 gene is located in a syntenic region with 11 CYP2 genes in zebrafish and nine in mangrove killifish.[29,31] Several of the CYP2 proteins found in this region oxidize arachidonic acid, as does CYP2J2,[29] and expression of the different CYP2 genes varies across tissue, developmental stage, and xenobiotic exposure,[31] suggesting diversification of CYP2 in response to specific metabolic needs.

CYP2 gene regulation in teleosts is largely unknown; the genes do not respond to AHR stimulation, and known mammalian CYP2 inducers such as phenobarbitol do not induce CYP2 in teleosts.[26,28,31] The CYP2 family likely important in responding to a range of xenobiotics that varies with the environmental conditions of the organism.

CYP families 3 and 4 are smaller and less diverse than family 2. CYP3 has 3 subfamilies, one of which is only found in teleosts. CYP3A is found in both teleosts and mammals and is important for testosterone and progesterone metabolism, and responds to various xenobiotics such a rifampicin and dexamethasone, while

the functions of CYP3B and C subfamilies are not well understood.[26,29] CYP4 genes are more common in mammals than fish and are thought to be mostly FA hydroxylases that play a role in metabolism of various signaling molecules. CYP4 expression is affected by exposure to fibrate drugs, and are potentially under the control of the PPAR transcription factors, as they respond to xenobiotics known to alter PPAR activity.[26,29]

1.2.3 Reactive Oxygen Species Metabolism

One byproduct of both FAO and the oxidative reactions performed by CYPs and other xenobiotic biotransformation enzymes is an increase in reactive oxygen species (ROS). Peroxisomes and mitochondria are the primary source of endogenous ROS.[24] While a basal level of ROS is necessary for normal cell functioning, particularly for cell-signaling through disulfide bond formation and breaking, excess ROS is damaging to RNA, ribosome, lipid membranes, and cellular protein. Generally, cells attempt to convert the oxygen radicals $O_2^{\cdot-}$ to H_2O_2 with superoxide dismutase to prevent oxidative damage by ROS. Enzymes to convert H_2O_2 to water, particularly catalase, peroxiredoxin, and glutathione peroxidase, then maintain H_2O_2 at the necessary levels.[32] However, increased production of ROS by β-oxidation or xenobiotic metabolism can overwhelm the system. Glutathione peroxidase, for example, relies on a supply of monomeric glutathione both for reducing H_2O_2 and biotransformation of lipophilic xenobiotics for excretion from the organism. PPARα activation has been shown to increase H_2O_2 production from ACOX1 and CYP4, but also reduces cellular levels of catalase, potentially leading to oxidative stress.[24]

1.2.4 Stress Response

A key pathway during stress response is the hypothalamus-pituitary-interrenal (HPI) axis, which mediates the release of corticosteroids such as cortisol, the primary corticosteroid involved in stress response in teleosts.[33–35] After synthesis, cortisol is released from interrenal cells in the head kidney into the plasma, where is circulates through the body. Cortisol is lipid soluble and thought to passively diffuse into cells, though additional mechanisms of transport have been suggested.[36] Once in the cytosol, cortisol binds to the glucocorticoid receptor (GR) transcription factor. GR is then translocated to the nucleus where it interacts with the DNA glucocorticoid-response elements to upregulate transcription.[33,36,37]

Cortisol is involved in regulating a variety of processes in hepatocytes as part of the stress response, including gluconeogenesis and transcription of a wide range of genes.[33,36]. Studies suggest the primary purpose of gene regulation through a GR within the liver is to control energy demand.[36,38] In particular, cortisol upregulates expression of phosphenolpyruvate carboxykinase (PEPCK), a key enzyme in gluconeogenesis,[36,37] as well as suppressors of cytokine signaling (SOCS) genes, which suppress the Janus

kinase-Signal Transducer and Activator of Transcription (JAKS-STAT).[36] JAKS-STAT signaling mediates energy demanding growth and immune pathways through expression of genes such as Insulin-like Growth Factor and cytokine-mediated lipopolysaccharide-stimulated immune response.[36,39] Additionally, cortisol exposure has been shown to decrease GR levels through protein degradation while increasing transcription rates of the GR gene.[36]

Non-genomic signaling has also been observed to occur in hepatocytes in response to cortisol signaling in trout. Changing the fluidity and topology of plasma membrane of liver cells, induces phosphorylation of proteins kinases A and C and Akt, which are involved in signaling pathways regulating glucose metabolism,[34] as well as the transcription factor cAMP response element-binding protein.[36] Similar changes in membrane were not observed when cortisol synthesis was blocked.[40]

1.2.5 MWWE and Liver

Exposure to MWWE is known to have a variety of impacts liver-related functions in exposed fish. Exposure studies have shown that fibrates and pharmaceuticals designed to alter FA metabolism in humans also increases FA β-oxidation in fish.[18,21] These pharmaceuticals alter PPAR activity in exposed organisms, much like in humans, and exposure to MWWE can alter fatty acid and glycogen content and cause changes in glucose and lipid metabolism.[9] The increased xenobiotic metabolism and FA metabolism can increase oxidative stress,[21] and changes in oxidative stress response occurs with exposure to MWWE occurs in gill and liver.[20] Kling *et al.* [41] found that exposure to different concentrations of brominated fire retardants could cause sex-specific changes in the liver proteome, with proteins involved in metabolism, such as α-enolase and aldehyde dehydrogenase 8, being down regulated only in females and males, respectively. This suggests that the proteome changes after exposure to MWWE may be gender specific.

Exposure to toxicants, such as those found in MWWE, are known to effect cortisol-mediated responses in teleosts. Extended exposure to pollutants can induce chronic stress, resulting in an impaired response to acute stressors.[33,35,42] The stress response allows the organism to undergo physiological adaptations to counteract threats to homeostasis; an impaired response counteracts the adaptive value and negatively impacts organism fitness.[43] In a field study of rainbow trout, an acute stressor showed significantly reduced response in cortisol levels, as well as circulating glucose and PEPCK activity, in fish with a high degree of MWWE exposure as compared to the unexposed control. Additionally, genes involved in cortisol synthesis were suppressed 24 h post-stressor after exposure to MWWE as compared to the control, suggesting that the HPI axis was not functioning effectively in response.[35]. Studies using caged trout near MWWE outlets has been shown to increase circulating cortisol as well as the amount of GR present in liver tissue.[35] CYP11A1, an enzyme involved in the synthesis of cortisol from cholesterol showed altered expression at

different levels of MWWE exposure.[42] Endocrine disruptors, another common type of MWWE contaminant,[2] have also been shown to interrupt function of the HPI axis.[35] Exposure studies have found that to atorvastatin impairs stress response, potentially due to altered cholesterol synthesis, along with lowering levels of triglyceride levels and alterations in FA catabolism.[21]

The variety of metabolic function, and the multiple ways the MWWE can impact them, makes the liver a key site for studying the impact of MWWE.

1.3 Proteomics

Determining the impact of MWWE on a organism can be difficult due to the broad range and interacting effects of contaminants. Testing expression changes in a small number of genes, such as the GR or P450scc, can be useful for understanding the impact on a particular function, but testing a large number of individual genes or proteins is time consuming and expensive. Additionally, the variability of MWWE composition and the interaction between components means the impacts of a contaminant mixture may not be characterized well enough to know which genes to test.[44]

To combat these challenges, various 'omics' methods, including transcriptomics, proteomics, and metabolomics, have become an increasingly common alternative to more targeted but less comprehensive approaches.[44] While the different 'omics' approaches have unique strengths and weaknesses, proteomics, the large-scale identification and quantification of proteins[45], has a few advantages that make it particularly suitable for understanding the impacts of MWWE. While transcriptomics experiments usually yield more comprehensive results, identifying a larger number of differentially regulated gene products[46], protein quantification data captures information on post-transcriptional and post-translational regulation that is not available from the transcriptome, such as changes in synthesis and degradation of proteins.[47] Proteins are responsible for the majority of the cell's function, and thus changes in the proteome more closely mimic the changing phenotype of the organism.[44,46] However, only a portion of the potentially 10,000 or more proteins expressed in a typical eukaryotic organism can be detected in a single mass spectrometry run.[47]

1.3.1 Protein Identification by Mass Spectrometry

The key technology for proteome analysis is mass spectrometry (MS). Proteins or peptides are charged by electrospray ionization (ESI) or matrix assisted laser desorption ionization (MALDI) and injected into the mass spectrometer, where they are separated by mass-to-charge ratio (m/z) using a mass analyzer such as a quadropole or Orbitrap, and the detected response as a function of m/z is quantified. By scanning across a range of m/z ratios a spectra of ion intensities will be produced, which can be used to

identify the injected ions. Since proteins are large and difficult to charge, the majority of experiments use a 'bottom-up' approach in which enzymatically cleaved peptides are analyzed, whereas 'top-down' proteomics approaches, which use intact proteins, are rare.[47] The most common enzymes used for cleavage of proteins into peptides are trypsin and LysC, which cleave only at the the C-terminal side of lysine and arginine residues.

While it is possible to identify proteins directly from the masses of their peptides – a process known as peptide mass fingerprinting – the number of unique peptides with overlapping masses limits the effectiveness of this technique for protein mixtures.[45,48] To overcome this limitation, a variety of techniques have been applied. Initially, proteomics experiments relied on separation of proteins by 2D gels followed by identification of individual protein spots with MALDI-time-of-flight (TOF).[44,45] Characterizing even a few hundred of the most abundant proteins with this methods is time consuming, and advancements in mass spectrometers and computing power has moved proteomics away from gel-based methods. The availability of liquid chromatography mass spectrometry (LC-MS) systems, in which complex peptide mixtures are separated by liquid chromatography and fed directly into the mass spectrometer by ESI, has allowed for the analysis of complex mixtures without gel-based separation.[46,47] This, along with the increasing speed of mass spectrometers, has allowed the identification of thousands proteins in a single mass spectrometry run, with the newest methods identifying up to 100 proteins/min.[49]

In addition to effective separation of peptides, proteomics also relies on tandem MS to identify the sequence of individual peptides.[48] Using two or more mass analyzers working in series, high abundance precursor ions are isolated by m/z for collision with an inert gas, causing them to fragment into smaller ions that are measured in the second mass analyzer.[47] Ideally, precursor peptide ions would be fragmented at each position in the peptide backbone, producing a spectra of ions separated by the mass of each residue, though the process of fragmentation is complex and certain fragments are less likely.[47] Fragmentation of tryptic peptides tends to produce y type ions, one or more residues cleaved at the amide bond proceeding from the C-terminus, though b type ions, which begin at the N-terminus, are also frequently observed.[48] An example fragment spectra and the corresponding peptide identification can be seen in Figure 1.1.

The spectra produced from the ion fragmentation are then used to identify the precursor peptide. The most common method is database searching,[50] in which the observed fragment spectra and the precursor ion m/z are compared to the theoretical fragmentation spectra of potential peptides from a protein database.[47] Since even a small database of 10,000 proteins can contain more than 500,000 potential peptides, the candidate peptides are limited to those with a calculated mass within a predefined range of the precusor ion m/z. This reduces the number of peptide sequences to be scored against each spectra, and thus the required computational time. Theoretical fragment spectra from these candidate peptides are then compared to the observed fragment spectra and scored using a method particular to the search engine,

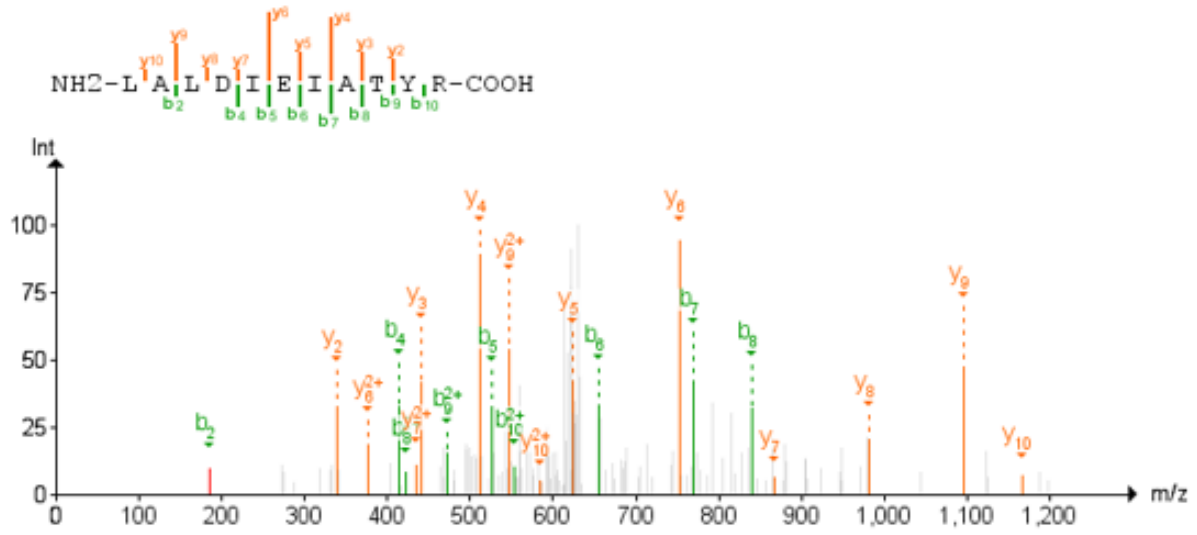


Figure 1.1: Example spectra used to identify the peptide LALDIEIATYR with high confidence. The y and b ions are indicated in orange and green, respectively. y ions predominate, but the majority of fragment ions cleaved at the amide bond and larger than a single residue are observed. Most of the larger fragment spectra (greater than 6 residues) are also found multiply charged, as indicated by the superscript of the peak labels.

and the highest scoring sequence is assigned, producing a peptide-spectrum match (PSM).[51]

It is also possible to sequence peptides *de novo* based on the fragment ion pattern. Theoretically, if the precursor is completely fragmented, the resulting ion spectra should contain peaks separated by the mass of each consecutive amino acid residue. Since tryptic peptides end with a lysine or arginine residue, the y_1 ion mass should always have a mass of 147 or 175. By identifying the subsequent ions, the sequence of the peptide can be identified from the difference between each peak. However, the combination of missing fragment peaks, multiply charged fragments, fragmentation of the backbone outside the amide bond, multiply fragmented ions, and contaminant ion spectra all make the clear identification of the full peptide sequences computationally challenging and error prone.[48] The most successful tools use *de novo* sequencing in combination with database search to improve identification rates.[50]

1.3.2 Scoring Peptide Identifications

Theoretically, each spectra produced during a mass spectrometry experiment should correspond to a peptide from the original sample; in practice, only 10-50% of PSMs represent true hits.[52] This is due to noise in the spectra, caused by factors such as incomplete fragmentation, or because the true peptide sequence is absent from the database.[51] Thus, after peptide-spectrum matching has been performed, some method for distinguishing correct and incorrect PSMs is needed. While it is possible to simply decide a

cut-off value for scores generated by the search engine, this provides no insight into the degree of confidence in the PSMs, or the number of incorrect matches at a particular threshold.[51]

For database searches, most methods are based on the target-decoy approach first described by Elias & Gygi [52], in which known 'decoy' peptide sequences, unlikely to exist in the sample, are added to the database. These decoy sequences compete with the expected target peptides for the highest score during the database search. Incorrect matches are assumed to be equally likely to occur with target and decoy peptides, while correct matches are significantly more likely to occur with target peptides. Since decoy peptides are assumed to be incorrectly matched, the false discovery rate (FDR) and a variety of other statistics can be estimated from the number of decoy peptides at a particular score, allowing a score cut-off to be selected at an estimated error rate.[52] Extensions of the target-decoy approach include PeptideProphet, which assigns confidences scores to PSMs based on which of two overlapping distributions - correct and incorrect matches - the PSM is more likely to belong to, with the ability to use decoys to better estimate the distributions.[51]

While the target-decoy approach provides an estimate of the FDR, it is limited to assessing PSMs as independent observations within a single experiment. However, on both the peptide and protein level, there are other factors to be considered in assigning confidence to peptide identifications. The simplest case is multiple observations of the 'same' PSM - i.e. an ion with the same m/z ratio eluted at similar times and matched to the same peptide - across multiple experiments. Intuitively, it would make sense to assign a higher confidence to PSMs observed in 3 mass spectrometry runs than a PSM observed in only one, even if the search engine assigned identical scores to both. Other similar scenarios include multiply charged ions from the same peptide, or even multiple identifications by different search engines employing different scoring algorithms. The iProphet software integrates results from multiple search engines, multiple observations of the same precursor ion, multiple observations across experiments, and multiply charged or post-translationally modified versions of the same peptide to increase or decrease the confidence of individual PSMs.[53] Finally, the ProteinProphet software calculates the confidence that a particular protein exists in a sample, with increasing confidence for proteins with larger numbers of high confidence peptides.[54]

1.3.3 Importance of the Protein Database

With the exception of *de novo* sequencing, the methods for identifying peptides and accurately assigning confidence outlined in the previous sections are dependent on the correct peptide sequences existing in the database being searched. All scoring functions rely at least partially on the similarity between the observed ion spectra and the theoretical ones from a database, so the presence of the correct sequence, or

a reasonably close approximation, is necessary for a high scoring match.[50,51,55,56] While it is theoretically possible to compare each spectra to all possible peptides, there are more than 10^{50} possible tryptic peptides between 8 and 40 amino acids long, making the required computational time impractical. Additionally, the target-decoy approach essentially relies on the ability to distinguish correct and incorrect hits based on the distribution of scores; if the number of false positive hits is too high, the distribution of high-scoring correct PSMs will be indistinguishable from randomly high-score false positives in both the target and decoy databases.[57,58]

Thus, the protein database must meet a number of requirements: large enough to contain all potential peptides in the sample, similar enough to the sample to be reliably matched to the experimental ion spectra, but small enough to prevent inflation of false positives or excessive computational time. For many of the organisms commonly studied by proteomics, particularly yeast, mice, and human samples, comprehensive and well-curated proteomes are available, with the ability to include or exclude isoforms; for aquatic organisms, zebrafish is similarly well studied. Unfortunately, the fathead minnow (FHM) genome sequence was first released in 2016[59], and initial annotations were only released in 2017[60], limiting the number of FHM specific proteins available in public databases. The challenges of database selection for FHM will be discussed in depth in Chapter 2.

1.3.4 Protein Quantification and Isobaric Mass Tagging

There are multiple methods for performing quantification of proteins during LC-MS. Label-free quantification, perhaps the most conceptually straightforward method, compares the intensity of precursor or fragment ion peaks across multiple MS runs. While conceptually simple, label-free quantification requires complex matching of the spectra from different samples, and requires an individual MS run for each sample in the comparison.[46,47] stable isotope labeling with amino acids in cell culture (SILAC) and other metabolic labeling techniques involve growing one of a pair of samples in the presence of an amino acid labeled with a heavy carbon or nitrogen isotope (i.e. ^{13}C or ^{15}N). When the combined heavy and normal samples are run on a high resolution mass spectrometer, the spectra from each sample differ by the mass of the isotopes, allowing the relative abundance to be determined from the relative height of the peaks. However, SILAC and other metabolic labeling methods require the ability to grow cells in highly controlled conditions, to ensure proper labeling of proteins by the isotope.[47] Additionally, only two samples can be compared at once.

Isobaric mass tagging provides a method for multiplexed comparisons in a single MS run. Following a tryptic digest, a chemical tag with an amine reactive group is added to the sample, where it bonds to the amine (N-) terminal of the peptides. Different samples receive tags that are structurally identical

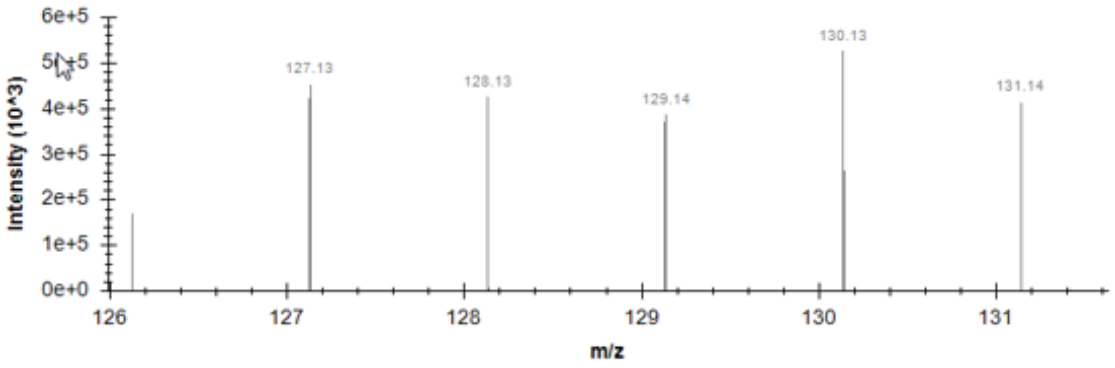


Figure 1.2: Example reporter ion spectra for TMT-10plex tags. Peaks corresponding to the 6 major reporter ion m/z of *126,127,128,129,130,131* can be seen, but the difference between the ^{13}C and ^{15}N report variants cannot be seen at this scale.

and equally massive (isobaric), but contain a unique arrangement of carbon, nitrogen, and oxygen isotopes across the structure. The peptides from different samples are then pooled together and run through LC-MS system. Since the tags are isobaric, similar peptides across samples are eluted and ionized at the same time. During fragmentation, however, the tags split into three pieces - a unique reporter group, with a characteristic m/z between 100–150 Da, a small balance group, and the amine reactive group, which remains bonded to the peptide. Reporter groups from each tag contain slightly different distributions of isotopes, causing small mass differences between them. This creates a distribution of reporter ions in which the m/z of the ion identifies the original sample, and the intensity of that ion is proportional to the relative abundance of the peptide (Figure 1.2). The balance group contains a complementary isotope distribution, so that the combined mass of the reporter group and balance group is the same for all tags. This mean the amine-bonded reactive group has an identical mass in all samples, so the fragmentation pattern produced for the peptides from the different samples is identical, which simplifies peptide identification. Since the peptide ion intensity is cumulative across the multiplexed samples, identification rates for low abundance peptides are increased, and the selected precursor ions are more consistent between different samples compared to label-free methods.[61] With TMT-10plex tags allowing up to 10 different samples to be combined in a single run, the number of experimental conditions and replicates can be significantly increased without requiring additional instrument time.

However, the need to accurately distinguish and quantify reporter ions when using isobaric mass tags introduces some drawbacks and experimental complications. The small mass differences between the reporter ions, particularly the 6.32 mDa difference between some TMT-10plex tags, requires the use of high resolution mass analyzers to properly resolve low m/z reporter ion peaks for quantification.[62] The most significant flaw in isobaric mass tagging is that quantification from the reporter ion spectra assumes

that all reporter ions are sourced from the same peptide, but experimental results have shown that the complex samples derived from tryptic digest of full proteomes can result in multiple precursor peptide ions to be selected simultaneously.[63] The reporter ion overlap between multiple peptides reduces the relative differences between reporter ions, underestimating the fold-change between different samples, an effect known as ratio compression.[61]

Multiple methods to combat ratio compression have been proposed, ranging from sample preparation to statistical analysis.[63,64] A common approach is to reduce complexity by prefractionating samples and running them separately, which reduces the number of peptides being eluted in a given time window. However, prefractionation cannot fully prevent co-eluting peptides and introduces other fraction-specific effects, as well as drastically increased instrument time to run multiple samples.[61] More recently, MS3-based methods have been introduced. Originally, reporter ions were fragmented and quantified during MS2, along with the peptide fragment ions. With newer instruments, by careful selection of instrument settings, only the peptide backbone can be fragmented during MS2, while the reporter tags remains intact on the new fragment ion.[65] Multiple fragment ions in the MS2 spectra are then simultaneously selected and fragmented again with higher energy, breaking the tags and producing reporter ion spectra from a limited set of peptides.[66] By limiting the signal from co-eluting peptides and other contaminants, the ratio between reporter ions is more accurately reported, significantly reducing compression.[65]

While MS3-based quantification significantly improves the accuracy of quantification, it incurs several trade-offs. The requirement of an additional MS scan for each isolated precursor ion reduces the total number of scans that can be performed in a given run, and requires the use of specialized triple mass spectrometers. The method also requires that MS2 spectra be analyzed in the less accurate linear ion trap, reducing the accuracy of the peptide fragment spectra. The combined effect is a decrease in total number of peptides, and thus proteins, identified from the sample.[65,66]

1.4 Conclusion

Investigation of the liver proteome is an important tool to better understand the impacts of MWWE exposure. The liver is a primary target for the hormones produce by the HPI axis, including cortisol, and helps regulate metabolism through glucose and lipid metabolism. Exposure to MWWE effects liver function, changing expression for a variety of genes involved in metabolism, homeostasis, and stress adaptations. This makes the liver a prime target for investigation of the effects of MWWE exposure on the stress response.

Chapter 2

Construction and Testing of a Fathead Minnow Protein Database

2.1 Introduction

Analyzing changes in protein expression is a key part of understanding in a biological organism. While advances in RNA sequencing technologies have made transcript level measurements increasingly accessible[67], mRNA levels are generally poorly correlated with protein abundance. A variety of cellular mechanisms, such as targeted protein degradation and post-transcriptional regulation, impact the correlation between mRNA and protein abundance.[68] Under some conditions protein abundance can be negatively correlated with mRNA levels, such as the glucocorticoid receptor (GR) in teleosts, reflecting additional controls on protein expression in addition to mRNA abundance.[36] Thus, investigating changes in the proteome allows researchers to capture the state of the molecular 'machinery' carrying out cellular function, rather than the 'blueprints' for those machines, providing an additional layer of insight into the biological function of the cell.

Currently the dominant method for proteomics analysis, shotgun proteomics, produces large scale identifications of proteins in the proteome by utilizing nanoLC and tandem mass spectrometry (MS) technologies.[46] A key step in the data analysis of proteomics experiments is identifying the original peptides from the mass spectra produced. Depending on research goals, it may also be important to identify proteins variants and modifications, which also affect the mass spectra produced.

A key factor in shotgun proteomics is the database used for protein identification. Generally, proteomics experiments rely on the availability of comprehensive protein databases for reliable peptide identifications,

limiting the application of proteomics when a database is not available.[69] While well-studied model organisms, such as zebrafish (*Danio rerio*), have comprehensive protein sequence information readily available for proteomics work, such is not the case for non-model organisms.[70] While the fathead minnow (*Pimephales promelas*) is commonly used for aquatic toxicology studies,[71] there is little protein sequence data available in public repositories such as the Uniprot.[72]

As mass spectrometry and sequencing technologies have become increasingly ubiquitous the field of proteogenomics, the integration of nucleotide sequence data and proteomics analysis, has arisen.[73] The combination of RNA sequencing data and proteomics had led to multiple tools for generate sample specific protein databases from both *de novo* and reference-based transcript assemblies.[67,69,74] Proteome data has also been combined with genome sequence data to produce protein databases for prokaryotic organisms,[75–77] or to identifying point mutations.[78] However, to the author's knowledge, while genome-based protein identification has been discussed since the early days of proteomics,[48,77], little has been done to specifically analyze the impact of genome-derived predicted proteins as a database source compared to publicly available protein databases. Predicted gene sequences from a sequenced genome have primarily been used for updating and re-annotating existing protein annotations or to add protein sequences to an existing species specific database,[79–81] with few attempts to use purely genome-derived sequences for a protein database.[70,82,83]

Many search engines exist to map MS spectra to corresponding peptides. Two popular tools are X!Tandem and Comet, which score observed mass spectra against predicted spectra computed from a database of proteins.[47] More recent tools attempt to improve on these methods to increase peptide identification rate. These include MSFragger, which performs fast searching of spectra with large mass variations[56], and *de novo* peptide sequencing tool PEAKS, which combines customized database search engines with *de novo* peptide sequencing methods.[50] To produce accurate and reproducible results, it is important to determine how these different tools perform, in terms of both the number of protein identifications, and the degree of similarity in the identifications, and the ability to identify modifications to proteins.

In this chapter, we describe the creation of a proteomics database using the publicly available gene annotations of the fathead minnow genome,[59,60] and analyse the effect on protein identification as compared to the general purpose UniProt database and the zebrafish reference proteome.[72] By using existing protein annotations, identified proteins can be directly associated with with annotation in the fathead minnow (FHM) genome. We also compare the performance of several popular search engines, including open-source tools X!Tandem, Comet, MSFragger, and the commercial program PEAKS on multiple mass spectrometry samples. Search engine comparisons were made using three datasets: the FHM-specific database and FHM MS data described in this thesis, as well as two other datasets previously generated in the McConkey Lab; one trout dataset generated using different MS equipment and a human dataset generated with the same

MS equipment as the FHM data.

2.2 Methods

2.2.1 Database Comparison

Three databases will be used for the purposes of database comparison. The first is the Fathead Minnow Predicted Proteome (FHMP), which is constructed from annotations of predicted protein-coding regions in the FHM genome sequence produced by Saari *et al.* [60]. Two sets of annotations from that work were used for database construction. One set of annotations is from the *de novo* gene predictions software Augustus, and the other from the alignment of zebrafish transcripts to the FHM genome using the program exonerate.[60] Construction of the FHMP is described below.

The UniRef90 Cyprinidae Proteome (U90CYP)[84] and the Uniprot Zebrafish Reference Proteome (ZRP)[72] are used for comparison. U90CYP is composed of the reference sequences for all proteins in UniProt with "Cyprinidae" in the taxonomy, clustered at 90% identity. The ZRP is the UniProt reference proteome for zebrafish.

Proteome Construction

The FHM genome sequence,[59] along with *ab initio* Augustus gene prediction and zebrafish transcript alignment annotations,[60] were downloaded from the Society for Environmental Toxicology and Chemistry (SETAC) website on October 24th, 2017.[85] An in-house python script (Appendix A.1) using the *gffutils* python package[86] was used to collect all the protein coding regions from the genome sequence data.

For each of the Augustus-predicted genes, coding sequence regions were concatenated into a single open reading frame (ORF). ORFs were checked for completeness, requiring an in-frame coding sequence with start and stop codons at each end, and any incomplete ORFs were dropped. The coding sequences were then translated into the corresponding protein sequences.

For the aligned zebrafish transcripts, the exons were concatenated, then translated into protein sequences using TransDecoder[87]. TransDecoder identified ORFs longer than 80 residues and scores them for likelihood of translation. Additionally, it searched the ORFs against a database of all UniprotKB teleost sequences using Blastp[88] and against the Pfam[89] domain database using hmmscan[90]. High scoring ORFs, along with those containing blast or Pfam domain hits, were translated to protein sequences.

The two interim protein sequence databases derived from the Augustus gene predictions and zebrafish transcript alignments were assessed for completeness using the Benchmarking Universal Single Copy Orthologs (BUSCO) software suite[91] with the Actinopterygii lineage dataset. BUSCO assess the completeness

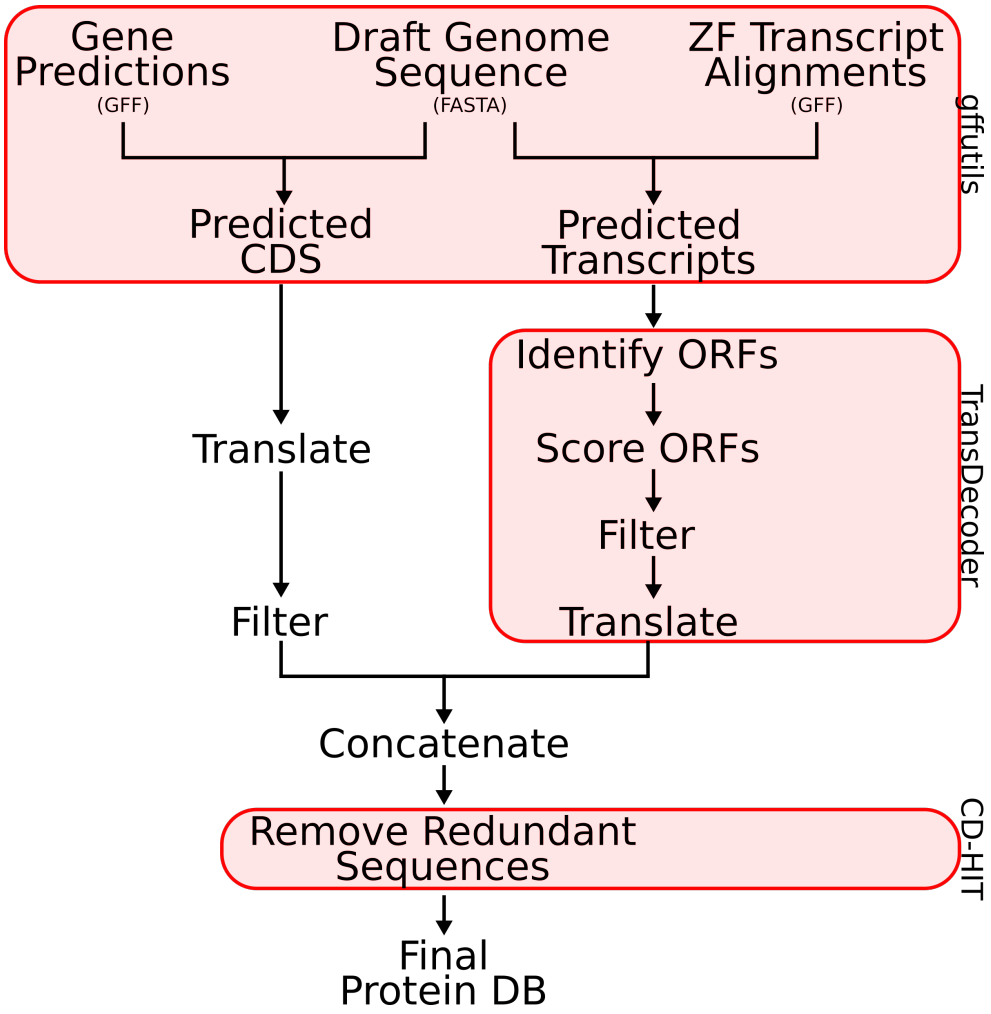


Figure 2.1: Construction of the FHMP from FHM genome annotations.

of genomes, transcriptomes, and proteomes by identifying and quantifying genes normally found as single copies in other members of the same taxonomic group. This serves as an estimate of what fraction of genes in the original organism are found in the dataset. Results were visualized using the provided BUSCO plotting tool. Additionally, DOmain-based General Measure for transcriptome and proteome quality Assessment (DOGMA)[92] was performed using conserved domain arrangements (CDAs) up to length three from the vertebrate CDA collection to analyze the functional portions of protein sequences. Analysis of CDA by DOGMA is an alternative quality measure to BUSCO. DOGMA identifies the frequency and order of sets of conserved Pfam domains to measure the completeness of a proteome.[92] Since domains usually have conserved and independent function, measuring domain frequency indicates functional, rather than sequence-level, completeness. The protein sequences from the annotations were then combined into a single, concatenated Fathead Minnow Predicted Proteome (FHMP) database.

Proteome Quality Testing

After proteome construction, additional reference databases were compared with the generated Fathead Minnow Predicted Proteome. The UniRef90 Cyprinidae Proteome (U90CYP)[84] and the Uniprot Zebrafish Reference Proteome (ZRP)[72] were downloaded for comparison to the FHMP. The proteomes, as well as the FHM genome, were analyzed for completeness using BUSCO with the Actinopterygii lineage dataset, as well as DOGMA using the vertebrate dataset. Since DOGMA does not work on complete genomes, DOGMA analysis was not performed on the FHM genome.

Effects of DB Size and Redundancy

Confident identification of proteins from mass spectrometry data requires peptides that map uniquely to that protein and no others. Redundant proteins were removed from the FHMP using cd-hit[93] to reduce repetitive protein sequences in the proteome. Proteins were clustered at a range of identity thresholds of 85%, 90%, 95%, and 100%, and the longest protein in each cluster was kept in the proteome as the reference sequence. BUSCO and DOGMA analyses were performed after clustering the FHMP at each threshold value to identify the effects of clustering on the proteome completeness.

In order to assess the frequency of unique peptides, an *in-silico* tryptic digest was performed with a custom python script (Appendices A.3, A.4) for the initial FHMP and at each identity threshold, as well as the ZRP and U90CYP. Tryptic peptides between 2 and 35 residues were retained, with up to one missed cleavage, for all protein sequences in the database. The number and length of all possible peptides (unique peptide sequences) in the proteome were calculated, as well as the number of occurrences of each peptide sequence. The uniqueness of the proteome was calculated as the fraction of single occurrence peptides over the total number of possible peptides in the proteome for peptide lengths 2 through 35. Additionally, a six-frame translation of the fathead minnow genome was performed *in-silico* using EMBOSS transeq[94] to create a database of all potential peptides, except those spanning exon juctions, and the set of peptides was compared to the set of peptide of the U90CYP and the FHMP.

To investigate the effects of proteome size and redundancy on a decoyed database, the complete UniProt Actinopterygii protein database was also downloaded and randomly sub-sampled at various sizes. Decoys versions of the databases were created by pseudo-random shuffling using the Trans-Proteomic Pipeline (TPP)[95] decoy generation tool to add decoy proteins, and an *in-silico* tryptic digest was performed. Pseudo-random shuffling generates decoys by 'cutting' each protein sequence into peptides after particular residues, shuffling the sequence between those residues, then joining the peptides back together in the same order. This creates randomized decoy proteins with similar amino acid compositions and length distributions as the target set. For decoy generation, the default residues were used for shuffling: glycine

and phenylalanine, except when they preceded asparagine. The database and subsamples were then clustered at 95% identity, and the decoy generation and tryptic digest were repeated. The overlap between decoy and target peptides, as well as the total number of peptides, was calculated based on peptide sequence for both the original, redundant databases and the versions clustered at 95% identity.

Following analysis of the proteome for completeness, duplication, and the effects of clustering, the final FHMP was clustered at 90% identity for subsequent analysis.

Effect of database on search results for a FHM dataset

In addition to the comparison of search engine performance on different datasets described in Section 2.2.2, the effect of different protein databases on protein and peptide identification rates was compared using the FHM MS dataset. Details on the MS data generation can be found in Section 3.2.2. A proteomics pipeline were run for the three different protein databases:

- FHMP at 90% identity: 37,454 sequences
- U90CYP: 42,248 sequences
- ZRP: 25,747 sequences

The 6-frame translation of the FHM genome was excluded from the comparison because 6-frame translations are known to perform poorly compared to other databases due to the over-abundance of 'incorrect' target sequences in the resulting database.[58] Decoys were added to each dataset by pseudo-randomly shuffling the sequences. The Comet and X!Tandem search engines were used for the peptides search against each database. Results from the different runs and search engines were scored and combined using the PeptideProphet, iProphet, and ProteinProphet data analysis tools in the TPP. Results were compared on the basis on overall number of proteins identified, the number of peptides, and the peptides matched to each spectra in the MS dataset.

2.2.2 Search Engine Comparison

In order to compare the effectiveness of different search engines and data analysis tools, proteomics pipelines were run on three liquid chromatography mass spectrometry (LC-MS) datasets generated in the McConkey lab (Table 2.1). The FHM MS data is the same as used for the FHMP database testing (Section 2.2.1), described in Section 3.2.2. The Human dataset was generated using the same MS equipment and conditions as the FHM MS data, while the Trout dataset was generated using the higher-accuracy Orbitrap rather

Table 2.1: Details for each sample used in the peptide search comparison

	Human	**Trout**	**Fathead Minnow**
Tissue	Breast cancer cell culture	Liver	Liver
Digestion	Trypsin/LysC	Trypsin	Trypsin/LysC
Mass Spec	Fusion Lumos	Q-Exactive	Fusion Lumos
MS Runs	1	3	3
MS2 Spectra	146,082	122,227	144,532
MS2 Source	Ion trap	Orbitrap	Ion trap
Database	Uniprot Human Proteome	Uniref100 Rainbow Trout	FHM Gene Predictions
Database size	93,591	49,519	23,191

than the ion trap used for the Human and FHM datasets.[66] This allows for consideration of database accuracy and the effects of different MS conditions when comparing search engine results.

The databases used were a FHM protein database generated from the genome using only the *de-novo* gene predictions clustered at 90% identity. The Uniprot human reference proteome[72] and the Uniref proteins found in *Oncorhynchus mykiss* clustered at 100%[84] were used as the human and rainbow trout databases, respectively. For consistency in search results, decoys were generated and concatenated to the FASTA database once for each database using PEAKS Studio[50] v8.5 and used for all search engines. The decoyed databases were searched using Comet[55], X!Tandem[96], MSFragger[56], and PEAKS[50]. Search parameters were kept as similar as possible for all search engines, allowing for up to one missed cleavage, with TMT10plex tags and carbamidomethylation as fixed modifications and oxidation of methionine as a variable modifications. For MSFragger, which is designed to search a large mass range, the precursor mass tolerance range was set to ±500 Da. The precursor tolerance for PEAKS, Comet, and X!Tandem was set to 10 ppm for the Human database search, and 15 ppm for the FHM database search, typical ranges for ion-trap MS spectra, while the higher accuracy Orbitrap spectra in the Trout dataset were searched at 0.1 Da. Additional parameters are included in Appendix A.2.

The rate of false positives was estimated from hits to decoy proteins in the database. An estimated 1% false discovery rate (FDR) was used as the cutoff for protein identification from each search engine. For X!Tandem, Comet, and MSFragger, post processing with PeptideProphet[51], iProphet[53], and ProteinProphet[95] was used for integration of peptide and protein identifications across runs, as well as post-translational modification (PTM) identification. For PEAKS, PTMs and modifications were identified with the PTM and SPIDER search engines. Scan numbers were used to identify which spectra were mapped to peptides by each search engine, and the identified peptides and proteins were matched by sequence and protein ID, respectively, to determine similarity between search engine results.

2.3 Results and Discussion

2.3.1 Generation of Protein Databases from Genomic Sequence

Constructing the FHMP database from the genome sequence and annotations yielded a total of 65,236 protein sequences, 42.5% (27,784) of which were derived from Augustus-predicted gene annotations, and 57.5%(37,454) from the locations of aligned *Danio rerio* transcript annotations.

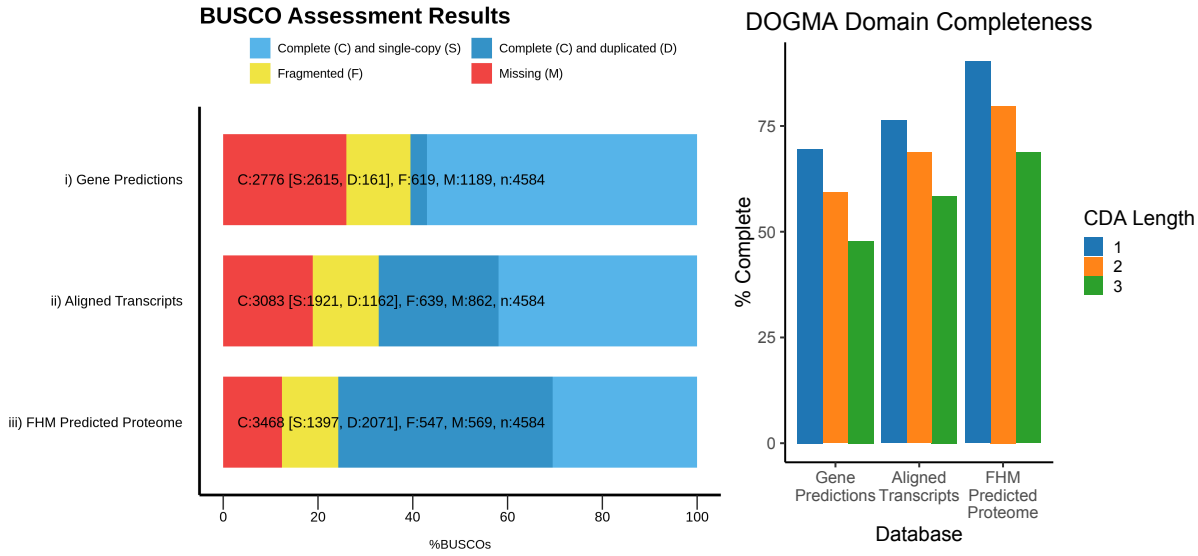
The Benchmarking Universal Single Copy Orthologs (BUSCO) dataset was used to compare the completeness of the proteins from each annotation set, as well as the combined FHMP. This assesses if protein sequences could be successfully be constructed from the genome sequence and annotations, and the redundancy of the protein sequences from the two different annotations.

Construction of the FHMP produced complete copies of 3468 of the single-copy orthologs, or 75.7% of the 4,584 genes in the Actinopterygii gene set. 569 BUSCO genes are missing, and another 547 are duplicated, while over 2000 of the complete BUSCO genes are found in duplicate (Figure 2.2aiii). The duplication is due to redundancy within the aligned transcripts and between the aligned transcripts and the Augustus-predicted genes. All but 385 proteins found in the final protein database are present in the aligned transcripts (Figure 2.2aii), and 1162 are present in more than one copy. Despite the redundancy, combining the two datasets still increased the total number of complete Actinopterygii single-copy orthologs.

Figure 2.3a compares the results of BUSCO analysis on the initial FHMP prior to removing redundant sequences, the full FHM genome, the UniRef90 Cyprinidae Proteome (U90CYP), and the Uniprot Zebrafish Reference Proteome (ZRP) against the 4,584 single-copy orthologs found in the OrthoDB v9 Actinopterygii database. The U90CYP is nearly complete, with 98.0% of genes found, but also contains 25.6% of complete genes in more than one copy. Given that Uniprot collects protein sequence from a wide variety of organisms, including both predicted and experimentally observed proteins, the high degree of completeness and a degree of duplication is not surprising.

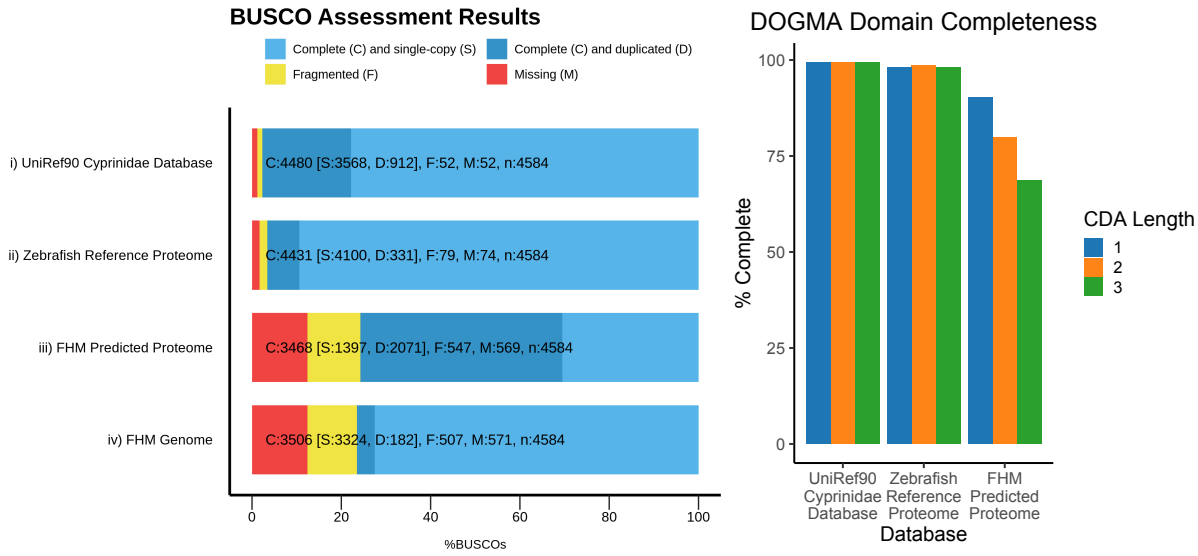
While BUSCO analysis of the manually curated zebrafish genome identified complete copies of 96.7% of the expected genes with a low rate of duplication, 23.5% of all BUSCO genes lack complete copies in the FHM genome sequence. This is consistent with the 74% of conserved eukaryotic genes found by the authors of the original genome sequencing. The missing genes are likely due to the fragmented nature of the genome, as only short-read sequencing technologies were used for genome assembly.[59] However, the FHM genome also shows a reduced rate of gene duplication compared to the ZRP, again likely due to the incomplete nature of the FHM genome.

The number of detected complete BUSCO genes decreased by 38 from the FHM genome to the FHMP (Figure 2.3ac and d), primarily as an increase in the fragmented genes. This is likely due to missing exons



(a) (b)

Figure 2.2: Assessment of FHMP completeness by BUSCO and DOGMA. a) Detected BUSCO genes in proteins converted from: i) *de novo* gene predictions ii) aligned zebrafish transcripts iii) the two sources combined with no redundancy correction. b) CDAs detected in proteins converted from *de novo* genes prediction, aligned zebrafish transcripts, and the two sources combined.



(a) (b)

Figure 2.3: Comparison of protein database completeness by BUSCO and DOGMA. a) Detected BUSCO genes in i) the UniRef90 Cyprinidae Proteome (42,248 protein sequences) (ii) the Uniprot Zebrafish Reference Proteome (25,747 protein sequences), iii) the Fathead Minnow Predicted Proteome (65,236 protein sequences), and iv) the Fathead Minnow (FHM) genome. b) Detected vertebrate CDA in the UniRef90 Cyprinidae Proteome ii) the Uniprot Zebrafish Reference Proteome, iii) the Fathead Minnow Predicted Proteome. Length indicates the number of domains in the CDA.

or early termination of proteins, resulting in the loss of one end of the sequence. Additionally, BUSCO uses Augustus with specific training parameters to predict the location of genes in the gene set, which has been shown to perform better than using pre-trained parameters from another species[91] as was done for the FHM genome annotations.[60] Both of these factor would result in more gene being detected by BUSCO than were present in the annotations to translate.

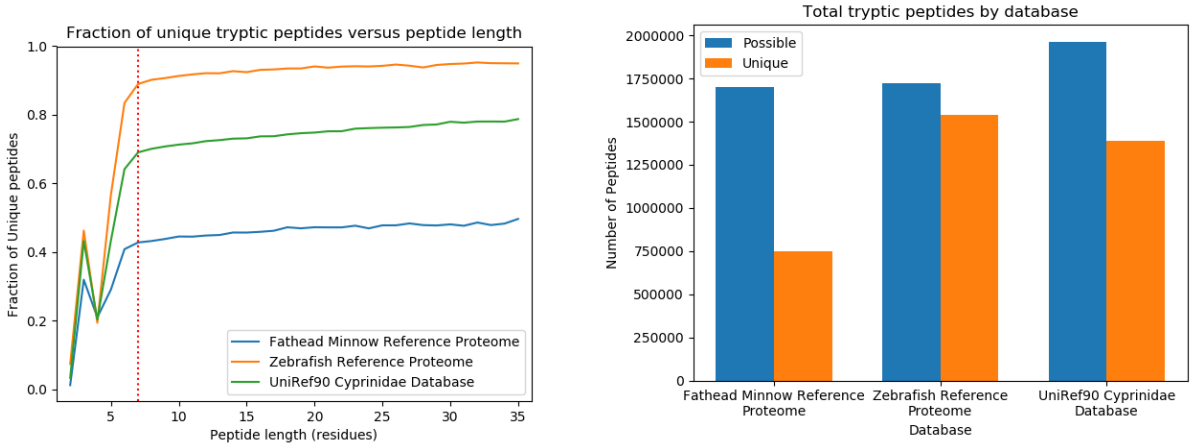
In addition to sequence-level BUSCO analysis, identification of conserved domain arrangements (CDAs) was performed using DOGMA. Since DOGMA identifies sets of protein domains, rather than whole genes, it estimates the degree to which functional portions of proteins are present. Despite the difference in method, the results of CDA and BUSCO analysis are largely consistent for the protein sets (Figure 2.3), with CDA completeness being slightly higher than the identification rates in BUSCO. For the FHMP, the overall average is consistent with BUSCO scores, however completeness decreases strongly with increasing CDA length. This further supports that the FHM genome, and the resulting protein database, is highly fragmented and producing incomplete protein sequences.

Reducing Database Redundancy Improves Peptide Search Space

The purpose of BUSCO is to benchmark genome assemblies and gene sets using near universally-present single-copy orthologs. Results indicate that 59.7% of all detected complete genes in the FHMP (Figure 2.3ad) are duplicates showing that the FHMP is heavily redundant.

In-silico tryptic digest of the U90CYP, ZRP, and FHMP also support that the FHMP and U90CYP databases are redundant (Figure 2.4a). The ZRP is well curated and contains a limited set of proteins meant to completely represent the Zebrafish proteome. While the Zebrafish reference is less than 50% of the size of the FHMP at 25,747 protein sequences, it contains more possible tryptic peptide than the FHMP, and over 90% of tryptic peptide sequences longer than seven residues occur only once in the database. The U90CYP and FHMP contain less than 70% and 50% unique peptides longer than seven residues, respectively. Thus, while the U90CYP and FHMP contain more proteins, there are fewer peptides that map uniquely to one protein compared to the ZRP. This reduces of the effectiveness of the database for shotgun proteomics, as correct peptide IDs are essentially lost due to the inability to distinguish between multiple proteins.

Removing redundant proteins using cd-hit[93] has a large impact on the rate of unique peptides greater than length 7 (Figure 2.5a). For the FHMP, removing identical protein sequences increased the fraction of unique peptide sequences by 10 percentage points, with almost no change in the quantity potential tryptic peptides (Figure 2.5b). Allowing for inexact matches, by reducing the required percent identity, showed diminishing returns beyond 90%. At that level approximately 70% of peptides greater than length



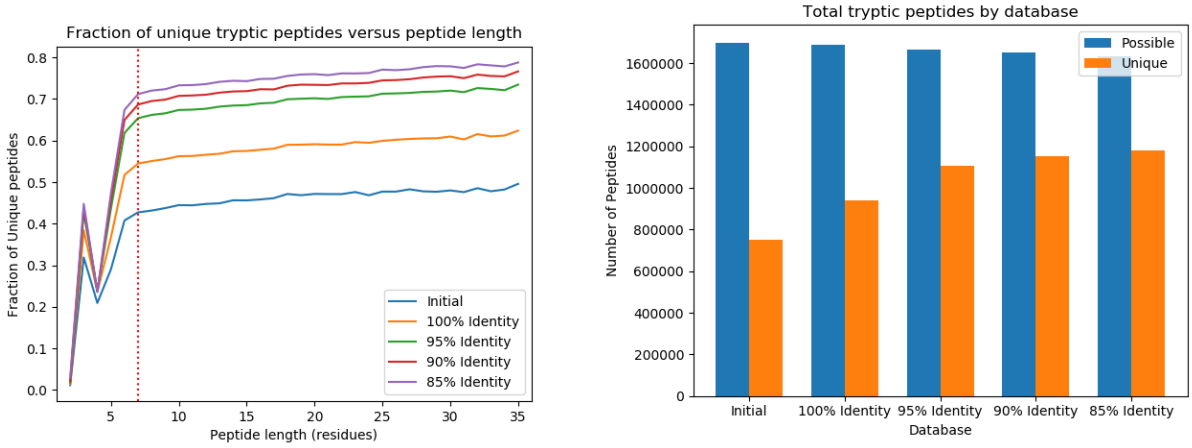
(a) (b)

Figure 2.4: Comparison of frequency of unique peptides in the different protein databases. a) Fraction of unique (single occurrence) potential tryptic peptides vs peptide size in the FHMP, U90CYP, and ZRP. The dashed red line indicates a length of seven, a standard cutoff for minimum peptide size during peptides searches. b) Number of possible peptides (unique sequences) and unique (single occurrence) peptides in the databases.

7 are unique in the database, almost 30 percentage points higher than the original database. Uniqueness is roughly equivalent to the U90CYP but still significantly behind the ZRP (Figure 2.4a).

Despite the greatly increased proportion of unique peptides, clustering at identity thresholds as low as 85% had minimal effect on the total number of peptides (Figure 2.5b), and the distribution of peptide lengths is unchanged (Figure A.1). BUSCO analysis of the database at various levels of clustering (Figure 2.6a) shows that the number of duplicated BUSCO genes is drastically reduced by removing similar protein sequences, with almost half of duplicated BUSCO genes being removed when identical sequences are clustered, and nearly two-thirds at 90% identity threshold. Despite the decrease in duplicated BUSCO genes, there is almost no change in the number of complete BUSCO genes, with only 15 of over 3000 being lost between after clustering the database at 90% identity. Additionally, the number of proteins in the database was reduced by 37%, removing over 21,000 protein sequences (Figure 2.6b) while the number of unique peptides available increased by around 300,000 peptides (Figure 2.5b). Thus, removing similar protein sequences by clustering with cd-hit significantly increases the number of proteins that can be uniquely identified, but the overall set of tryptic peptides available for search has not been reduced, and there is almost no loss in functional protein sequence.

Of the 37,454 proteins in the FHMP after clustering at 90% identity, 23,137 proteins are from the gene predictions and the remaining 14,317 sequences are from the aligned transcripts. As the identity threshold for clustering was lowered, the removed proteins were primarily from the transcript alignment (Figure



(a) (b)

Figure 2.5: Comparison of frequency of unique peptides in the Fathead Minnow Predicted Proteome after clustering a various identity thresholds. a) Fraction of possible tryptic peptides that occur only once in the database vs. peptide size. Different lines indicate clustering at various identity thresholds. The dashed red line indicates a length of seven, a standard cutoff for minimum peptide size during peptides searches. b) Quantity of possible and uniquely occurring peptides after removing redundant proteins at various identity thresholds.

2.6b). Since cd-hit retained the longest sequence during clustering, this suggests that transcript alignment may not capture the full length of some genes, or that Augustus over-extends gene prediction. Investigating difference between the gene predictions and transcript alignments may be a productive method for further improving the FHM genome annotations.

Reducing redundancy has effects in addition to increasing the number of unique peptides in the database. The majority of peptide search engines rely on a target-decoy strategy to distinguish correct and incorrect peptide identifications.[52,77] There are two assumptions of the target-decoy method that are dependent on database size and redundancy. First, the target-decoy method assumes effectively zero overlap between the set of target and decoy peptides used in the search, an assumption which held true for peptides greater than length 8 in the relatively small database of proteins used by Elias and Gygi when they originally described the method.[52] Second, it assumes that target and decoy peptides occur at a known ratio, typically 1:1.[52,58]

There are a variety of methods for decoy generation in protein database, the most common of which is protein reversal.[97] However, many proteomics search engines, including the TPP and PEAKS, perform pseudo-shuffling of protein sequences to generate a more random distribution of decoy peptides with similar length and amino acid compositions. Since every protein sequence will be shuffled independently, duplicate sequences in a redundant protein database will introduce multiple decoy peptide sequence for a single target sequence. This skews the decoy-target ratio, increasing the number of overlapping peptide sequences in

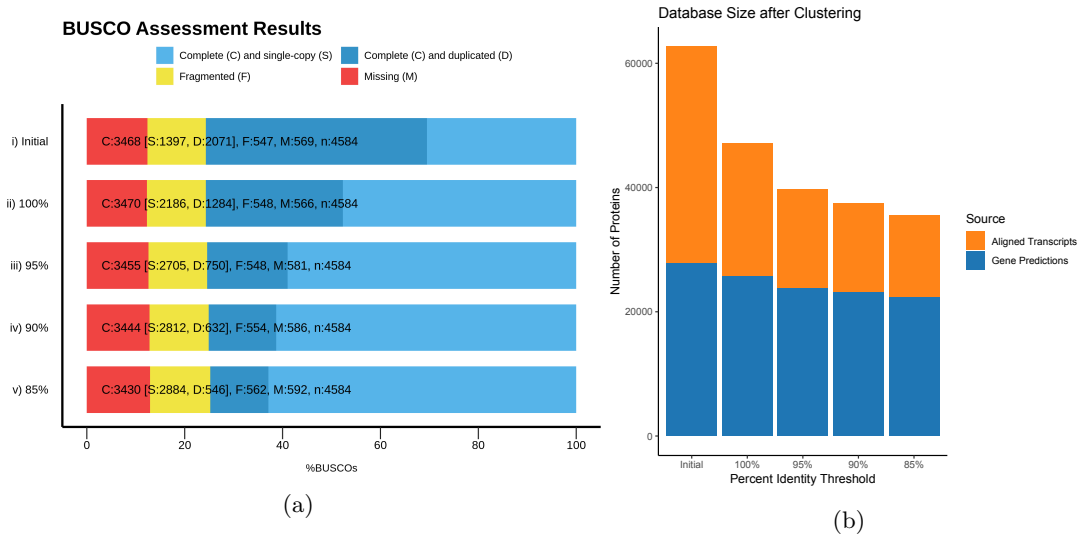


Figure 2.6: Effects of clustering on FHMP size and completeness. a) BUSCO analysis after clustering at various identity thresholds. i) 'Initial' the original set of protein sequences identified during database creation, ii-v) clustering at the indicated percent identity with cd-hit. b) Number of proteins in the database after clustering at various identity thresholds. Color indicates is the protein sequence was derived from an Augustus-predicted gene or from an aligned zebrafish transcript.

the target and decoy set and the chance of high scoring decoy hits. The impact of database redundancy can be seen in Figure A.3, in which the Uniprot actinopterygii protein set was sub-sampled at various fractions, and decoys were added before and after redundancy removal at 95% identity threshold. Despite the larger number of proteins in the non-redundant database, there were fewer unique peptides, and the overlap between the target and decoy peptide sets increases faster in redundant protein sets than the non-redundant versions. Since the number of false positives and subsequent search statistics are estimated from the number of high-scoring decoy hits,[52] an inflated decoy peptide set can result in decreased search sensitivity.[57] Partly, this is because the FDR is calculated from a expected ratio of false positive target and decoy peptide hits, i.e. is the ratio of target:decoy peptides is 1:1, the FDR at a given score threshold is twice the number of decoy peptides passing the scoring threshold. However, change in the ratio of decoy to target peptides can be accounted for when calculating FDR.[52,58] The more significant issue is the increased overlap between target and decoy peptides, as the target-decoy method assumes decoy hits are incorrect.

The FHM-specific Protein Database is Highly Unique

Though the FHMP contained fewer BUSCO genes than the ZRP and U90CYP, an *in-silico* tryptic digest of the protein sequences (Fig. 2.7) shows that it contains a large fraction of peptides that do not occur in the other potential databases. Most notably, the U90CYP protein databases contains a relatively small proportion of peptides that are likely to occur in a tryptic digest of FHM tissue, with less than 25% of over

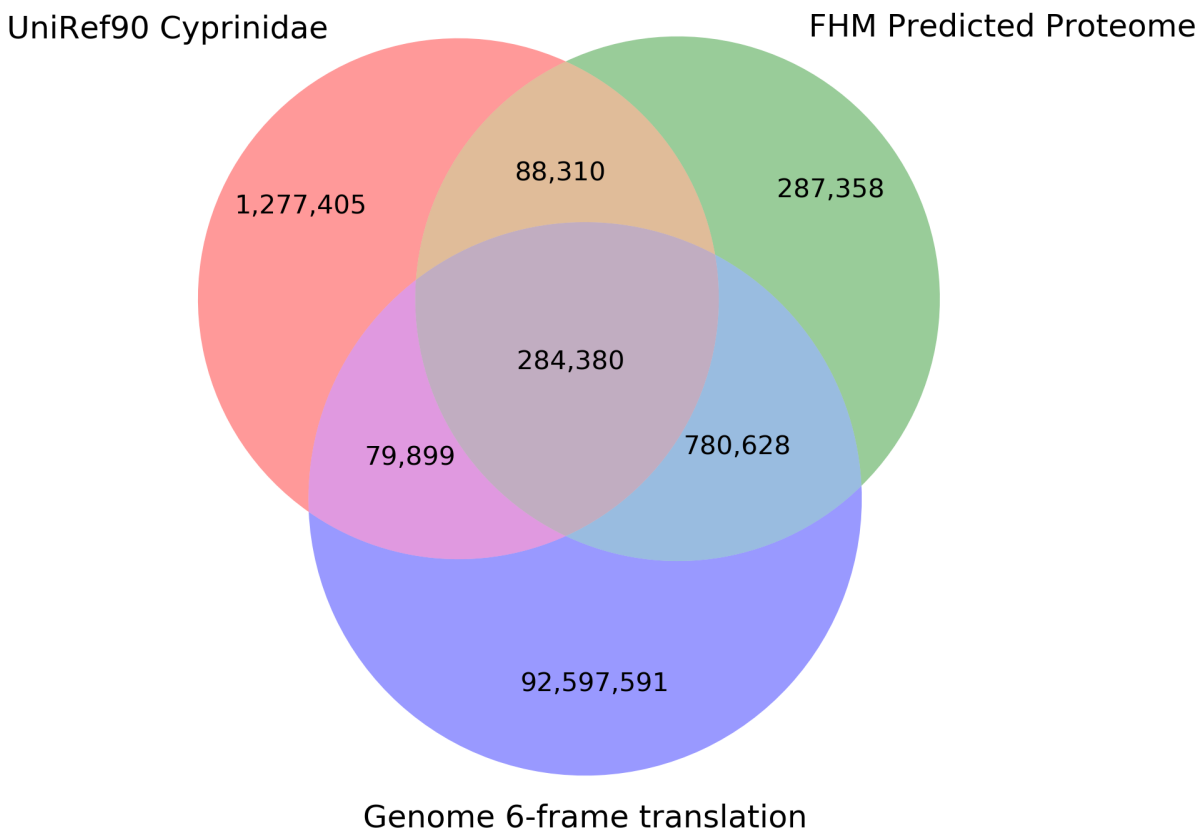


Figure 2.7: Overlap of tryptic peptides longer than 7 residues found in a 6-frame translation of the FHM genome, the U90CYP, and the FHMP. Circles are not weighted to peptide count.

1.5 million peptides being found in either the 6-frame translation or the FHMP tryptic peptide sets. As can be seen in Figure 2.3a, BUSCO analysis showed a nearly complete set of genes, and the true fathead minnow proteome most likely has orthologs to the vast majority of protein present in the U90CYP. The low proportion of overlap is likely due to substitutions and indels accumulating in different species over time, rather than the gain and loss of large numbers of proteins. Thus, despite likely representing a more complete proteome than the FHMP in terms of the set of proteins present, the U90CYP contains fewer correct peptides. Using the U90CYP for the search would result in the majority of spectra being searched against peptides sequences with at least one incorrect residue. Since the search engines rely on matching theoretical mass spectra to observed ones, incorrect residues reduce the score of spectra, reducing the score of matched peptides.

The 6-frame translation of the FHM genome could be used as a database for peptides search, since it would, by definition, contain the vast majority of possible tryptic peptide sequences found in proteins

encoded in the genome. However, Blakeley *et al.* [58] showed that databases using 6-frame translation of genome sequences significantly reduced the sensitivity of protein identification, due to the inflation of decoy peptides relative to the target database. This is related to the previously mentioned second assumption of the target-decoy approach, that decoys occur at a known ratio to target peptides. The vast majority of 6-frame peptides do not undergo translation, due to low gene density and minimal overlap of coding regions.[58,98] However, during creation of a decoyed database, peptides that do not actually exist in the sample of interest still add to the number of decoys, since they also undergo shuffling or reversal. Similar to decoy peptides, target peptides generated from sequences outside the coding regions cannot generate true hits, only false positives. This increases the number of false positives, and thus the false discovery rate, at a fixed score threshold. Adding extra peptide sequences that have little chance of being present in the sample reduces the sensitivity of database search, and Blakeley *et al.* [58] suggest gene prediction tools, including Augustus, as a method to combat database inflation. This issue is present, albeit to a lesser extent, in the U90CYP as well.

Additionally, 26% of the peptides in the FHMP are not found in the 6-frame translation of the genome at all. This is likely due to peptides spanning exon-exon junctions, as there are roughly 10 'non-genomic' peptides for each protein, a little above the average of approximately 9 introns per coding genes found in teleost genomes.[98]

Thus, while BUSCO analysis suggests that the FHMP is an incomplete representation of the true FHM proteome, it is significantly more accurate than more complete protein databases, such as the U90CYP, and better suited to database search than a 6-frame translation of the genome.

Protein Search Results comparison

To confirm that the FHMP performed better than the U90CYP and ZRP, proteins identification were performed with the TPP using both Comet and X!Tandem search engines. Summary results of the search can be found in Figure 2.8, and are consistent with the observed differences in the database. A total of 3,237, 3,190, and 3698 proteins were identified when using the U90CYP, ZRP and FHMP respectively. While U90CYP is larger and more complete by both DOGMA and BUSCO metrics (Figure 2.2), 7455 (19.5%) more peptides, and 461 (14.2%) more proteins were identified when the FHMP database was used to search the same data at the same FDR.

Database PSM score comparison

The increased protein identifications are due to a variety of factors. Figure 2.9 shows the distribution of peptide-spectrum match (PSM) scores calculated by Comet after searching against the FHMP and

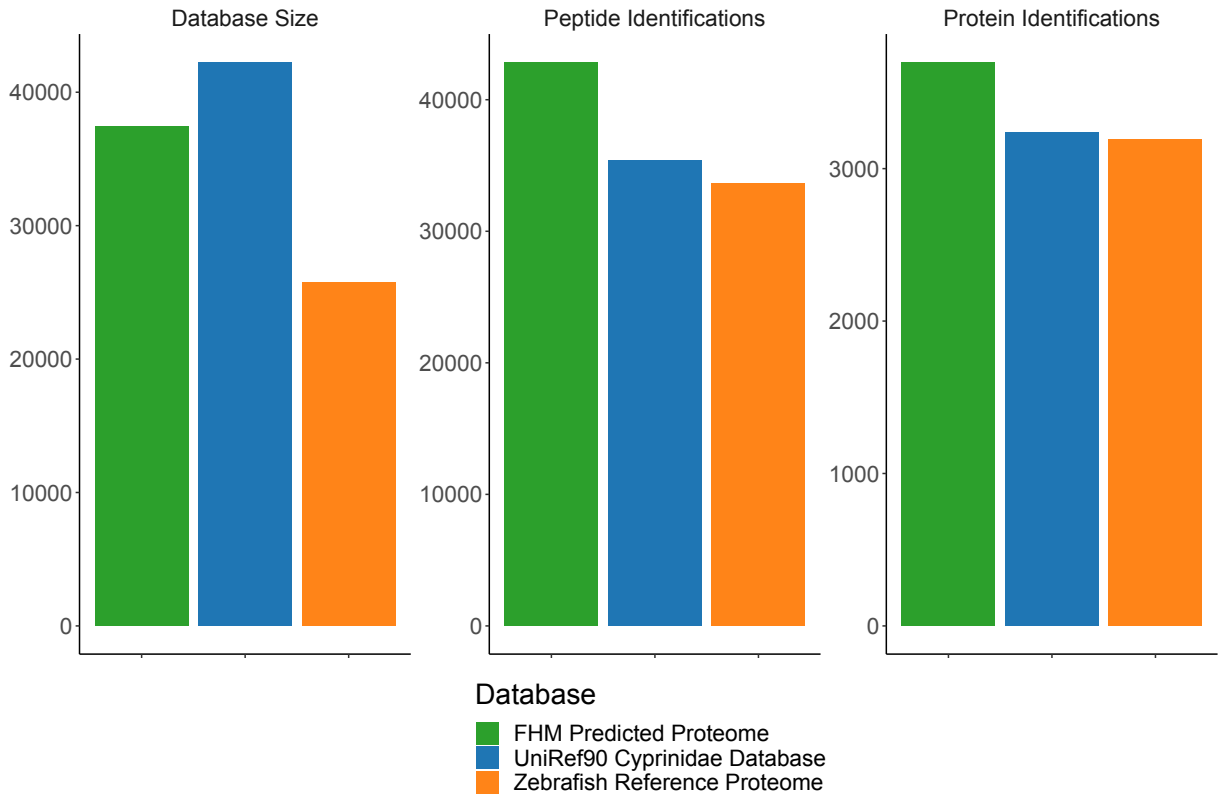


Figure 2.8: Total size of the UniRef90 Cyprinidae Proteome (U90CYP), Uniprot Zebrafish Reference Proteome (ZRP), and Fathead Minnow Predicted Proteome (FHMP) databases and the quantity of unique proteins and peptides identified from each database.

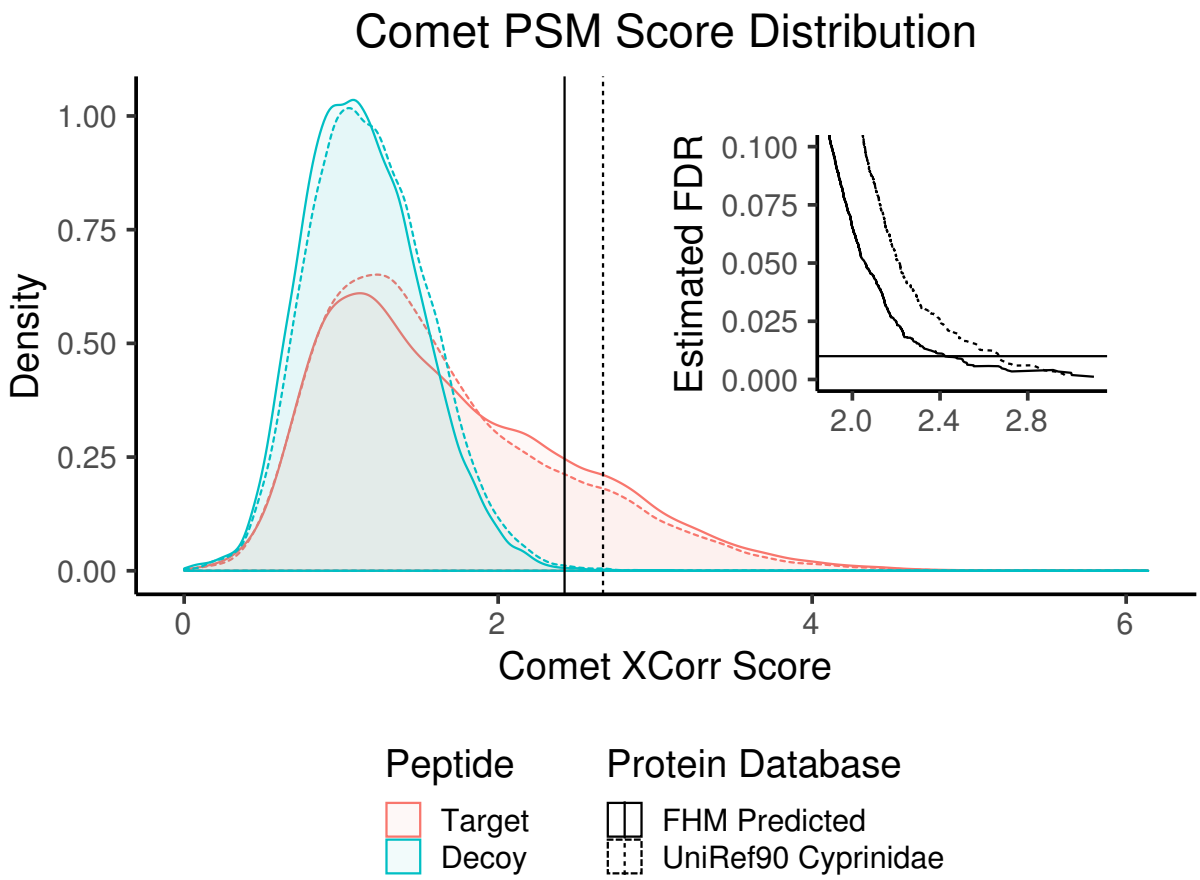


Figure 2.9: Comet XCorr score distribution from the same mass spectrometry run after separate searches against the decoyed FHMP (solid) and U90CYP (dashed) protein databases. The distribution of decoy peptides scores is indicated in blue, and target peptide scores in pink. The vertical lines indicate the score cut-off for an estimated 1% false positive (FP) rate in each distribution. The inset shows the estimated FDR for the two different score distributions.

U90CYP databases for one mass spectrometry run containing 50,361 spectra. In both searches the total number of matched spectra is similar, 48,688 for the FHMP versus 48,701 for the U90CYP. However, while the distribution of target and decoy peptide scores follows a similar pattern for both databases, the target peptide scores in the FHMP are more positively skewed than in the U90CYP. 660 more target spectra were also identified when searching the FHMP than U90CYP, and the number of matching ions and the total mass difference between theoretical and observed spectra were slightly higher and lower, respectively (Table A.1). This suggests higher similarity between the experimental and theoretical spectra when identifying PSMs, consistent with the expectation that the FHMP contains more accurate peptides sequences based on the greater overlap in *in-silico* tryptic peptides (Figure 2.7).

The mean decoy peptide score was also slightly higher after searching the U90CYP than in the FHMP

(Figure 2.9 and A.1). Since the score cutoff is calculated from the overlap between high scoring and target peptides, increased decoy scores reduces the sensitivity of the search by increasing the required score threshold for a given FDR. While the individual difference is small, the combined effect of decreased target score and increased decoy score reduces the number of PSMs passing a 1% FDR cutoff. Despite slightly more PSMs being generated when searching against the U90CYP, only 4052 PSMs exceed the 1% FP score cutoff, while there are 5822 PSMs over the score cutoff after searching against the FHMP.

Since the U90CYP and FHMP were used for searching against the same mass spectrometry run, the correlation in scores for particular spectra can also be compared. As expected, when the same target peptide is identified from both databases, the Comet scores produced are identical (Figure 2.10a). Similarly, when both databases produce a decoy hit, the resulting score is very similar, though the FHMP tends to score slightly higher for the same scan (Figure A.4).

However, when the search engine identifies different peptides for the two databases for the same spectra, the FHMP tends to score significantly higher. Figure 2.10b shows the score correlation when two different target peptides are identified for the same spectra. For a portion of the hits, the score is essentially identical, as in Figure 2.10a, likely due to the presence of peptide sequences with minor differences, such as leucine to isoleucine variants, which produce indistinguishable or highly similar fragmentation spectra. For dissimilar peptides the FHMP tends to score higher, and has more peptides passing the 1% FDR score-cutoff, consistent with the higher accuracy expected from the FHMP.

The largest set of spectra was those in which a decoy peptide was identified from one database, and a target peptide from the other (Figures 2.10c and 2.10d). The FHMP produces 1,866 more PSM identifications from the target set than the U90CYP, and a greater proportion of the PSMs score over the 1% FDR score threshold. However, there are still many high-scoring target PSM matches from the U90CYP that don't have a corresponding high-scoring match in the FHMP database (Figures 2.10b and 2.10d). In these cases, the correct peptide sequence is likely missing from the FHMP as both the genome and the annotations are incomplete (Figure 2.2a). Figure 2.7 also shows there are nearly 80,000 peptides sequences that are present in both the U90CYP and the FHM genome but not present in the FHMP, suggesting that more than 5% of the total set of peptides is missing. Since many of these spectra score over the 1% FP rate score cutoff when searched against the U90CYP database, they may be biologically relevant peptides that are not being identified due to the incompleteness of the FHMP.

Despite the number of potentially missing spectra, using the FHMP still produces more peptide and protein identifications than the U90CYP database.

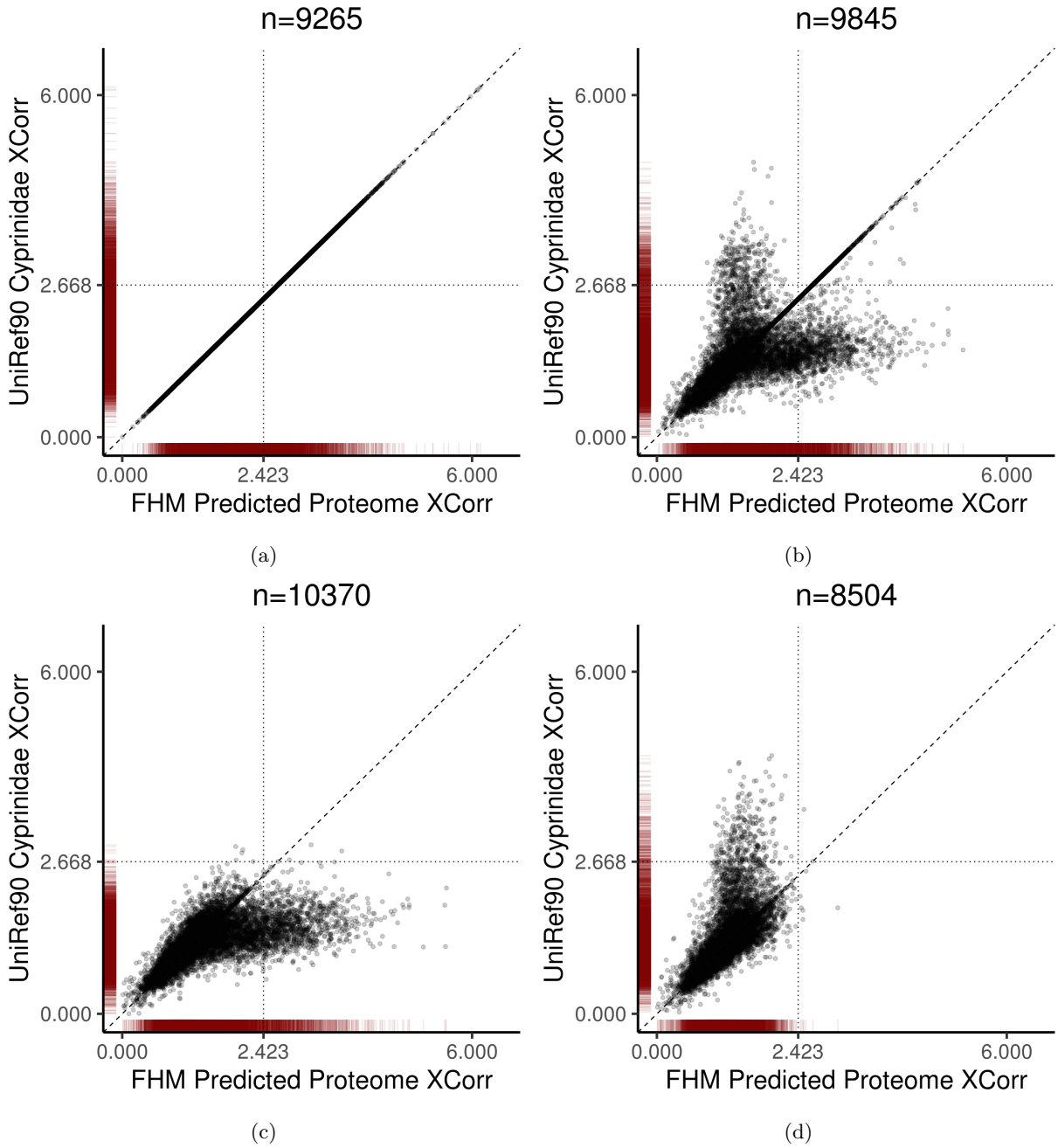


Figure 2.10: Comparison of PSM scores between the FHMP and U90CYP databases, from the same mass spectrometry run. a) PSMs with identical peptide sequences, b) PSMs which matched to different target peptides in the different databases, c) PSMs that matched a target peptide in the FHMP and a decoy in the U90CYP and d PSMs that matched to a decoy peptide in the FHMP and a target peptide in the U90CYP. The number of PSMs in each plot are indicated above, the dashed diagonal line indicates the 1:1 ratio between scores, and the dotted horizontal and vertical line indicates the 1% FDR score cutoff for the U90CYPand the FHMP, respectively.

Table 2.2: Search method summary for the tools used.

Name	Search Method
MSFragger	Open-window database-spectra matching
Comet	Narrow-window database-spectra matching
X!Tandem	Narrow-window database-spectra matching
PEAKS 8.5	De novo spectra sequencing, database search

2.3.2 Search Engine Comparison

In addition to comparing different protein database sources, the effectiveness of different search engines for identifying proteins and peptides from mass spectrometry data was compared. Search engines employ different methods for identifying peptides from mass spectra, which are summarized in Table 2.2. Comet and X!Tandem are both long-standing tools still commonly used for shotgun proteomics, which rely on matching observed MS2 fragmentation spectra to theoretical spectra from peptides in a database.[99,100] To reduce the amount of theoretical fragment spectra to be searched, the precursor intensity from the MS1 spectra is used to select potential peptides for spectra search. MSFragger is a newer tool which uses novel indexing of theoretical fragmentation patterns to improve the search speed, allowing orders of magnitude increases in speed to search a wider range of theoretical spectra and increase the number of high scoring peptide identifications.[56]. Finally, PEAKS 8.5 is a commonly used commercial software that performs *de novo* peptide sequencing from the MS2 spectra and then performs a database search with the *de novo* peptides to identify proteins.[50]

The total PSMs, peptides, and proteins identified by each tool from each sample can be found in Table 2.3. With the exception of MSFragger, the number of PSMs per input spectra was highest in the Trout sample, and lower in the Human and FHM searches. Generally, less than 50% of input spectra will be matched to a peptide during database search.[101] The Trout spectra acquisition was performed on a Q-Exactive, while the other samples were both acquired on a Fusion Lumos mass spectrometer in SPS mode (See Table 2.1). This means that the Trout MS2 spectra were generated in the higher accuracy Orbitrap, rather than the lower accuracy ion trap.[66] The Human and FHM samples traded lower accuracy fragmentation spectra, which reduced the rate of PSMs identification, for more accurate reporter ion spectra for quantification. However, despite the reduced number of spectra, the number of unique peptides and proteins identified was higher in the Human sample, possibly due to the simpler extraction and preparation methods for isolating protein from human cell cultures compared to fish liver samples.

In all datasets, Comet identified the largest number of proteins, and only lagged behind PEAKS for the number of unique peptides in the Trout and FHM datasets where the protein database is less complete. Generally, MSFragger performed the worst, falling behind the narrow-window search engines X!Tandem

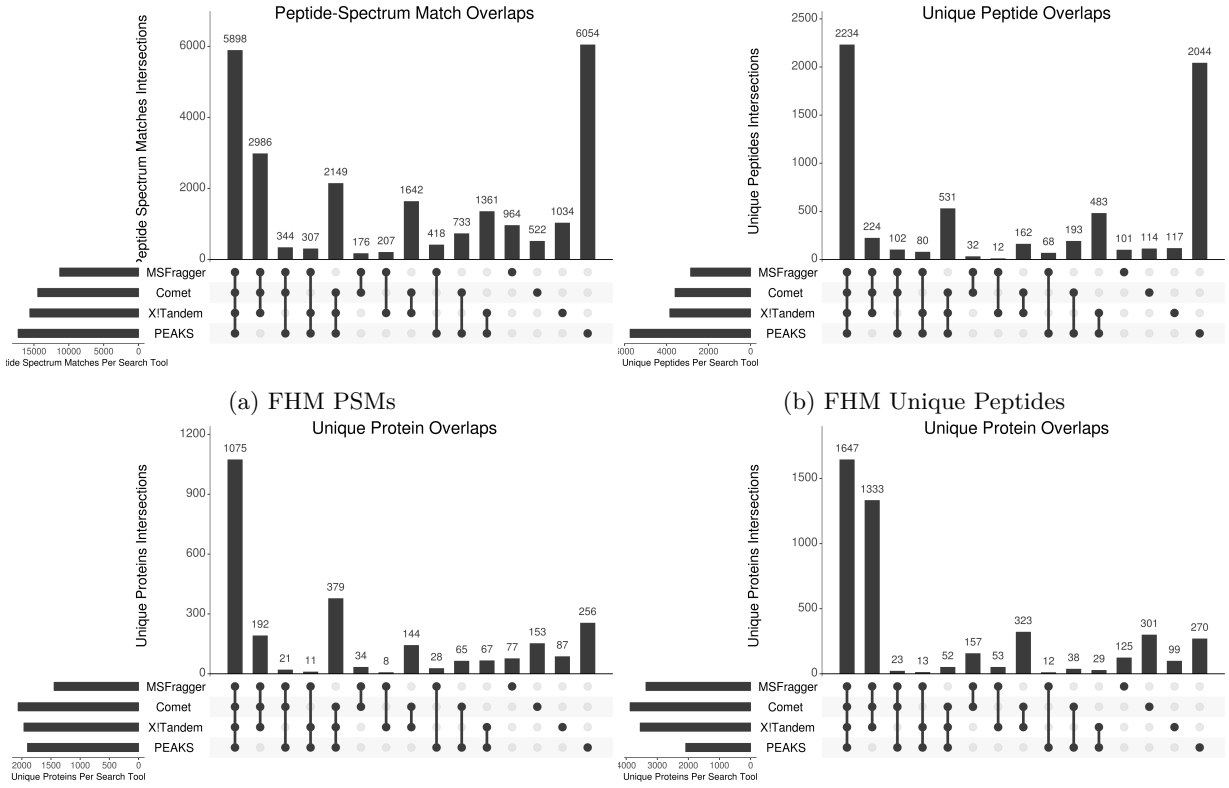
Table 2.3: Number of PSMs, peptides, and proteins identified by each search engine in each sample. For PSMs the percentage of ID'd spectra is indicated in brackets.

		Sample					
	Search Engine	Human		Trout		FHM	
PSMs	MSFragger	28,914	(20%)	11,965	(10%)	11,300	(8%)
	Comet	36,936	(25%)	50,638	(41%)	14,450	(10%)
	X!Tandem	32,377	(22%)	34,487	(28%)	15,584	(11%)
	PEAKS	24,071	(16%)	51,527	(42%)	17,264	(12%)
Unique Peptides	MSFragger	18,632		3,211		2,853	
	Comet	23,604		11,786		3,592	
	X!Tandem	21,950		8,184		3,843	
	PEAKS	16,494		13,152		5,735	
Unique Proteins	MSFragger	3,363		817		1,446	
	Comet	3,874		2,048		2,063	
	X!Tandem	3,549		1,636		1,963	
	PEAKS	2,084		2,010		1,902	

and Comet in all cases. The poor performance of MSFragger is surprising, since the primary difference between it and the narrow window search engines is not the method for identifying and scoring PSMs, but the number of theoretical fragment spectra it compares to. As such, the search results should be at worst relatively similar to the Comet and X!Tandem results. In the Trout dataset the performance of MSFragger was particularly poor, identifying only 817 proteins and less than 1/5th the PSMs of PEAKS and Comet. Given the higher accuracy of the spectra, this is contrary to both expectation and the results of the other search engines. The relatively novelty and ongoing development of MSFragger at the time of testing, and the inverse effect of spectra accuracy on the frequency of PSMs, suggests there may be software modifications or setting specific to particular mass spectrometers or samples that need to be adjusted to effectively identify proteins with MSFragger.

Since it is possible to incorporate search results from different software (e.g. with iProphet), the choice of search engine is not limited to a single tool, but to the set of tools that are possible to combine. Since the protein identifications were performed using the same protein database and spectra for each dataset, the overlap in PSMs, unique peptides, and proteins could also be compared. Figures 2.11 and A.5 show the degree of overlap within each set of features. Notably, all search engines identify a core set of proteins comprising approximately $\frac{1}{2}$ of all unique protein identifications. Additionally, each search engine identifies a unique set of proteins that are not found by the other search engines (Figures 2.11c, 2.11d, and A.5e).

When searching the Human sample, for which the most comprehensive proteome is available, all database-spectra matching tools outperformed the *de novo* PEAKS method, identifying up to 50% more spectra and nearly twice as many proteins (Table 2.3). With a comprehensive and accurate set of proteins



(a) FHM PSMs

(b) FHM Unique Peptides

(c) FHM Unique Proteins

(d) Human Unique Proteins

Figure 2.11: Upset diagrams of the overlap between a) FHM PSMs, b) unique peptide, c) protein IDs, and d) Human Protein IDs). The vertical bar chart shows the size of each overlapping set indicated by the circles underneath (e.g. in c), 192 identical proteins were identified MSFragger, Comet, and X!Tandem). The horizontal bars indicates the total quantity of each feature identified by that search engine.

to match against, database search is simpler and more accurate, particularly with the lower accuracy mass spectra produced from the ion trap. By contrast, in both the FHM and Trout samples, PEAKS identified the most PSMs and unique peptides. Comparing the overlap between the matched scan and peptide sequence for the FHM (Figures 2.11a and 2.11b) and Trout (Figures A.5b, 2.11b) samples shows that PEAKS *de novo* sequencing identified a large set of PSMs and peptides that the three database-spectra matching tools could not identify, and a smaller set of unique proteins and peptides in Human samples (Figures A.5a and A.5c. Since PEAKS doesn't rely on the presence of the peptides in a database, the less comprehensive proteomes available for Trout, and FHM in particular, weren't as limiting on PSMs identification rates. Despite identifying more peptide sequences, PEAKS produced fewer unique proteins than Comet in all samples, and only surpassed X!Tandem in the Trout sample. Since there is no corresponding protein in the database, the unique peptides did not contribute to increased protein identifications.

This highlights both a strength and weakness of the PEAKS *de novo* method - while it is able to identify

a large number of *de novo* peptides in the absence of a comprehensive database, the peptides still need to be compared to some other set of sequences to identify proteins. While this could potentially identify proteins not present in the original database, e.g. by BLAST search against a larger collection, calculating FDR and quantifying expression become more complicated. Additionally, if searches are performed against continuously updating online collections such as UniprotKB later replication of experiment results becomes more difficult because the database may change between runs.

2.4 Conclusions

Due to the lack of a comprehensive protein database containing fathead minnow specific proteins, I have described a method for database construction from genome sequence and predicted gene locations. This database contains a set of peptides that are not found in Uniprot, and thus represents species specific peptides more likely to be found in fathead minnows, while containing a significant proportion of the gene sequences and protein domains that are expected in a ray-finned fish. The reduced database size and redundancy has increased the sensitivity of mass spectrometry based protein identification using a target-decoy method for FDR calculations.

The genome-derived protein database was assessed on the basis of single-copy orthologs, the set of potential tryptic peptides, and the degree of redundancy within the database. BUSCO analysis showed that the genome-derived database contained the majority of single-copy orthologs found in the genome. The set of peptides derived from the genome predictions was a subset of those found in a 6-frame translation of the FHM genome, and was similar in size in those found in the Zebrafish Reference Proteome. However, only a small portion of the peptides were found in the U90CYP, indicating the protein database is species specific. Finally, the initial database constructed was highly redundant, but clustering with cd-hit significantly increased the fraction of unique peptides without significant effect on the set of potential peptides sequences or predicted single-copy orthologs.

The genome-derived database is limited primarily by the quality of the gene predictions. Despite a large number of aligned Zebrafish gene transcripts, the majority of the complete protein sequences are derived from the Augustus gene predictions. While the majority of U90CYP tryptic peptides are not present on the FHM genome 6-frame translation, there is still a significant number that are present, but missing in the FHMP. Due to incomplete genome and annotations, the set of FHM protein sequences is incomplete as well, and some biologically significant MS spectra may be missed during protein identification.

However, the FHMP still outperformed the U90CYP protein database for protein identification from the same MS dataset. The FHMP produced more protein identification at the same error rate, and more high quality PSMs and peptides.

Comparison of different search tools showed that de novo sequencing identifies peptides and proteins not found by spectra matching programs, but database-spectra matching tools generally perform better when a comprehensive database is available. While PEAKS was able to identify more PSMs and peptides in the FHM and Trout datasets, the lack of protein sequence information still prevented more proteins from being identified. While *de novo* peptide information could be used for further protein identification, it would increase the complexity of downstream analysis. This further highlights the importance of effective database construction for protein identifications.

Comparison of the set of protein IDs from each search tool also showed that unique proteins were identified from all search tools. This means that a single best search tool is unlikely to exist, and integrating results from multiple search engines is important for achieving best coverage. For the purpose of analyzing the FHM proteome for quantitative analysis, the Comet and X!Tandem search engines will be used, as both can be integrated effectively with established tools, specifically, the Trans-Proteomic Pipeline (TPP).

Chapter 3

Hepatic Proteome Analysis of Fathead Minnows in the Bow River

3.1 Introduction

With an estimated 300-500L of wastewater generated per capita, per day in Canada, municipal wastewater effluent (MWWE) is one of the primary sources for anthropogenic chemical compounds in aquatic environments[1]. The resulting effluents are a complex mixture of chemicals, potentially including polyacyclic aromatic hydrocarbons (PAHs), metals, pharmaceutical and personal care products (PPCPs), and a wide variety of other contaminants. PPCPs in particular are of growing concern, as the efficacy of wastewater treatment plants (WWTPs) in removing PPCPs can vary dramatically with treatment method and waste source. Some pharmaceuticals may be resistant to degradation, causing them to persist in aquatic environments[11,102]. PPCPs also encompass a wide variety of chemical compounds designed to be biological active at low concentrations, and the effects of exposure on aquatic organisms is unknown[2].

Metabolic changes in aquatic organisms have often been found after exposure to MWWE or its constituent components. The liver is the site of a wide variety of lipid and glucose metabolism functions important for energy regulation and mobilization.[103] In fish, the liver can store up to 10-20% of total body lipids and is the primary site for the synthesis of lipids and lipoproteins, making it an important target for contaminants that affect lipid metabolism, such as statin and fibrate pharmaceuticals and a wide variety of other organic pollutants.[10,104,105] In particular, the peroxisome proliferator-activated receptors (PPARs) regulate a wide variety of metabolic processes, such as fatty acid oxidation and elongation, glucose metabolism, and lipogenesis.[22] PPARs are known to be sensitive to exposure to a variety

contaminants[9] and peroxisome proliferation has been proposed as a biomarker for xenobiotic exposure.[106]

The liver is also a key site for biotransformation of xenobiotics, and is the primary site for expression of cytochrome P450 isozymes, which target a wide range of xenobiotics for biotransformation and excretion.[107] Biotransformation increases the oxidative stress the cell undergoes due to increased production of reactive oxygen species in the peroxisome. The large number of processes for energy usage, stress response, and homeostasis regulation taking place in the liver make it a key site for identifying potential biomarkers of exposure to environmental contaminants.[27,104]

One known response to MWWE exposure is impairment of the cortisol mediated stress response in fish, leading to a chronic stress condition.[35,42] Cortisol regulates a variety of processes as part of stress response by binding to glucocorticoid receptor (GR), a transcription factors which regulates a wide range of processes.[33] One of the primary targets of cortisol action is in hepatocytes, where it regulates a variety of pathways to control energy usage during the stress response and maintain homeostasis. Cortisol signaling in hepatocytes alters expressions of many energy related genes, such as phosphenolpyruvate carboxykinase (PEPCK), a key enzyme in gluconeogenesis; as well as immune related genes such as suppressors of cytokine signaling (SOCS) and Janus kinase-Signal Transducer and Activator of Transcription (JAKS-STAT) related genes.[14] Altering these functions is an important mechanism allowing fish to respond to stressors in the environment and return to homeostasis.[33]

While many studies have investigate the impact of single contaminants on the liver, less has been done to study impacts of chronic exposure on organisms in complex mixtures that occur in aquantic environments.[12] Exposure to low levels of contaminants in complex mixtures can have a variety of interacting, often sub-lethal effects, altering behavior and a wide range of biological functions in the organism, ultimately reducing fitness.[2,15] Studies of exposure to complex mixtures in natural environments often focus on reproductive endpoints and impacts, but exposure to complex contaminant mixtures will impact a wide variety of pathways, including metabolic and energy related pathways in the liver.[12,42] While not directly lethal, impacts on these function can reduce the ability of aquatic organisms to respond to stressors, negatively impacting their fitness.[35]

The purpose of the study presented in this chapter was to investigate the changes that occur in the hepatic proteome of fathead minnow (*Pimephales promelas*) caged in sites upstream and downstream of the three WWTPs along the Bow River in Calgary, Alberta, Canada. This was performed as part of a larger project investigating the impacts of MWWE exposure on stress response and reproductive health, the mechanisms of it effects, and potential biomarkers for MWWE exposure. Shotgun proteomics and isobaric mass tagging were used to identify changes in proteins between the different sites. Enriched pathways and molecular functions were identified in the differentially expressed proteins, and gene set enrichment analysis was used to compare expression profiles between sites.

3.2 Methods

3.2.1 Exposure and Sampling

Field exposure and sampling was carried out alongside the field stress experiment described in Lazaro-Côté *et al.* [12]. Briefly, a caging study in the Bow River, Calgary, Alberta, was carried out from September to October 2016, using adult fathead minnows in 5 sites across the Calgary urban gradient (Fig. 3.1). Two sites, Bearspaw Dam (BEAR) and Cushing Bridge (CUSH) are located upstream of the 3 WWTPs in Calgary, and the other three, Glenmore (GLEN), Highway 22X (H22X), and Highwood River (UPHI) are located downstream. Of the three WWTPs in Calgary, Bonnybrook produces the largest volume, releasing 4.5 m^3/s of MWWE, 8.3% of the flow of the Bow River. Fish Creek and Pine Creek are smaller plants, releasing 0.4 m^3/s (0.6%)and 1.1 m^3/s (1.9%), respectively.[12] All three WWTPs perform tertiary treatment of MWWE, employing ultraviolet light for disinfection. Bonnybrook and Pine Creek perform biological removal of nitrogen and phosphorous, while Fish Creek uses chemical treatment to remove phosphorous.[108]

While the BEAR site is located upstream of urban runoff, CUSH receives urban and rural runoff from more than 100 storm water outfalls as well as Nose Creek. The GLEN site is located 15 m downstream of the Bonnybrook WWTP outflow, with minimal mixing, while H22X is 250 m downstream of Fish Creek WWTP, and H22X 18 km downstream of the Pine Creek outflow. Water quality parameters at the sites can been seen in Table B.2. The mean water temperatures at the GLEN site (18.4 °C) was significantly higher than all other sites (12.0–12.5 °C), and was conductivity was approximately 3-fold higher (928.5 µS/cm compared to 298.3–397.4 µS/cm). Dissolved oxygen concentrations and pH were significantly lower, 6.06 mg/L and 7.23 compared to 9.63–11.04 mg/L and 8.00–8.25, respectively.[12]

Adult fathead minnows were obtained from the University of Lethbridge, and acclimatized in-laboratory to Bow River temperatures for two weeks. 35-36 adult fathead minnows were then randomly assigned to each of the two cages at each site for 26 days. Galvanized steel minnow traps were used as cages, with a 150 mm by 100 mm piece of PVC pipe provided as shelter. Cages were cleaned every two days to avoid obstruction of water flow through the cages, and no food was provided to the fish. Temperature, dissolved oxygen, conductivity, pH, and turbidity were measured at each site prior to cleaning.

Following exposure fish were sacrificed and dissected on site on the day of collection between 9:30am and 11:00am. A subset of 5-10 fish per cage (11-19 fish per site) were collected from each cage and sacrificed by immersion in 1 g/L MS-222 buffered with sodium bicarbonate. Carcasses were weighed and the fork length measured, then immediately dissected. Head kidney, gonads, and liver tissue were collected and stored at approximately −80 °C. A total of 30 liver samples, 6 liver samples from each site, were used for proteome analysis, with an overall sex ratio of approximately 1:1 (Table B.3).

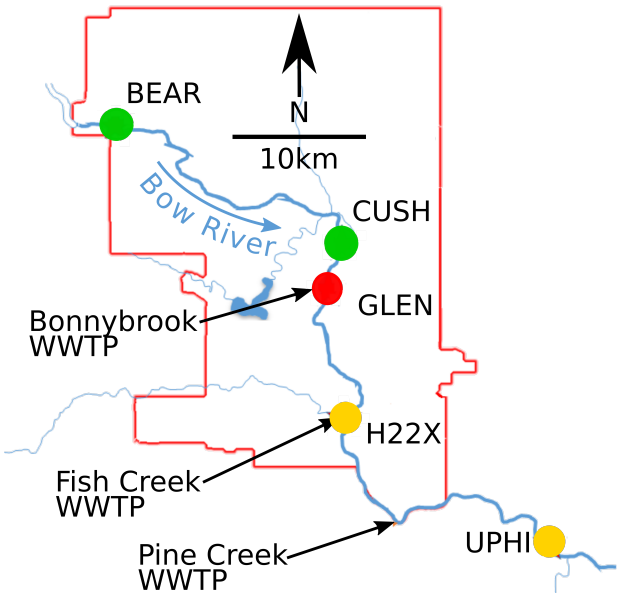


Figure 3.1: Locations of the 5 caging sites along the Bow River. WWTP locations are indicated by arrows and the Calgary city limits are indicated in red.

3.2.2 Proteomic Analysis

Protein Extraction, Labeling, and Mass Spectrometry

Liver samples were processed using a modified version of filter aided sample preparation[109] previously implemented in the McConkey Lab. Briefly, each sample was lysed in a buffer of 4% (w/v) SDS, 100 mM HEPES/HCl pH 7.6, 100 mM 1,4-Dithiothretiol, and MS-SAFE protease inhibitor (Sigma-Aldrich Canada Co., Oakville, ON, Canada, Catalog #MSSAFE-1VL). Samples were manually homogenized on ice using a plastic pestle prewashed with methanol, then incubated at 95 °C for 5 min, sonicated in 10 s pulses for 1 min, followed by another 5 min incubation and 95 °C for 5 min, and a 10 min centrifugation at 14000x rcf to clarify the extract. Proteins were then precipitated using a Calbiochem Protein Precipitation kit (Millipore Canada, Etobicoke, ON, Canada, Catalog #539180-1KIT) following the manufacturers instruction, and the pellet allowed to dry at room temperature for 5 min. The pellet was then resuspended and solubilized in 200 µL of 8 M urea in 100 mM HEPES/HCl pH 8.5 (UA) until completely resuspended.

Once resuspension had completed, protein was loaded into a Amicon Ultra 3kDa 1.5mL centrifugal filter (Millipore Canada, Etobicoke, ON, Canada, Catalog #UFC500396) for digestion. Filters were washed immediately before use with two volumes 500 µL of 60% methanol, then two washes with 500 µL of ultrapure H_2O. 100 µg of protein was added to each filter, and washed with 200 µL of UA, followed by a 20 min incubation with 200 µL of 50 mM iodoacetamide. This was followed by three more washes with 200 µL of

UA, then 3 washes with 200 µL of 100 mM triethylammonium bicarbonate (TEAB). Proteins were digested overnight at 37 °C with 0.2 µg/µL of trypsin/LysC (Promega, Madison, WI, Catalog #V5073), with tubes wrapped in parafilm to reduce evaporation.

After digestion, peptides were eluted twice by adding 60 µL of 100 mM TEAB to the filter, followed by centrifugation at 30 min at 14 000 rcf. Peptide concentration was measured using a Pierce Quantitative Colorimetric Peptide Assay (Thermo Fisher Scientific, Rockford, IL, Catalog #23275) in triplicate. For each sample run, 80 µg was transferred to a clean microfuge tube, and the sample volume adjusted to 100 µL. Tandem Mass Tags (TMT) labeling was performed according to manufacturer's instructions, with labels randomly assigned to each sample in the run. Samples were then combined into one tube and dried under a vacuum, then stored at −20 °C prior to mass spectrometry analysis.

Mass Spectrometry

Samples were sent to the SPARC BioCentre Molecular Analysis facility for mass spectrometry analysis on the Thermo Orbitrap Fusion-Lumos in SPS (MS3) mode. Three separate mass 1 hour spectrometry runs were performed, each containing 10 different samples, two from each site.

3.2.3 Protein Identification

Database Construction

Due to lack of a fathead minnow-specific protein database, fathead minnow genome sequence data[59] was downloaded from the SETAC website[85], along with annotations for Augustus-predicted gene sequences and aligned zebrafish transcripts with corresponding Ensembl IDs.[60] Protein database construction is described in detail in Section 2.2.1. Before searching, randomized decoy proteins were added to the database using the Trans-Proteomic Pipeline (TPP) decoy generation script[95].

Database Search

RAW data files were converted to mzML using ProteoWizard[110]. Database searches were performed using Comet[55] and the TPP-specific X!Tandem[95] against the previously described Fathead Minnow Predicted Proteome (FHMP) database, with added randomized decoy proteins. For both search engines, the enzyme was set to trypsin, with one missed cleavages allowed for each peptide. Oxidation of methionine was set as a variable modification, and carbamidomethylation of cysteine and the TMT balance tag of 229.1629 on N-terminal and lysine residues as static modifications. Ions in the m/z range of 125 to 132 were ignored to prevent TMT reporter ions from being included in the peptide identification. Monoisotopic masses were

used for the search for both peptide and fragment ions, with a mass tolerance of 15 ppm for the parent peptide and 0.6 Da for the fragment.

PeptideProphet[51] was then used to assign confidence scores to the peptide IDs in each run. Results from each search engine across the three runs were integrated using iProphet[53], and protein IDs and groups were assigned and given a confidence score using ProteinProphet[54]. The minimum protein probability was set at 0.7552, equivalent to an estimated false discovery rate (FDR) of 1%.

3.2.4 Differential Expression Analysis

Protein Normalization and Quantification

TMT reporter ion intensities for each peptide-spectrum match (PSM) were identified using the TPP Libra module[95] individually for each mass spectrometry replicate. The full set of PSMs, along with confidence scores, raw reporter ion intensities, and the ID of the originating protein were exported from the TPP web user interface to a text file. Additionally, protein IDs from all protein groups identified at an FDR of 1% were also exported. A custom python script (Appendix B.2) was then used to filter the PSM list to remove all peptides that matched to multiple proteins. As part of filtering, the genomic region of proteins that matched an identical set of peptides were compared, and any proteins that matched to the same sequence scaffold with greater than 50% overlap were considered to be duplicate identifications of the same protein, and only ID of the longest sequence was used for filtering. Next, an in-house R script (Appendix B.1) was used for further filtering of PSMs, as well as quality control, normalization, and differential expression analysis. All PSMs with an iProphet confidence score less than 90% were discarded. Multidimensional scaling and of the peptide-level log-transformed intensities distributions were visualized with R[111] and ggplot2[112] to compare sample similarity and intensity distributions.

Next, for each mass spectrometry run, PSMs with fewer than 2 unique TMT reporter ions were discarded. A list of housekeeping genes identified from RNAseq data in human samples was matched to zebrafish symbols, and all then mapped to the fathead minnow (FHM) protein IDs. All identified PSMs matching to proteins from the housekeeping list were correlated across samples using Pearson correlation to identify potential outlier samples. The reporter ion intensity for all peptides of the same protein were summed, and normalized across the whole run with the *limma*[113] *justvsn* function (Appendix B.3). The number of PSMs and unique peptide sequences for each protein was recorded. The Pearson correlation of protein expression across all samples for protein in the housekeeping list was calculated, then the median expression for each protein was subtracted from each sample. All proteins with less than 2 unique peptide sequences were dropped from the list for each run. Finally, the protein-level quantification data was then merged by protein ID, and any proteins not present in all of the three runs were removed. Multidimensional

scaling and intensity distribution were calculated on the normalized protein log-fold change expression.

Differential Expression

Differential expression analysis was carried out with the R package *limma*. A model matrix was specified with a coefficient for each site, as well as the sex of each sample as a blocking factor. The *duplicateCorrelation* function was used to calculate correlation between biological replicates in the different mass spectrometry runs, and the expression data was fit to a linear model for each protein. The 5 sites were divided into three groups: Upstream, consisting of BEAR and CUSH; Outflow, the GLEN site; and Downstream, the H22X and UPHI sites (Fig. 3.2). Contrasts were created to compare the three groups, and the differential expression was calculated using empirical Bayes within *limma*. Proteins were considered significant with a p-value cutoff based on a 5% FDR calculated using the method of Benjamini-Hochberg[114], with multiple testing correction performed on all contrasts combined using the 'global' method of the *decideTests* function.

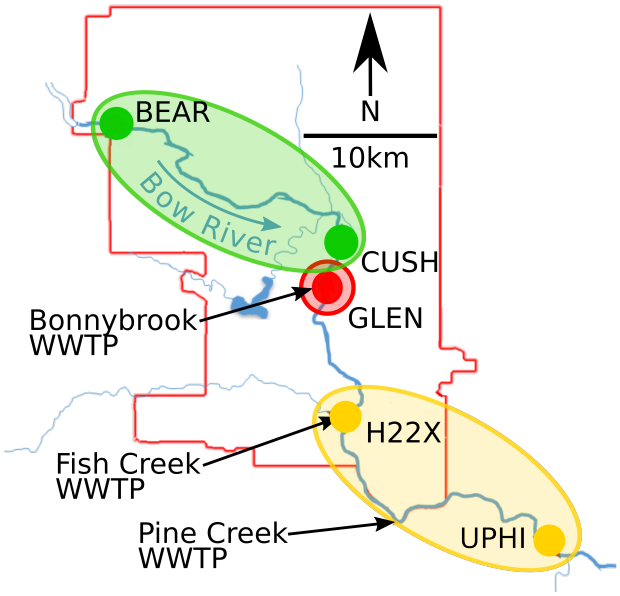


Figure 3.2: Grouping of the caging sites for differential expressions analysis. Green circle: Upstream Group, containing sites BEAR and CUSH, red circle: Outflow group, containing the GLEN site, yellow circle: Downstream group, containing the H22X and UPHI sites.

Following differential expression analysis, protein annotation information was collected from on-line resources to to better determine what proteins were differentially expressed (DE), and their putative function. Ensembl[94] IDs for each protein were searched using the R biomaRt package[115], and the gene description and GO terms associated with each gene automatically downloaded and annotated to the

protein. The proteins were then identified in the ZFIN database[116], and the ZFIN gene name was recorded for each gene. Known or predicted interactions between DE proteins were visualized into Cytoscape using StringApp[117].

Molecular Function Enrichment Analysis

Once gene symbols were identified for all the differentially expressed genes, the Ensembl *Danio rerio* transcript id was downloaded for each gene and compiled into a list. The transcript list was uploaded to the DAVID[118,119] functional annotation tool as a gene list with the ZFIN gene name as the identifier. The genes were searched for molecular functions that enriched in a statistically significant manner compared to their occurrence in the background protein list, with *Danio rerio* as the background gene set. The functional annotation chart produced by DAVID was downloaded and imported into the R *FGNet* package[120], which groups genes with molecular functions identified by DAVID as statistically enriched into overlapping clusters of shared molecular function. This allows for easier identification of particular molecular functions within the tissue sample that are changing, allowing the analysis of changes beyond the effects of single genes.

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA)[121] was performed with the normalized protein expression data for all samples. Samples were assigned phenotypes according to sampling site groups (Figure 3.2). Gene set for GO[122] Molecular Functions and Biological Process, and Reactome[123] pathways were scored for enrichment for the Outflow group vs the Upstream group, and the Downstream group vs the Upstream group. Cytoscape[124] was used to visualize enriched gene sets with an enrichment score cutoff based on a 20% FDR from either comparison.

3.2.5 Contaminant Concentration Measurements

Measurement of select PPCPs was described in previously in Lazaro-Côté *et al.* [12]. Water contaminant levels were averaged according to the group used in the analysis.

3.3 Results and Discussion

3.3.1 Fish morphometrics and mortality

The survival rate of the caged fish was 93.0% at BEAR, 90.1% at CUSH, 100% at GLEN, and 70.4% for both H22X and UPHI. Generally, survival rate was consistent the between cages at different sites, with the exception of UPHI, where the majority of losses occurred in one cage.[108] For fathead minnows used in proteome analysis, there was significant differences in the condition factor ($F_{(4,25)} = 4.7831, p = 0.0053$) between sites, but no significant differences in fish mass ($F_{(4,25)} = 1.5201, p = 0.2267$) or fork length ($F_{(4,25)} = 0.7425, p = 0.5721$) (Table B.3). Condition factor at the CUSH and GLEN sites was significantly higher than at the BEAR site, but not H22X or UPHI. No other significant differences between sites were found. Analysis by Lazaro-Côté *et al.* [12] of all 165 fish caged as part of the larger study, including the ones used for proteome analysis, did find that fish mass and fork length was significantly higher at the GLEN site compared to the other sites. Condition factor was also significantly different, higher at CUSH compared to all sites except GLEN, and higher at GLEN than at BEAR and UPHI.

Due to the limited number of fish available from the study, an even sex ratio was not possible for all sites. While CUSH, GLEN, and H22X had an equal number of male and female (3/3) fish, BEAR and UPHI had more male fish, 2/4 and 1/5, respectively. For the groups used for analysis, the ratio (F:M) was 5:7 in the Upstream group, 3:3 in the Outflow group, and 4:8 in the Downstream group.

3.3.2 Proteins identified from FHM genome annotations

For database construction, 36,911 ZFIN transcript alignments and 43,345 predicted gene sequences were present in the FHM genome annotations. After translation, a total of 62,707 protein sequences were identified; filtering redundant proteins at 90% identity with CD-HIT reduced the total database size to 37,454 putative protein sequences.

3.3.3 Run results

The database search of the three MS runs identified a combined total of 3,698 proteins with two or more peptides in any of the three runs, using an estimated FDR of 1%. From the 91,524 total PSMs identified during database search, 36,722 have an iProphet confidence score greater than 90% and uniquely match to one high-confidence protein. The distribution of Log_2-transformed intensities is relatively consistent across runs, with a mean Log_2-intensity of approximately 13 and interquartile range between 10 and 15 for most samples (Figure 3.3).

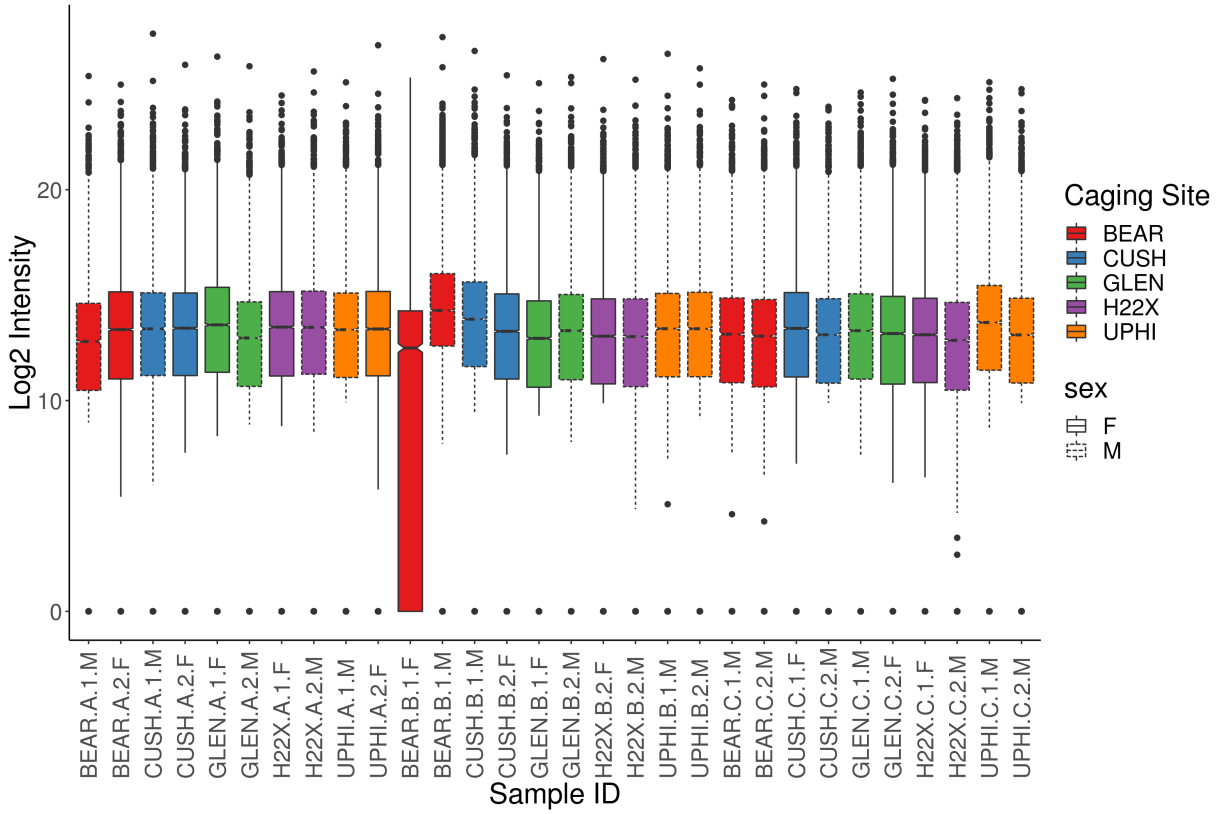
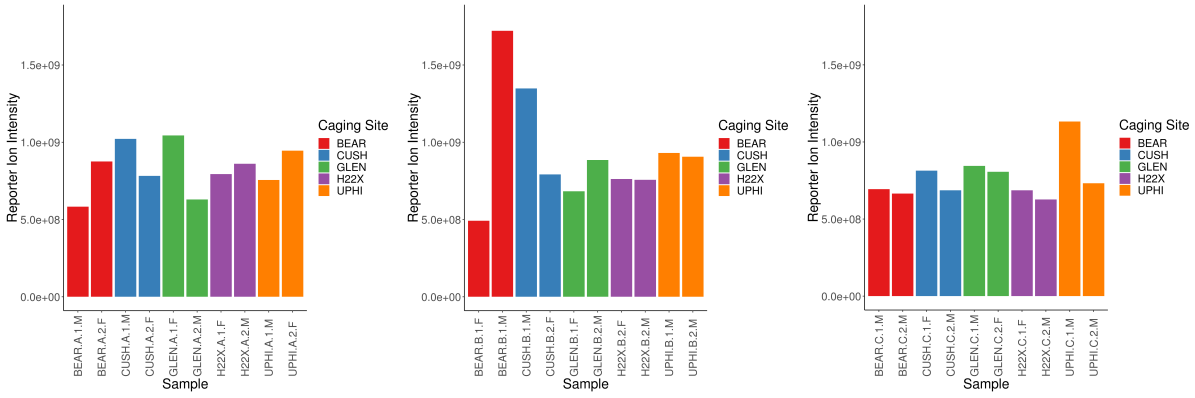


Figure 3.3: Distribution of Log_2-transformed PSMs intensities. Samples are sorted by mass spectrometry run (A-C), then caging site and cage number, as indicated in the sample label at the bottom. Color and line style indicate the caging site and sex of the sample, respectively.

In the second mass spectrometry run (run B), 1 sample (BEAR.B.1.F) shows an interquartile range starting at zero, indicating that a large number of PSMs had no reporter ions in that sample; additionally, the mean intensity is lower than other samples. Comparison of the total reporter ion intensity for the different runs (Figure 3.4) shows that while the number of ions reported for that sample was comparable to samples in other runs, two other samples in the same run have elevated reporter ion intensities. These samples would likely have been present in the run at a higher ratio than other, causing them to 'crowd-out' the other samples and causing underreporting of the samples. Alternately, the higher abundance samples may have more low-abundance peptides detected, increasing the number of zero-intensity peptides in other channels. However, discarding all PSMs with fewer than two TMT reporter ions recorded did not noticeably reduce the total ion intensity of any samples (Figure B.1), suggesting that increased low-abundance peptides from over-represented samples is not responsible for the missing reporter ions from samples BEAR.B.1.F.

Intra-run correlation of peptide intensities between samples for the list of 226 housekeeping genes



(a) (b) (c)

Figure 3.4: Sum of reporter ion intensity for the 10 reporter tags in each of the 3 mass spectrometry runs.

Table 3.1: Number of PSMs in the three runs at various stages of filtering.

Run	Total PSMs	High Confidence	>2 Reporter Ions
A	31,015	12,111	10,461
B	30,575	12,529	11,045
C	29,934	12,082	10,339

was high for all runs, with a minimum Pearson correlation of 0.89 for any pair of samples(Figure 3.5). Overall, the different mass spectrometry runs produced similar results, with consistent quantities of PSMs at different stages (Table 3.1) and all samples producing total reporter ions within a 3-fold range (Figure 3.4.

3.3.4 Normalization

After filtering, the PSM reporter ion intensities were summed by protein and normalized by variance stabilized normalization (VSN) within each run individually. Pearson Correlation of expression of housekeeping-proteins was similar between samples, ranging from 0.92 to 0.99, with the exception of sample H22X.C.2.M in run C, which had correlation scores between 0.88 and 0.94 (Figure B.8). The log fold change of protein expression also had a larger interquartile range and wider distribution than other samples in the same run (Figures B.9 and B.10). Multidimensional scaling (MDS) grouped samples by site but with sample H22XC.2.M as a clear outlier from the rest (Figure B.11). Finally, a heatmap clustering the samples by similarity of expression places the H22X.C.2.M sample as a clear outlier from all other samples (Figure B.12). The H22X.C.2.M sample is considered to be an outlier and was excluded from differential expression analysis. Since the removed sample was male, the ratio of females to males (F:M) change to 5:7, 3:3, and 4:7 in the Upstream, Outflow, and Downstream groups, respectively.

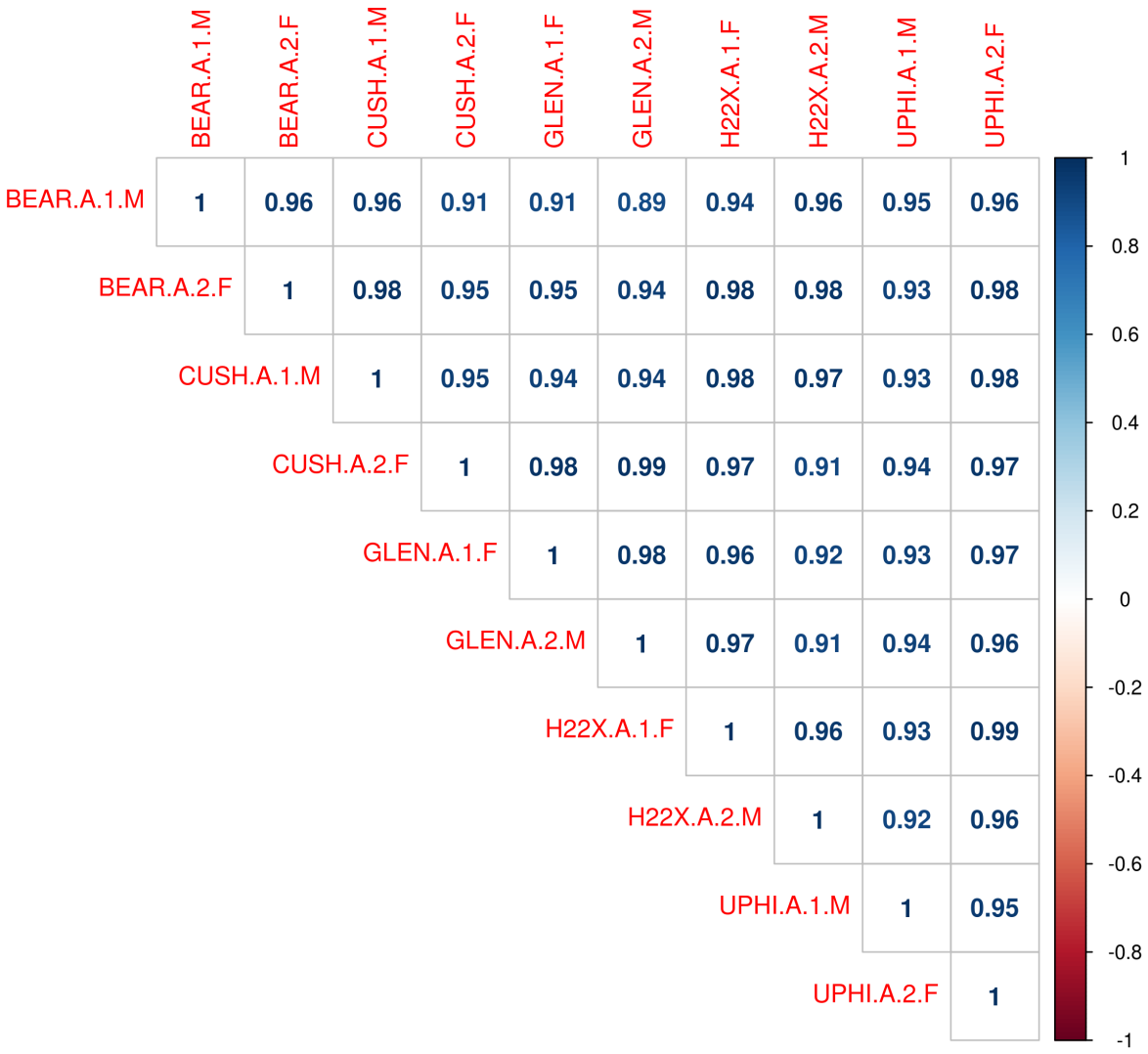


Figure 3.5: Intra-run correlation of PSM reporter ion intensities for mass spectrometry run A. Figures for runs B and C are in Appendix B.3

Filtering and normalization analysis were performed on the run C samples again, with the H22X.C.2.M sample excluded. With the outlier removed the correlation between housekeeping-protein expression with samples was consistently 0.92 or greater in all runs (Figure 3.6, and Figures B.4 and B.5 in Appendix B.3).

The most heavily effluent exposed site, GLEN, has the greatest deviation from the two upstream sites, BEAR and CUSH, with UPHI placed in between. The H22X site, even after the removal of the outlier sample, is less strongly grouped than other sites, and overlaps the BEAR, CUSH, and UPHI groups. While the samples are separated by site, there is no clear separation of samples by other factor, such as sex, or cage. The significant overlap between the BEAR and CUSH groups, and the dispersion of the UPHI group support the decision to group multiple sites together for analysis. (Figure 3.7).

3.3.5 The Hepatic proteome is altered downstream of WWTPs

For differential expression analysis, a total of 895 proteins were identified in all three runs by at least two unique peptides. 164 proteins were identified as differentially expressed in at least one of the three comparisons. Of those, 140 proteins were DE between the Upstream and Outflow groups, and 23 proteins were DE in the Downstream group compared the Upstream group. Finally, 92 protein were DE between the Outflow and Downstream groups. The overlap in differentially expressed proteins between comparisons can be seen in Appendix B.7. The full list of DE proteins is shown in Table 3.2.

Clustering of samples by expression level of DE protein (Figure 3.8) shows a similar pattern to Figure 3.7. The GLEN site samples forms a cluster with similar expression pattern showing clear differences from other sites. The CUSH, BEAR, H22X, and UPHI site samples are less clearly distinguished, however, the H22X and UPHI site samples tend to cluster together, as do the BEAR and CUSH sites. The expression changes in the BEAR and CUSH are also more strongly contrast the GLEN sites changes, while H22X and UPHI are more similar. There are no clear patterns of clustering by other factors such as sex or cage.

Pathway enrichment analysis of the differentially expressed proteins in the Outflow site (Figure 3.9) revealed that differentially expressed proteins at the outflow site were significantly enriched for lipid and amino acid metabolism, drug and xenobiotic metabolism, and RNA splicing functions (Figure 3.10). The Downstream group, with fewer DE proteins, had only four enriched pathways, comprised of amino acid and purine metabolism, drug metabolism, and spliceosome relate proteins. While differential expression in the proteome compared to the upstream groups was highest in the Outflow group, in the 18 proteins found differentially expressed in both the Outflow and Downstream groups, all proteins agree in direction of changes, and the magnitude of change is consistently reduced in the downstream group (Table 3.2). Additionally, three of the four KEGG pathways enriched in the Downstream group are also enriched at Outflow group (Figure 3.9). This suggests that there is a similar effect, but with reduced magnitude,

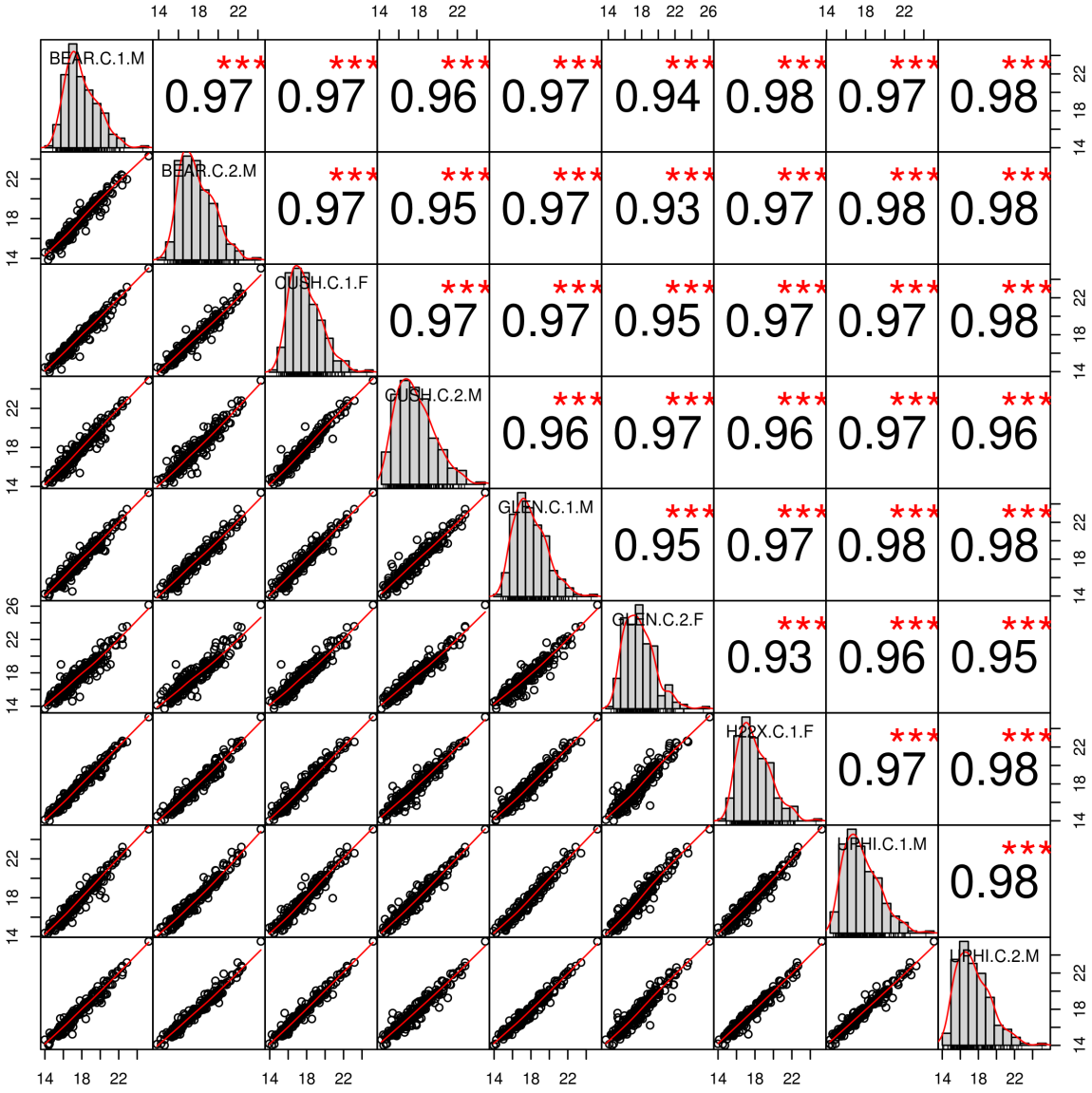


Figure 3.6: Intra-run correlation of normalized protein intensities for mass spectrometry run C. Figures for runs A and B are in Appendix B.3

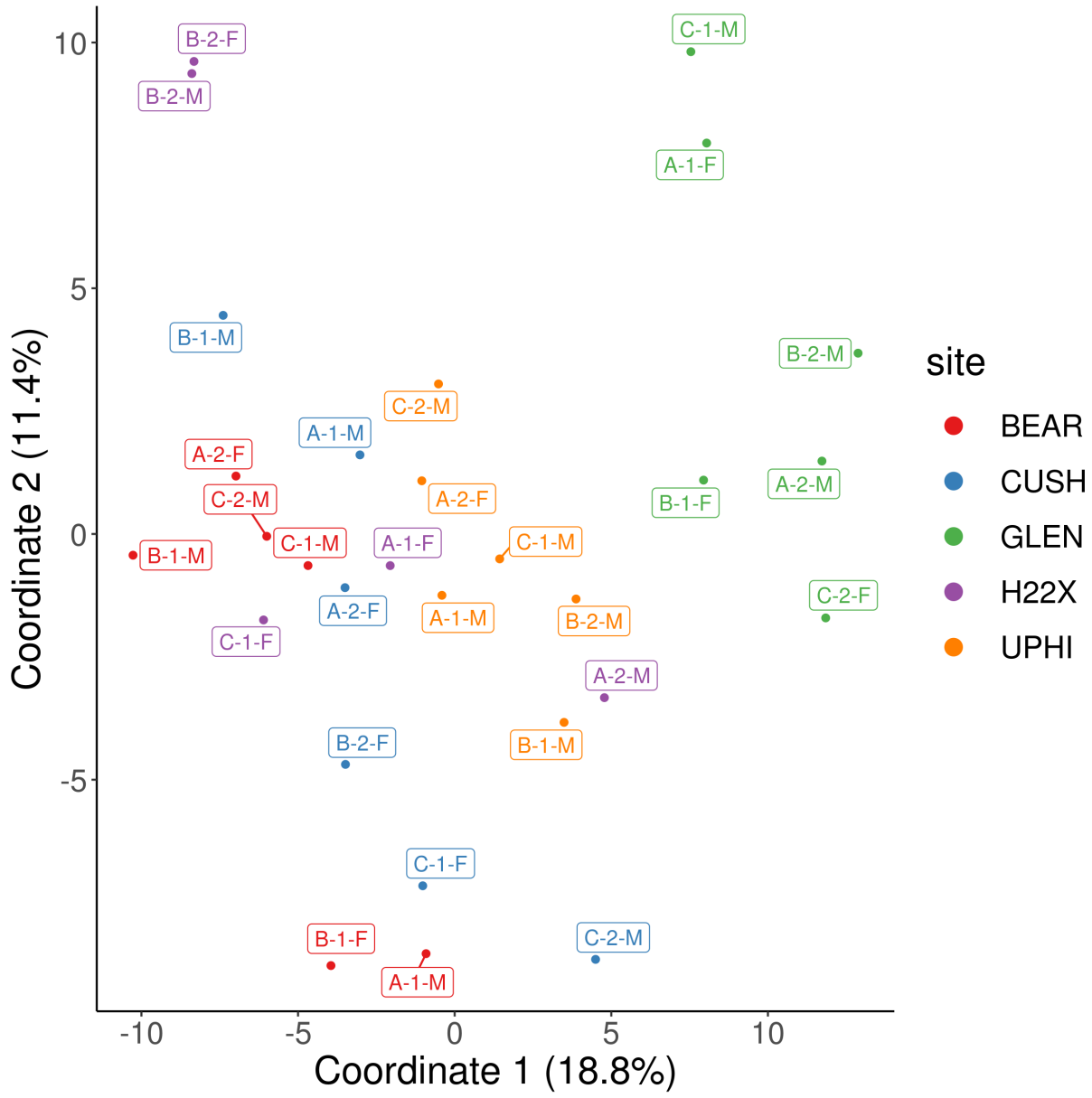


Figure 3.7: MDS plot of normalized protein Log_2 fold-change. The label next to each point indicates which of the three mass spectrometry runs (A-C), sample cage (1 or 2), and the sex (M or F) of the sample.

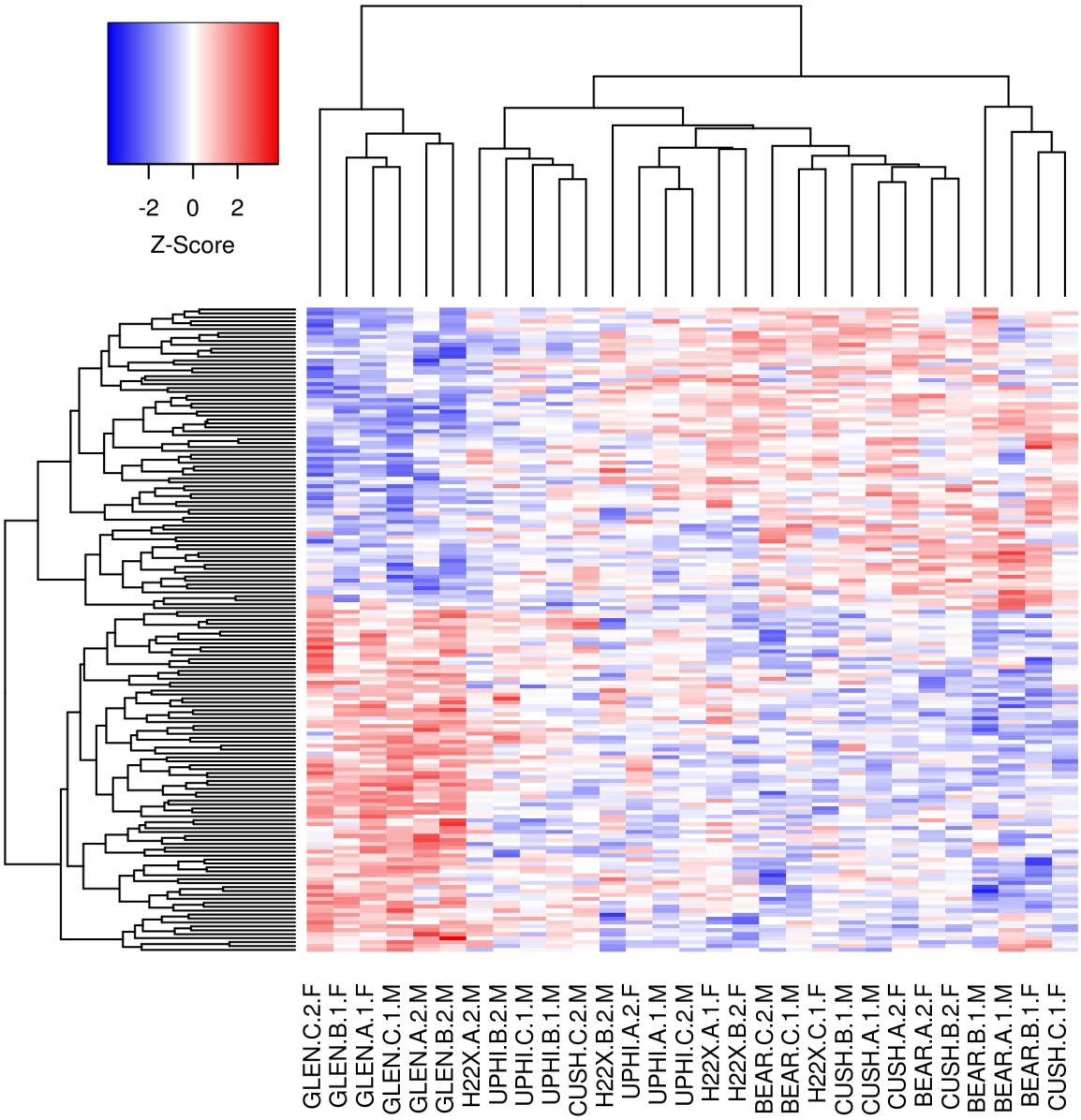
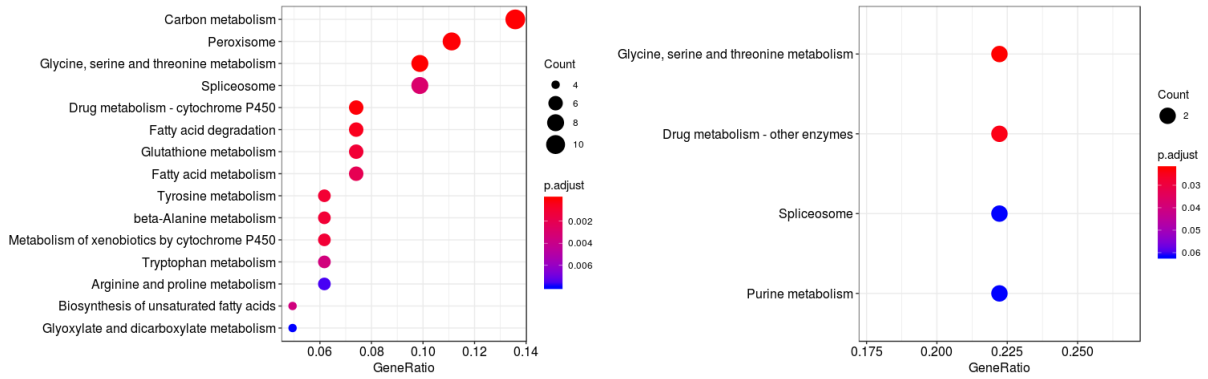


Figure 3.8: Heatmap of samples clustered by differential expression pattern. Sample naming follows the same scheme as Figure 3.7.)



(a) (b)

Figure 3.9: Enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in significantly differentially expressed proteins in a the Outflow group and b the Downstream group. Only the top 15 most enriched KEGG pathways are show in figure a. The size of each node indicates the number of proteins with that term, and the color indicates the Benjamini-Hochberg adjusted p-value for the enriched term.

occurring in the Downstream group (H22X and UPHI sites) and the Outflow group (GLEN site). For comparison, measurement of select PPCP found they were mostly indetectable upstream of the WWTPs, had similar concentration to undiluted MWWE at the GLEN sites, and reduced concentration further downstream (Table B.1).

Table 3.2: Log_2 fold-change and raw p-values of differentially expressed proteins in the Outflow and Downstream groups compared to the Upstream group. Proteins with p-values listed are differentially expressed compared to the Upstream group. The p-value of proteins which are significantly differentially expressed between the Outflow and Downstream group are listed in the 'Between' column.

Symbol	Description	Outflow		Downstream		Between
		logFC	p-value	logFC	p-value	p-value
CYP2AA4	cytochrome P450, family 2, subfamily AA, polypeptide 4	5.00	5.58e−5	0.40	-	1.78e−4
CYP2X10.2	cytochrome P450, family 2, subfamily X, polypeptide 10.2	4.76	8.24e−9	1.79	8.78e−4	2.41e−5
DHDHL	dihydrodiol dehydrogenase (dimeric), like	2.24	3.85e−7	0.69	-	9.53e−5
AOX6	aldehyde oxidase 6	1.96	1.07e−3	0.43	-	-
RIDA (HRSP12)	reactive intermediate imine deaminase A homolog	1.68	3.43e−5	0.80	-	-
GAMT	guanidinoacetate N-methyltransferase	1.46	3.45e−8	0.53	2.38e−3	5.05e−5
UGDH	UDP-glucose 6-dehydrogenase	1.45	2.27e−5	0.33	-	5.78e−4
CYP4T8	cytochrome P450, family 4, subfamily T, polypeptide 8	1.41	3.62e−7	0.79	1.18e−4	-
DIO1	deiodinase, iodothyronine, type I	1.40	1.81e−8	0.52	1.57e−3	3.56e−5
OCIAD2	OCIA domain containing 2	1.40	5.50e−4	0.08	-	1.14e−3
BECN1	beclin 1, autophagy related	1.33	7.37e−4	0.07	-	1.43e−3
APOA2	apolipoprotein A-II	1.30	3.71e−5	0.21	-	3.39e−4
CBSA	cystathionine-beta-synthase a	1.26	2.13e−4	0.70	-	-
PTGR1	prostaglandin reductase 1	1.26	1.38e−7	0.01	-	1.85e−7
CYP2K21	cytochrome P450, family 2, subfamily k, polypeptide 21	1.25	1.02e−6	0.17	-	1.14e−5
RBP7B	retinol binding protein 7b, cellular	1.24	1.44e−6	0.38	-	2.40e−4
MLSL	zmp:0000000758	1.21	1.18e−4	0.23	-	1.23e−3
GSTP1	glutathione S-transferase pi 1	1.16	6.83e−4	0.49	-	-
MAO	monoamine oxidase	1.15	3.01e−4	0.06	-	6.25e−4
UCP1	uncoupling protein 1	1.14	3.21e−3	0.84	-	-
AIFM2	apoptosis-inducing factor, mitochondrion-associated, 2	1.14	1.87e−8	0.10	-	1.20e−7
MAO	monoamine oxidase	1.12	2.32e−5	0.05	-	6.25e−4
GSTA.2	glutathione S-transferase, alpha tandem duplicate 2	1.12	2.50e−5	0.03	-	4.49e−5
CYP2AD2	cytochrome P450, family 2, subfamily AD, polypeptide 2	1.10	2.31e−5	0.46	-	-
QDPRA	quinoid dihydropteridine reductase a	1.09	2.96e−7	0.28	-	3.45e−5
PGD	phosphogluconate dehydrogenase	1.06	5.31e−9	0.03	-	1.24e−8
AGXTA	alanine-glyoxylate aminotransferase a	1.00	1.30e−5	0.22	-	3.33e−4
CROT	carnitine O-octanoyltransferase	0.98	4.05e−5	0.32	-	3.12e−3
zgc:172341	D-aspartate oxidase	0.95	1.94e−4	0.19	-	2.16e−3
BDH1	3-hydroxybutyrate dehydrogenase, type 1	0.93	9.28e−6	0.33	-	1.66e−3
MTHFD1B	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1b	0.93	2.37e−3	0.32	-	-

Symbol	Description	Outflow		Downstream		Between
		logFC	p-value	logFC	p-value	p-value
AGXTB	alanine-glyoxylate aminotransferase b	0.93	7.48e−5	0.22	-	1.56e−3
GPX4A	glutathione peroxidase 4a	0.91	4.73e−5	0.36	-	-
BHMT	betaine-homocysteine methyltransferase	0.90	4.56e−4	0.59	3.85e−3	-
GSTO1	glutathione S-transferase omega 1	0.89	5.83e−4	0.14	-	3.17e−3
SQRDL	sulfide quinone reductase-like (yeast)	0.89	2.72e−4	−0.01	-	2.86e−4
ACOX1	acyl-CoA oxidase 1, palmitoyl	0.88	1.12e−4	0.15	-	9.50e−4
ASMTL	acetylserotonin O-methyltransferase-like	0.86	4.57e−7	0.44	3.79e−4	3.47e−3
SELENBP1	selenium binding protein 1	0.86	1.78e−6	0.28	-	4.12e−4
zgc:56493	zgc:56493	0.86	9.65e−4	0.00	-	1.10e−3
UOX	urate oxidase	0.85	7.55e−5	0.38	-	-
SCARB2B	scavenger receptor class B, member 2b	0.83	1.09e−4	0.05	-	2.52e−4
HTATIP2	HIV-1 Tat interactive protein 2	0.82	1.94e−5	0.05	-	5.37e−5
ABCC2	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	0.81	2.81e−5	0.06	-	9.02e−5
SLC27A2B	solute carrier family 27 (fatty acid transporter), member 2b	0.79	6.67e−4	0.34	-	-
MGST1.1	microsomal glutathione S-transferase 1.1	0.78	6.53e−5	0.49	1.46e−3	-
AASS	aminoadipate-semialdehyde synthase	0.78	4.02e−4	0.18	-	4.63e−3
ACO1	aconitase 1, soluble	0.77	1.87e−3	0.38	-	-
si:ch211-93f2.1	Predicted to have carboxylic ester hydrolase activity	0.74	3.53e−3	−0.23	-	3.08e−4
si:dkey-91i10.3	Orthologous to human CYP27A1	0.71	6.57e−6	0.11	-	7.68e−5
CYP3C4	cytochrome P450, family 3, subfamily C, polypeptide 4	0.69	1.61e−3	0.12	-	-
ALDH16A1	aldehyde dehydrogenase 16 family, member A1	0.66	4.51e−4	0.23	-	-
HGD	homogentisate 1,2-dioxygenase	0.65	1.34e−3	0.24	-	-
APOBA	apolipoprotein Ba	0.62	3.29e−4	0.14	-	3.87e−3
ACOX3	acyl-CoA oxidase 3, pristanoyl	0.59	2.09e−3	0.31	-	-
GOT1	glutamic-oxaloacetic transaminase 1, soluble	0.58	1.63e−3	0.31	-	-
ACO1	aconitase 1, soluble	0.58	1.11e−3	0.26	-	-
SCP2A	sterol carrier protein 2a	0.58	1.18e−3	0.22	-	-
PRDX6	peroxiredoxin 6	0.58	2.69e−3	0.18	-	-
AMPD2A	adenosine monophosphate deaminase 2a	0.56	2.97e−3	0.28	-	-
EHHADH	enoyl-CoA, hydratase/3-hydroxyacyl CoA dehydrogenase	0.55	8.46e−4	0.37	-	-
SDSL	serine dehydratase-like	0.55	3.57e−4	0.16	-	-
PPID	peptidylprolyl isomerase D	0.54	3.52e−3	0.25	-	-
COMTA	catechol-O-methyltransferase a	0.51	1.65e−3	0.15	-	-
RTN4A	reticulon 4a	0.49	1.47e−3	0.21	-	-
NME2B.1	NME/NM23 nucleoside diphosphate kinase 2b, tandem duplicate 1	0.46	1.34e−3	0.42	5.41e−4	-

Symbol	Description	Outflow		Downstream		Between
		logFC	p-value	logFC	p-value	p-value
CANT1B	calcium activated nucleotidase 1b	0.46	3.51e−4	−0.12	-	2.18e−5
RPL4	ribosomal protein L4	0.44	2.73e−3	0.15	-	-
SEC14L7	SEC14-like lipid binding 7	0.44	4.68e−3	0.12	-	-
LARP1	La ribonucleoprotein domain family, member 1	0.42	3.56e−3	0.09	-	-
HSD17B4	hydroxysteroid (17-beta) dehydrogenase 4	0.42	1.74e−3	0.03	-	3.59e−3
MIPEP	mitochondrial intermediate peptidase	0.41	9.36e−4	0.04	-	2.65e−3
IDE	insulin-degrading enzyme	0.39	8.25e−4	0.29	2.61e−3	-
CSNK2A1	casein kinase 2, alpha 1 polypeptide	0.36	2.31e−3	−0.01	-	2.17e−3
EFTUD2	elongation factor Tu GTP binding domain containing 2	−0.29	3.24e−3	0.11	-	1.43e−4
MATR3L1.1	matrin 3-like 1.1	−0.32	8.17e−4	−0.05	-	3.72e−3
UPF2	UPF2 regulator of nonsense mediated mRNA decay	−0.33	2.03e−3	−0.01	-	2.80e−3
RBM12B	RNA binding motif protein 12B	−0.35	5.98e−4	−0.23	-	-
PKP3B	plakophilin 3b	−0.36	1.56e−3	−0.26	4.70e−3	-
ELAVL1	ELAV like RNA binding protein 1	−0.37	4.41e−4	−0.25	3.57e−3	-
RBM12	RNA binding motif protein 12	−0.37	4.19e−3	−0.26	-	-
GLG1A	golgi glycoprotein 1a	−0.39	6.88e−5	−0.22	3.16e−3	-
PRPF8	pre-mRNA processing factor 8	−0.40	1.30e−3	−0.03	-	2.84e−3
DMGDH	dimethylglycine dehydrogenase	−0.42	2.38e−3	−0.23	-	-
CDC5L	CDC5 cell division cycle 5-like (S. pombe)	−0.42	4.72e−3	−0.46	4.08e−4	-
SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	−0.43	1.21e−3	−0.04	-	3.38e−3
SRP54	signal recognition particle 54	−0.44	2.22e−3	0.15	-	1.37e−4
SPNA2	spectrin alpha 2	−0.44	2.37e−4	−0.21	-	-
ILF3B	interleukin enhancer binding factor 3b	−0.45	1.11e−3	−0.25	-	-
SAE1	SUMO1 activating enzyme subunit 1	−0.46	1.18e−4	−0.04	-	3.59e−4
PCBP2	poly(rC) binding protein 2	−0.46	1.17e−5	−0.05	-	6.53e−5
PLECB	plectin b	−0.46	2.50e−3	−0.33	-	-
PUM3	pumilio RNA-binding family member 3	−0.47	3.65e−3	−0.16	-	-
PRPF6	PRP6 pre-mRNA processing factor 6 homolog (S. cerevisiae)	−0.48	1.14e−3	−0.03	-	2.28e−3
FLNB	filamin B	−0.48	7.77e−5	−0.18	-	-
SLC25A1B	slc25a1 solute carrier family 25 (mitochondrial carrier; citrate transporter), member 1b	−0.49	2.37e−3	−0.13	-	-
MICAL2B	microtubule associated monooxygenase, calponin and LIM domain containing 2b	−0.49	2.84e−3	−0.25	-	-
PLSCR3B	phospholipid scramblase 3b	−0.49	8.79e−4	−0.32	-	-
CSE1L	CSE1 chromosome segregation 1-like (yeast)	−0.50	5.68e−4	0.17	-	1.88e−5
SF3B1	splicing factor 3b, subunit 1	−0.50	2.44e−4	0.00	-	2.73e−4

Symbol	Description	Outflow		Downstream		Between
		logFC	p-value	logFC	p-value	p-value
H2AFY2	H2A histone family, member Y2	−0.51	9.90e−5	−0.09	-	9.16e−4
PAPOLA	poly(A) polymerase alpha	−0.52	2.04e−3	−0.12	-	-
ANO9A	anoctamin 9a	−0.53	4.43e−4	−0.31	-	-
CS	citrate synthase	−0.54	7.84e−4	−0.14	-	-
CTNNBL1	catenin, beta like 1	−0.54	2.83e−4	−0.17	-	-
MGEA5	meningioma expressed antigen 5 (hyaluronidase)	−0.56	2.64e−5	−0.07	-	1.57e−4
ADAR	adenosine deaminase, RNA-specific	−0.56	4.23e−5	−0.11	-	6.30e−4
zgc:64106	Similar to retinol dehydrogenase 11-like (Ictalurus punctatus)	−0.57	2.56e−4	0.00	-	3.24e−4
API5	apoptosis inhibitor 5	−0.57	9.89e−5	−0.04	-	2.91e−4
KRT8	keratin 8	−0.57	4.62e−3	−0.54	1.52e−3	-
ABCB11B	ATP-binding cassette, sub-family B (MDR/TAP), member 11b	−0.59	1.08e−3	0.09	-	2.92e−4
RBM19	RNA binding motif protein 19	−0.59	3.02e−4	−0.29	-	-
PRPF3	PRP3 pre-mRNA processing factor 3 homolog (yeast)	−0.59	8.73e−4	−0.52	4.91e−4	-
KHDRBS1A	KH domain containing, RNA binding, signal transduction associated 1a	−0.61	7.05e−5	−0.13	-	1.20e−3
RBM45	RNA binding motif protein 45	−0.64	6.35e−7	−0.10	-	9.77e−6
ACADVL	acyl-CoA dehydrogenase, very long chain	−0.67	2.72e−3	−0.28	-	-
PTBP2A	polypyrimidine tract binding protein 2a	−0.68	1.68e−6	−0.31	2.38e−3	2.83e−3
ALDH9A1B	aldehyde dehydrogenase 9 family, member A1b	−0.68	8.64e−4	−0.37	-	-
VAT1	vesicle amine transport 1	−0.69	3.87e−3	−0.69	8.66e−4	-
PRMT1	protein arginine methyltransferase 1	−0.71	1.63e−3	0.22	-	1.14e−4
ATP5O	ATP synthase, H+ transporting, mitochondrial F1 complex, O subunit	−0.75	3.86e−3	0.01	-	3.80e−3
ACADVL	acyl-CoA dehydrogenase, very long chain	−0.76	2.82e−3	−0.37	-	-
GAPVD1	GTPase activating protein and VPS9 domains 1	−0.77	2.32e−3	−0.18	-	-
FLNA	filamin A, alpha (actin binding protein 280)	−0.79	2.98e−3	−0.29	-	-
BZW1B	basic leucine zipper and W2 domains 1b	−0.81	3.67e−4	−0.17	-	3.47e−3
IQGAP1	IQ motif containing GTPase activating protein 1	−0.81	1.97e−3	−0.17	-	-
FDPS	farnesyl diphosphate synthase	−0.82	1.24e−3	−0.22	-	-
SLC29A1B	solute carrier family 29 (equilibrative nucleoside transporter), member 1b	−0.84	1.78e−3	−0.04	-	3.08e−3
CFL1L	cofilin 1 (non-muscle), like	−0.84	3.65e−5	−0.38	-	-
SCINLA	scinderin like a	−0.95	1.82e−4	−0.11	-	7.70e−4
ANXA4	annexin A4	−0.97	9.54e−5	−0.06	-	2.45e−4
CTSH	cathepsin H	−0.99	1.26e−3	−0.39	-	-
POLR2GL	polymerase (RNA) II (DNA directed) polypeptide G-like	−1.00	1.80e−4	−0.10	-	6.41e−4

Symbol	Description	Outflow		Downstream		Between
		logFC	p-value	logFC	p-value	p-value
CYP51	cytochrome P450, family 51	−1.01	2.39e−3	−0.02	-	3.12e−3
ANXA11B	annexin A11b	−1.06	1.33e−6	−0.34	-	3.07e−4
HKDC1	hexokinase domain containing 1	−1.15	1.44e−5	−0.19	-	1.69e−4
MAPRE3B	microtubule-associated protein, RP/EB family, member 3b	−1.34	1.21e−10	−0.39	1.33e−3	1.28e−7
LGALS2B	lectin, galactoside-binding, soluble, 2b	−1.42	6.03e−5	−0.57	-	-
CTSD	cathepsin D	−1.46	3.78e−5	−0.34	-	9.32e−4
MPC2	mitochondrial pyruvate carrier 2	−1.52	3.24e−3	−0.06	-	-
PAICS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazole succinocarboxamide synthetase	0.34	-	0.33	4.62e−3	-
SNX1A	sorting nexin 1a	0.06	-	−0.29	4.56e−3	4.71e−3
MRPS22	mitochondrial ribosomal protein S22	−0.09	-	−0.32	1.72e−3	-
si:ch211-175m2.5	Predicted to have peroxiredoxin activity	−0.31	-	−0.39	2.11e−3	-
SFPQ	splicing factor proline/glutamine-rich	−0.49	-	−0.54	4.65e−4	-
CYP2AA9	cytochrome P450, family 2, subfamily AA, polypeptide 9	1.51	-	−0.19	-	2.51e−3
KLHL15	kelch-like family member 15	1.26	-	−0.11	-	4.36e−3
SORD	sorbitol dehydrogenase	0.63	-	−0.02	-	4.23e−3
ALDH8A1	aldehyde dehydrogenase 8 family, member A1	0.61	-	−0.07	-	3.94e−3
GLRX3	glutaredoxin 3	0.60	-	−0.11	-	2.24e−3
PHYH	phytanoyl-CoA 2-hydroxylase	0.49	-	−0.21	-	1.87e−4
VWA8	von Willebrand factor A domain containing 8	0.44	-	−0.18	-	2.41e−3
CYP3A65	cytochrome P450, family 3, subfamily A, polypeptide 65	0.43	-	−0.21	-	1.51e−3
OXCT1B	3-oxoacid CoA transferase 1b	0.41	-	−0.19	-	2.04e−3
si:ch211-93f2.1	Predicted to have carboxylic ester hydrolase activity	0.41	-	−0.46	-	3.08e−4
PFN2L	profilin 2 like	0.36	-	−0.13	-	9.93e−4
G6PD	glucose-6-phosphate dehydrogenase	0.30	-	−0.04	-	3.87e−3
PSME4B	proteasome activator subunit 4b	−0.20	-	0.12	-	2.97e−3
CNOT1	CCR4-NOT transcription complex, subunit 1	−0.26	-	0.22	-	2.66e−4
VPS4B	vacuolar protein sorting 4 homolog B	−0.28	-	0.04	-	4.66e−3
VWA5A	von Willebrand factor A domain containing 5A	−0.30	-	0.19	-	3.50e−3
PCYOX1	prenylcysteine oxidase 1	−0.31	-	0.26	-	2.82e−3
SEC24D	SEC24 homolog D, COPII coat complex component	−0.34	-	0.44	-	4.17e−3
PCNA	proliferating cell nuclear antigen	−0.51	-	0.17	-	3.62e−3

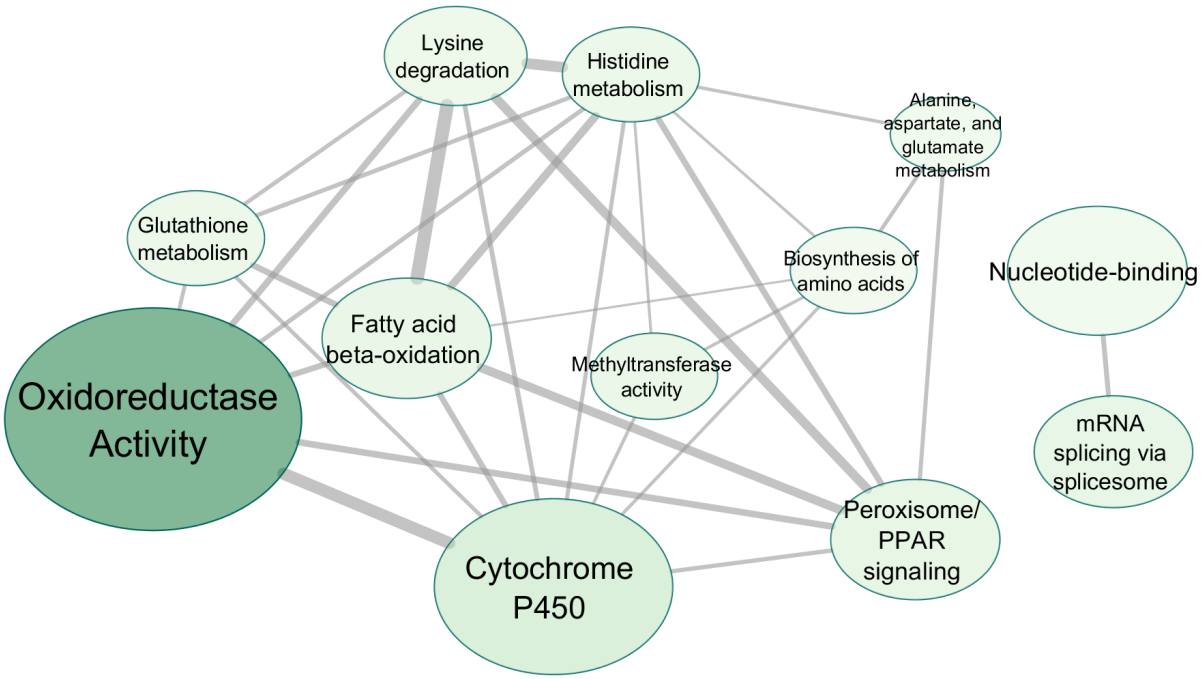


Figure 3.10: Network of DAVID Functional Annotation Cluster from significantly differentially expressed proteins in the Outflow group. The size of each node indicates the number of proteins with that term, and edges indicate overlapping groups. Edge thickness is proportional to the Jaccard similarity of the groups.

3.3.6 Differentially expressed proteins in the Outflow Group

Further analysis of the Outflow group using DAVID for a combined analysis of GO, KEGG, and Reactome terms showed that the DE proteins had range of molecular functions (Fig. 3.10, Table 3.2) consistent with a general xenobiotic stress response. Of particular interest, Figure 3.10 shows that a large group of proteins with overlapping oxidoreductase, and Cytochrome P450 functions was upregulated. This is consistent with exposure to a wide range of different contaminants, which induce metabolic changes and increased oxidative stress in the organism in order to biotransform xenobiotics for inactivation and removal[125]. In addition, proteins with PPAR signaling and carbon metabolism molecular functions were also significantly enriched compared to the upstream group. PPAR signaling regulates many metabolic functions, particularly lipid metabolism, which plays a key role in maintaining energy homeostasis in fish[10]. Common pharmaceuticals, including statins and fibrates like gemfibrozil, which have previously been found in the Bow River[11] as well as a part of this study (Table B.1), are known to impact PPAR signaling and lipid metabolism in environmentally relevant exposures.[10,21] However, the increased nutrition in the WWTP outflow,[12] along with the increased temperature, (Table B.2) may play a significant role in the observed changes in the Outflow group.

Phase I Xenobiotic Metabolism

A variety of proteins are commonly found in the response to xenobiotic challenge, usually involved in Phase I or II biotransformation, or combating the oxidative stress induced by degrading these compounds[27]. A key type class of protein for the transformation of xenobiotics is the cytochrome P450s, heme-binding enzymes that catalyze mono-oxygenase reactions on both exogenous and endogenous chemicals.[26] While the commonly used biomarker CYP1A proteins were not differentially expressed in any of the comparisons, a range of CYP2, 3, and 4 family proteins were significantly differentially expressed. Cytochromes are diverse in function, with families one through four are generally considered as having xenobiotic metabolizing function[29]. Previous studies in medaka have found that family two cytochrome P450s are widely regulated in response to exposure to benzo[a]pyrene[126] and water-soluble crude oil components.[127]

CYP2AA4 was the most strongly upregulated protein Outflow group, with a 31.9-fold increase in expression (Table 3.2). The CYP2AA family is most similar to the CYP2X family, and is fish-specific. The function of the CYP2AAs if unknown, previous test have suggested that the CYP2AA family may respond to phenobarbitol-type inducers.[128] CYP2X10.2 was the second most strongly differentially expressed protein in the Outflow groups, showing a log2FC of 4.76, and the most strongly differentially expressed in the Downstream group, with a log2FC of 1.79 (Table 3.2). The CYP2X family is found in fish but not in mammals[29], and limited information is available regarding their function. Studies of the CYP2X1 protein in channel catfish found limited similarity in function and substrate specificity to other CYP2 family proteins[129]. However, studies of goldfish have found upregulation of CYP2X10 in fish exposed to wastewater effluent. Interestingly, the study compared fish in two lakes, and upregulation occurred in the lake that was both exposed to municipal wastewater effluent and experienced increased temperatures of 5-10 °C over the ambient temperature[130], similar to what was observed at the GLEN site (Table B.2). This suggests a potential link between the expression of CYP2X10 and temperature.

Five other CYP proteins, CYP2AD2, CYP2K21, CYP3C4, CYP4T8, and CYP51 were also differentially expressed in the Outflow Group compared to Upstream. CYP51, a key enzyme in the synthesis of cholesterol,[26] was downregulated, while the remaining were upregulated.

CYP2AD2 is part of a cluster of family 2 cytochromes occurring next to the conserved *HOOK1* gene that have undergone expansion in *Danio rerio* into the CYP2N, CYP2P, CYP2AD, and CYP2V subfamilies, compared to the single CYP2J2 found in humans[29]. This expansion has also been found multiple species of killifish[28,131], and can also be partially observed in the fathead minnow genome. While the specific function of all these cytochromes are not known[29], CYP2J2 in humans is an epoxygenase that converts arachidonic acid into epoxyeicosatrienoic acid (EET) and other signaling molecules involved in regulation of apoptosis, inflammation, and metabolism through interaction with PPARs[132]; the CYP2N subfamily

in *F. heroclitus* has been shown to have a similar function[131]. The CYP2AD2 and orthologs have also been found to be xenobiotic responsive, being increased expression in response to bisphenol A (BPA), and 4-octylphenol (OP)[31], and has been shown to degrade benzamphetamine in killifish[131].

Fish CYP4s are thought to be fatty acid hydroxylases like their mammalian orthologs, and to play a smaller role in xenobiotic metabolism.[29] However, exposure to clofibrate and other PPAR inducer has been known to alter expression of CYP4T8 orthologs and impact fatty acid hydroxylation in some species, suggesting a link between CYP4s, fatty acid metabolism, and xenobiotics.[26] The functions of the remaining DE CYPs, CYP2K21 and CYP3C4, are unknown.[26,29]

Besides the cytochrome P450s, several other proteins involved in xenobiotic metabolism were differentially expressed. Aldehyde oxidase 6, AOX6, is upregulated nearly 4-fold, and has broad substrate specificity for the oxidation of various drugs in concert with different P450 enzymes as well as monoamine-oxidase (MAO),[133] which was upregulated more than 2 fold. Another potentially interesting set of non-CYP450 proteins for xenobiotics degradation is the combined upregulation of DHDHL and COMTA. DHDHL, or dihydrodiol dehydrogenase (dimeric) like, an ortholog of the human DHDH protein[116], is upregulated 4.7-fold. DHDH oxidizes polycylic aromatic hydrocarbons to catechols.[134] Catechol-O-methyltransferase a (COMTA) is also upregulated in the Outflow group, and degrades neurotransmitters and other catechol-based chemicals, including a wide variety of drugs, through methylation.[135] This may indicate the use of two sequential modifications of a xenobiotic to maximize chances of inactivation and removal.

Notably, the four most strongly differentially expressed proteins between the Upstream and Outflow groups are CYP2AA4, CYP2X10.2, DHDHL, and AOX6, suggesting a significant demand for increased xenobiotic metabolism. The large change in CYP2 family protein expression, particularly CYP2X10.2, suggests their potential as a biomarker, however, the quantity and variety of the CYP2s in fish, and the variability of their response to xenobiotics[26,29,31] suggests that selecting any particular CYP2 for testing would be difficult.

Phase II Xenobiotic Metabolism, Oxidative Stress, and Glutathione Metabolism

The glutathione S-transferases (GSTs) MGST1.1, GSTP1, GSTA.2, and GSTO1 were also upregulated between 1.7 and 2.2-fold. GSTs are important for xenobiotic removal because they catalyze the conjugation of reduced glutathione (GSH) to diverse compounds to reduce activity and increase solubility of the compound for excretion. GSTP1, GSTA.2, and GSTO1 are cytosolic GSTs that also function as GST peroxidases, removing reactive oxygen species (ROS) to prevent oxidative damage.[136]

A side effect of biotransformation of xenobiotics is increased ROS,[133] and several oxidative stress proteins were found to be significantly upregulated in the Outflow group. In particular GPX4A, glutathione

peroxidase 4a, and PRDX6, peroxiredoxin 6, both have known antioxidant behaviour. Glutathione peroxidases play an important role in removing H_2O_2 by catalyzing the formation of dimeric glutathione (GSSG) from the reduced form (GSH), simultaneously converting hydrogen peroxide into water, and glutathione peroxidase 4 (GPX4) is a well characterized GPx known to protect against oxidative attack on lipids.[27,137] In their review of redox regulation in peroxisomes, Walker et al. suggest that, due to the inactivation of catalase when exposed to elevated levels of hydrogen peroxide, glutathione peroxidases may be the primary mechanism for H_2O_2 during acute oxidative stress, such as exposure to hypolipidemic drugs such as fibrates or statins.[32] Peroxiredoxins also play an important role in oxidative stress[32], but PRDX6 proteins are unique in having both peroxidase and phospholipase activity, allowing them to protect phospholipids in lipid membranes against the oxidation by preventing the formation of damaging hydroperoxides.[138]. In the Outflow group, GPX4A showed a 2-fold increase in expression, and PRDX6 a 1.7-fold increase (Table 3.2), suggesting increased need for antioxidants in the fish at the GLEN site.

In addition to the enzymes to catalyze ROS and xenobiotic transformations, maintaining sufficient levels of GSH is necessary for maintaining ROS homeostasis. CBSA, cystathionine-beta-synthase a, performs the first step in the conversion of homocysteine to cysteine for glutathione and protein production and was upregulated 2.4-fold.[139] GSSG can also be reduced by to GSH by NADPH, which is catalyzed by glutathione reductase (GSR). NADPH is itself regenerated by phosphogluconate dehydrogenase (PGD) and glucose-6-phosphate dehydrogenase (G6PD).[140] PGD was significantly upregulated, while the GSR and G6PD were upregulated, but not significant (Table 3.2 and Figure B.13). This demonstrates an increase in both the proteins to carry out xenobiotic and ROS reactions, and proteins producing the necessary endogenous compounds for those reactions. Field studies of MWWE have found changes in oxidative stress proteins after exposure,[20,130] and changes in glutathione and oxidative stress parameters has been widely studied as a biomarker for a wide variety of contaminants, including metals,[27] polycyclic aromatic hydrocarbons and other organic pollutants[125], and MWWE exposure.[141] However, response of oxidative stress proteins and glutathione metabolites is usually highly variable with species, contaminant, dose, and degree of exposures.[27,142]

Changes in Lipid Metabolism and the Tricarboxylic Acid Cycle

A number of genes related to lipid metabolism were DE in the Outflow group compared to the Upstream group. Multiple peroxisomal β-oxidation proteins were significantly upregulated, including the acyl-CoA oxidases, acyl-CoA oxidase 1 (ACOX1) and ACOX3, by 1.8-fold and 1.5-fold, respectively, as well as enoyl-CoA, hydratase/3-hydroxyacyl CoA dehydrogenase (EHHADH) by 1.5-fold (Table 3.2. ACOX1, also known as palmitoyl acyl-CoA oxidase 1, is responsible for beta-oxidation of long-chain fatty acids in the peroxisome, while ACOX3, along with EHHADH, are part of the degradation pathway for branched-

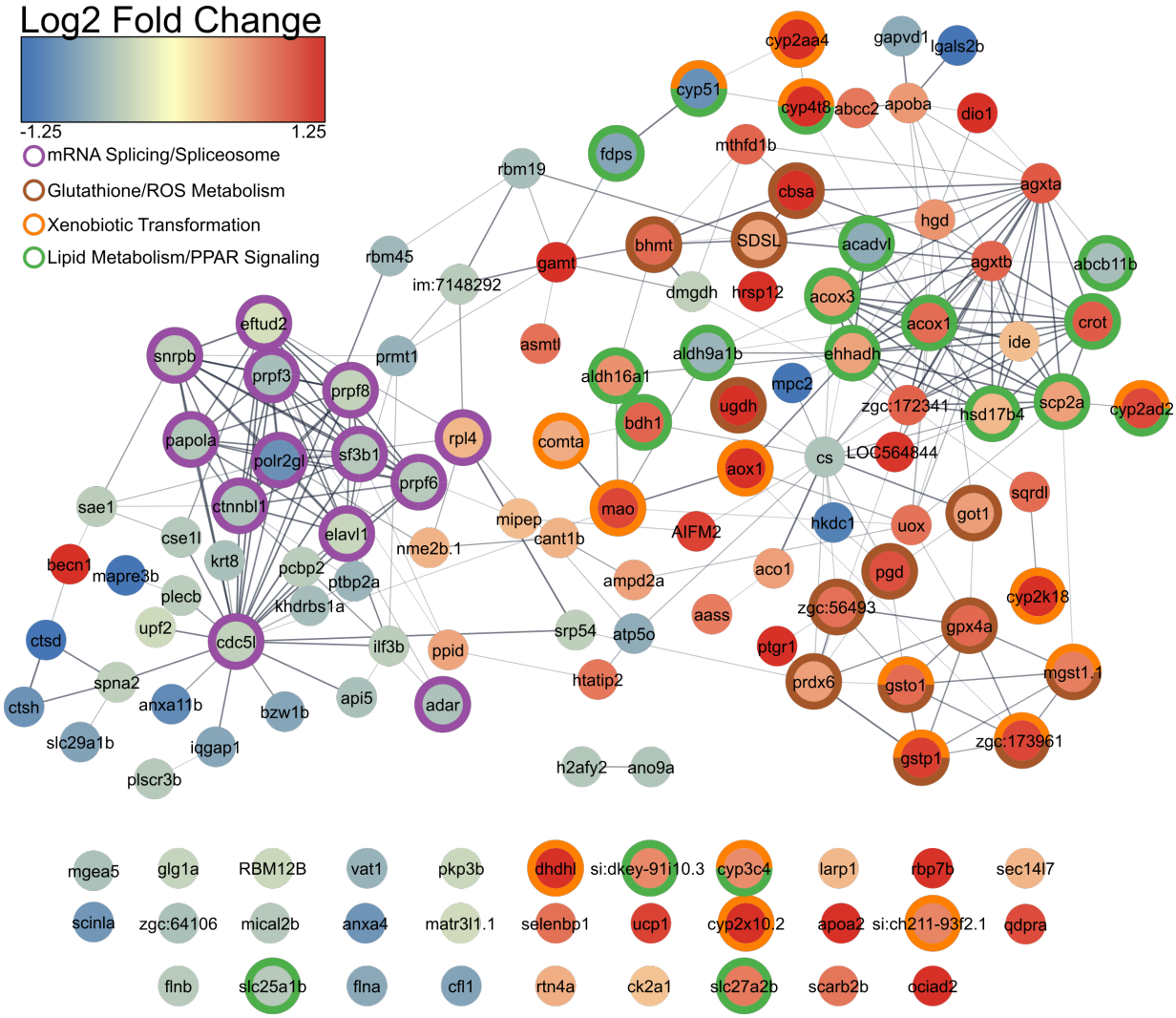


Figure 3.11: Search Tool for Retrieval of Interacting Genes/Proteins (STRING)[143] interaction network of significantly differentially expressed protein in the Outflow group. Nodes are coloured by Log_2 fold-change, and the outer rings indicate enriched pathways for select genes. Edges between nodes proteins with known or suspected interactions.

chain fatty acids.[10] SLC27A2b (solute carrier family 27 member 2b) was upregulated 1.7-fold, and converts fatty acids imported into the peroxisome into acyl-CoA for β-oxidation.[144] ACOX activity in the peroxisome is a known marker of β-oxidation in the peroxisome, and is a rate limiting step, though ACOX activity and gene transcription are known to be disconnected.[9] Both of these processes produce hydrogen peroxide as a byproduct, increasing the oxidative burden on the organism,[32] which may be linked with the increased expression of GPX4a and PRDX6, and the various GSTs. Acyl-CoA oxidase activity has has been previously suggested as a possible indicator of fatty acid (FA) metabolism changes after exposure for a variety of compounds in aquatic environments, including fibrates and perfluoroalkyl acids.[10] Other proteins involved in peroxisomal β-oxidation are also upregulated. However, the changes observed in lipid metabolism and oxidative stress are consistent with differential expression of genes with increased food availability,[145], which is expected with the increase nutrient downstream of MWWE outflow.

Other proteins differentially regulated in the Outflow group are also linked to lipid metabolism. Apoliprotein A-II (APOA2), which is involved in transport of lipids, as well as retinol binding protein 7b (RBP7b), which is both activated by and a co-activator of PPARγ,[146] were upregulated. GOT1 (increased 1.5-fold) is the cytoplasmic variant of glutamic-oxaloacetic transaminase, which is necessary for the regeneration of the peroxisomal NADPH needed for peroxisomal β-oxidation. Also upregulated was the protein carnitine O-octanyltransferase (CROT), part of the export pathway of shortened FAs from the peroxisome to the mitochondria via carnitine shuttling[144]. Interestingly, while CROT is upregulated nearly twofold in the Outflow group, a carnitine biosynthesis protein, aldehyde dehydrogenase 9a1b (ALDH9A1B) was downregulated. Given the evidence of increased peroxisomal fatty acid oxidation, increased carnitine biosynthesis would be expected, though aldh9a1b downregulation has been observed with long term exposure to mixtures of persistent organic pollutants.[147]

An important regulator of lipid metabolism is the PPAR signaling pathway, particularly through PPARγsignaling. PPARγis known to regulate expression of ACOX1, along with a variety of other other proteins, to increase rate of fatty acid oxidation (FAO) in the peroxisome for energy utilization and FA elongation. PPAR transcription factors are activated by a wide variety of ligands, such cellular lipids and phospholipids, as well statins, fibrates, and a wide range of organic pollutants.[10] For example, the EET created from arachidonic acid by CYP2J2 homologs,[131] is a known agonist of PPARs[148], as are polyunsaturated fatty acids (PUFAs) created during FA elongation.[149] Expression of arachidonic acid and many of the signaling molecules derived from it are altered by ROS levels,[150], and many be degraded by lipid peroxidation under oxidative stress.[151] With increased lipid metabolism there may be increased need for production of these signaling molecules and reversal of oxidative damage, such as performed by PRDX6. However, as with the increase ACOX expression, the changes in APOA2, RBP7b, and CROT are consistent with changes in PPAR and FA metabolism expected with increased nutrients.[145]

Metabolic protein expression changes beyond fatty acid β-oxidation were also observed in the Outflow group, including proteins involved in the tricarboxylic acid cycle (TCA) and related pathways. Citrate synthase (CS), which regulates entry of acetyl-CoA into the TCA and influences lipid and cholesterol synthesis,[152] was downregulated 1.5-fold, and MPC2, a protein necessary for the import of pyruvate into the mitochondria for oxidation[153] was down-regulated 2.9-fold. SLC25A1B, a mitochondrial citrate transporter, and acyl-CoA dehydrogenase, very long chain (ACADVL), a mitochondrial beta-oxidation protein,[10] were down-regulated well, as was ATP5O, a mitochondrial ATP synthase(Table 3.2). Combined, the downregulation of these protein suggest reduce β-oxidation and TCA activity in the mitochondria of fish in the Outflow group.

Cholesterol synthesis and transport was also altered the Outflow group. As previously mentioned in Section 3.3.6, CYP51 was downregulated 2-fold, additionally, farnesyl diphosphate synthase (FDPS), was downregulated 1.7-fold, along with ABCD11B, down 1.5-fold. CYP51 and FDPS are necessary for the synthesis of cholesterol.[154,155] ABCB11B is a bile salt exporter pump, performing the rate-limiting step in excreting cholesterol and cholesterol metabolites along with various toxic lipophilic substances, and allowing for the digestion of various fats and vitamins in the intestine.[154] Cholesterol and bile acid synthesis processes rely on products from both the peroxisome and mitochondria, and the can be altered by peroxisomal and mitochondrial dysfunction,[24,144,156] and FDPS is susceptible to inhibition by statins.[155]

Given the increased fish size and nutrient availability at GLEN compared to the upstream sites, as well as the high levels of glucose and lipids in fish there,[12] we would expect protein expression and lipid content profiles inverse to those found in starved fish. This generally includes downregulation of proteins involved in β-oxidation, oxidative stress response, and the TCA, as well as reduced cholesterol and lipid content.[145] In the Outflow group, oxidative stress proteins, and peroxisomal β-oxidation proteins are observed, as expected, and increases in lipid content were observed at the GLEN site compared to others[12]. However, downregulation is seen in several proteins involved in cholesterol synthesis and bile export, which is necessary for dietary fat absorption.[154] Changes in lipid metabolism have been commonly observed after MWWE exposure. Increased peroxisomal proliferation, fatty acid metabolism, and changes in PPAR or associated genes is particularly associated with fibrate exposure.[10,21] Metabolomic and proteomic analysis of goldfish in WWTP outflow in Cootes Paradise Marsh, Ontario, Canada, found increased plasma carnitine metabolites and decreased bile acid synthesis in MWWE exposed fish, along with changes in pathways associated with lipid accumulation that varied with degree and length of exposure.[16]

Additionally, the downregulation of CS and MPC2, which control entry of acetyl-CoA and pyruvate into the TCA, and decreases in ACADVL suggest a downregulation of mitochondrial β-oxidation and the TCA. The TCA is both the main source of ATP for eukaryotes and produces a large number metabolites for many biosynthetic and metabolic processes, including continued peroxisomal β-oxidation,[24,144] so we

would expect upregulation of TCA proteins compared to less well fed fish.[145] Thus, the changes in the Outflow group are difficult to explain solely as the result of increased nutrient availability.

Peroxisome and mitochondria function and regulation are closely related, and dysfunction in one organelle is known to affect the other. Increased peroxisomal β-oxidation due to statins or other PPAR inducers could lead to alterations to ROS, lipid, or other metabolite concentrations that impact mitochondrial function or lead to compensatory regulation of mitochondrial function.[24] Taken together, the differential expression of lipid metabolic proteins at the GLEN site suggest that there are complex interacting factors acting on lipid and other metabolic functions. While some changes are consistent with increased nutrient availability, it is unlikely to be the sole factor.

Splicing and Cell Cycle Regulation

Multiple key proteins in the spliceosome complex were downregulated in the Outflow group (Figure 3.12). Members of the U2, U4, and U5 small nuclear ribonucleoproteins (snRNPs) were all downregulated.The snRNPs are core components of the spliceosome, and as such are regulators of gene translation.[157]. Core member of the Prp19 complex, CDC5L and CTNNBL1, were also downregulated (Table 3.2) The Prp19 complex is a core spliceosome member, and evidence from studies in yeast suggest that it also has roles in genome maintenance, protein modification and degradation, and gene transcription.[158]

A number of proteins involved in regulation of apoptosis were differentially expressed in the Outflow group compared to the upstream group. Apoptosis inhibitor 5 (API5), a highly conserved protein which inhibits apoptosis in response to DNA damage signaling,[159] was downregulated approximately 1.5-fold, while the pro-apoptotic protein AIFM2 (apoptosis inducing factor, mitochondrion-associated 2) was increased two-fold (Table 3.2). Two anti-apoptotic annexin proteins, ANXA4 and ANXA11b, with roles in cell structure and cytokinesis[160,161] were downregulated at the GLEN site. Additionally, histone protein ANO9A, a homolog of Histone H2B[116], was downregulated at the GLEN site. Histone H2B has been found to be bound to double stranded breaks to help induce DNA repair[162], suggesting a reduced capability for DNA repair. Upregulation of apoptosis is commonly found after exposure to contaminants and is often considered as a biomarker of exposure.[27]

The study of MWWE exposure in Cootes Paradise Marsh referenced in Section 3.3.6 also found inhibition of pathways related to organism growth and entry into S-phase in exposed fish. Changes in cell death related pathways were also observed, with the mostly heavily-exposed caged fish showing increase in apoptotic pathways in the liver, while the wild fish showed upregulation of pathways involved in inhibition of apoptosis. This suggests adaptation to MWWE effects over time, as the wild fish were exposed to MWWE for much longer than the caged fish.[16] Exposure to WWTP effluent was also shown

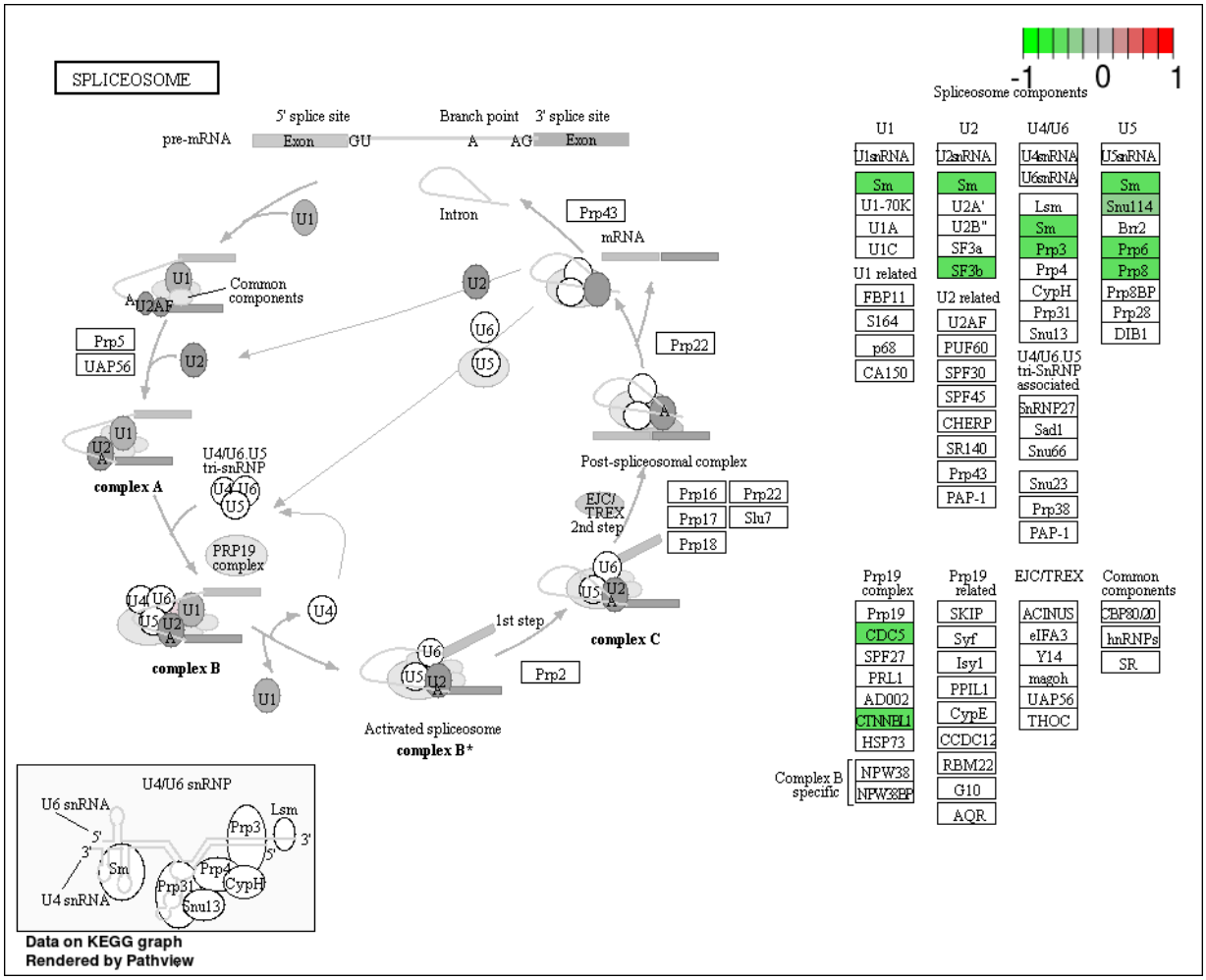


Figure 3.12: Downregulated splicesome proteins in the Outflow group.

to increase DNA damage in a dose-dependent manner, regardless of whether it underwent primary or secondary treatment.[163]

Impacts of Temperature

A key difference between the GLEN site and other caging sites is the 6 °C difference in water temperature due to the waste water treatment plant effluent (Table B.2). Oxidative stress is a known effect of temperature shock in teleosts, but is generally found in a response to a cold temperature shock rather than elevated temperatures.[164] Wen *et al.* [165] found increased expression of UCP1 and oxidative stress protein such as glutathione S-tranferase (GST) omega-1 (GST), and decreased expression of FLNB, a signaling protein, with a 14 °C decrease in temperature, while similar changes are observed in the Outflow group with a 6 °C increase. FHM are robust organisms, capable of surviving a wide range of environmental conditions and temperatures, and the water temperatures at all sites are within ranges normal for the FHM.[166] Thus, any temperature specific effects on protein expression are likely minimal.

One notable temperature related protein is UCP1, upregulated 2.2-fold in the Outflow group, and 1.8-fold in the Downstream group, though the downstream p-value is over the significance threshold after FDR adjustment (Table 3.2). UCP1 is mainly expressed in liver, but the response to thermal shock varies, with expression increasing, decreasing, and remaining unchanged with temperature change depending on tissue and species,[165,167] UCPs are 'uncoupling' proteins, allowing proton leakage through the mitochondrial inner membrane, producing heat, reducing ROS formation, and disconnecting ATP production from oxidative phosphorylation.[167] Rather than responding only to temperature changes, UCPs may also be connected to ROS response through PPARs, reducing mitochondrial ROS formation and alleviating oxidative stress.[164] Thus, the upregulation of UCP1 may be a connected to increased ROS, rather than temperature differences. Increased expression of UCP may be connected to the downregulation of mitochondrial proteins previously discussed in Section 3.3.6.

3.4 Reduced Differential Expression in the Downstream Group

Only 23 proteins are differentially expressed between the Downstream Group and the Upstream group (Table 3.2). Most of the differentially expressed proteins are also differentially expressed in the Outflow group, and the direction of change of proteins differentially expressed in both groups is consistent, and only varies by magnitude, with the Downstream groups showing reduced effect size. Only four enriched KEGG pathways were found (Figure 3.9), and three of them are also enriched in the Outflow group compared to upstream.

3.5 Gene Set Enrichment Analysis

GSEA was performed on the 895 proteins used for differential expression analysis using combined gene sets from GO Biological Process, Molecular Function, and Reactome Pathways. 105 gene sets were enriched between the Outflow and Upstream groups, and 112 were enriched between the Downstream and Upstream groups with a FDR of 20%.

GSEA identified overlapping clusters of gene sets that align with many of the enriched pathways identified in the enriched genes. Figure 3.13 shows cluster of positively enriched gene sets for xenobiotic transformation, peroxisomal β-oxidation, and various amino acid and other metabolic functions in the Outflow group compared to Upstream, while mRNA splicing and RNA binding clusters were negatively enriched.

Additional gene sets can also be identified in the GSEA data. A cluster of overlapping gene sets for cytoskeleton arrangement and actin binding was also enriched in the Upstream group compared to both the Outflow and Downstream groups, as was a cluster contains gene sets corresponding to 'M Phase' and 'Signaling by Rho GTPases.' Changes in actin binding and cytoskeleton organization are consistent with the downregulation of MAPRE3B and KRT8 in both the Outflow and Downstream group compared to upstream. Rho GTPases are important regulators of actin, and mitotic phase involves massive reorganization of the cell.[168] This suggests that cellular growth and organization is impacted downstream of the WWTPs.

GSEA also revealed that similar gene set enrichment occurred in the Downstream group compared to the Upstream, despite few differentially expressed genes after FDR adjustment. As can be seen in Figure 3.13, the vast majority of gene set positively or negatively enriched in the Upstream group show the same expression pattern in the Downstream group. This suggests that the changes observed in the GLEN site are continuing downstream at a reduced level, which is consistent with the observed concentrations of various contaminants at the sites (Table B.1).

Finally, 4 gene set clusters show opposing expression patterns between the Outflow and Downstream groups compared to Upstream. The clusters labeled 'CDH1 Replication Control', 'Protein Glycosylation', 'Translation Regulation', and 'Protein Modification' are all downregulated in the Outflow group, but upregulated in the Downstream. Cadherin-1, or E-cadherin is important in regulating cell division and degrading proteins during the end of mitosis, and plays a role as a tumor suppressor. Signaling pathways related to E-cadherin involve ubiquitination, phosphorylation and other protein modifications, as well a transcriptional regulation[169], and gene set clusters with these functions are differentially enriched between the Outflow and Downstream groups as well. It is possible that the increased oxidative stress and xenobiotic exposure at the GLEN site is impacting cell division through oxidative damage to DNA or other effects.

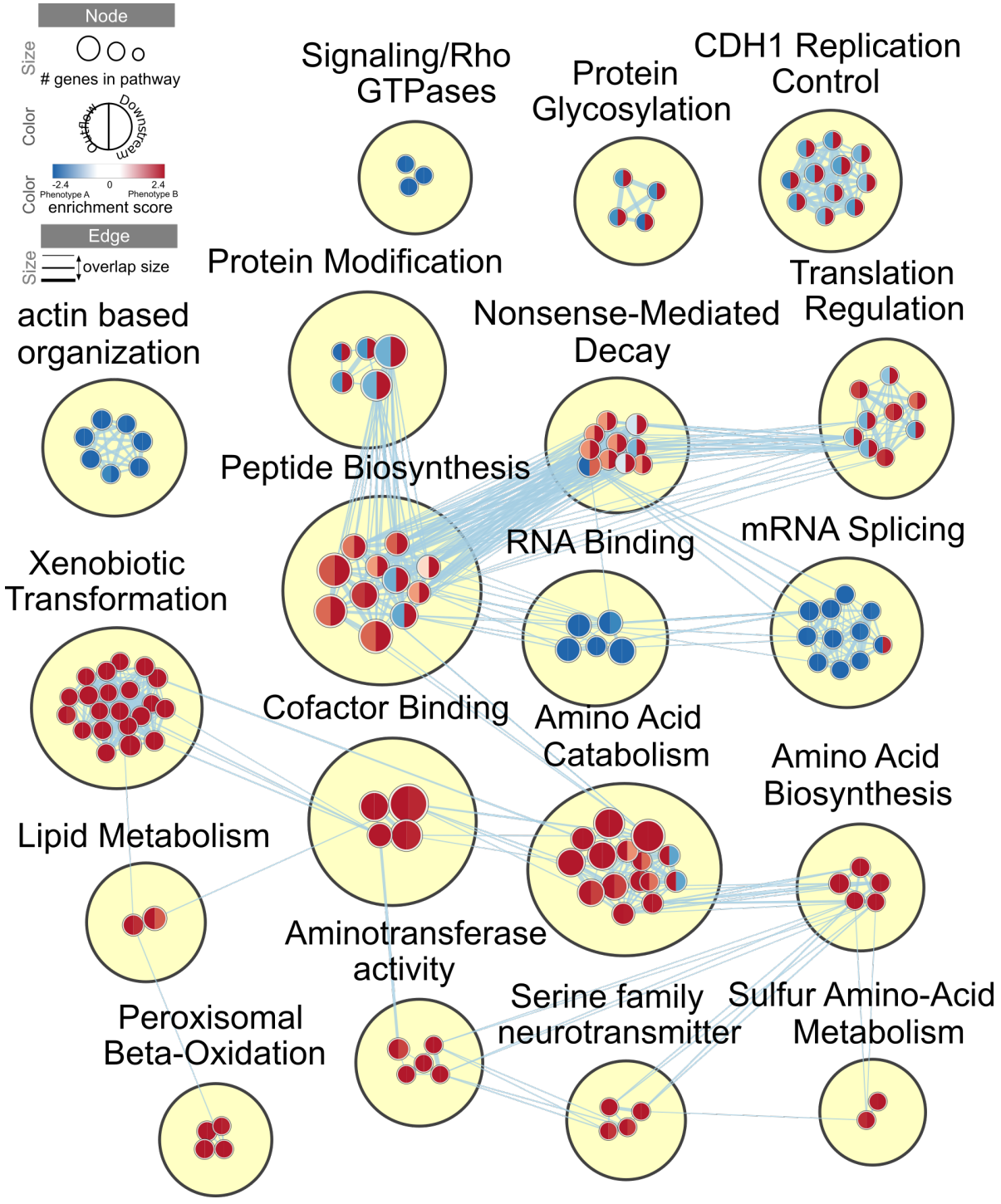


Figure 3.13: Comparison of enriched gene sets at an FDR of 20%. Gene sets from GO biological process and molecular function, as well as Reactome pathways were scored for enrichment using GSEA against the 895 proteins with expression data, and sets were then combined using automatic annotation analysis. The left side of each node indicates the enrichment of that set in the Outflow group vs the Upstream group, and the right side indicates the Downstream group vs Upstream.

3.6 Conclusions

This study has identified 3698 protein sequences from the FHM liver proteome by mass spectrometry. From the 895 proteins within the thresholds for quantification, 164 were significantly differentially expressed between the three Upstream, Outflow, and Downstream groups. GSEA and functional analysis of differentially expressed proteins shows that proteins involved in xenobiotic transformation, oxidative stress, and lipid metabolism were increased in sites downstream of the WWTPs, while proteins involved in mRNA splicing and cytoskeleton arrangement were decreased. Despite multiple confounding factors between sites, including variations in temperature and nutrient availability, the response was generally consistent with the concentrations of various PPCPs at the sites.

While many of the observed changes are consistent with both effluent concentration and other studies of MWWE and contaminant exposure, care should be taken when interpreting the results. While the comparison is between sites upstream and downstream of the WWTPs, a variety of factors other than MWWE exposure could impact the hepatic proteome. This includes the contaminants from other sources such as urban and rural runoff and stormwater outfalls, and changes in nutrient availability and water quality downstream of the WWTPs. Since proteome changes can be sex-specific, the different ratio of male to female fish across sites may impact the observed changes. Additionally, while the grouping of the caging sites for differential expression analysis improves the statistical power by reducing the number of comparisons, it also means that sites with differing conditions are combined into one group. In particular, CUSH receives urban and rural runoff from Calgary and surrounding communities that is not present in the other upstream site, BEAR. Lazaro-Côté *et al.* [12] found that whole-body glycogen content and condition factor was significantly higher for fish caged at CUSH compared to BEAR, while the water quality parameters and select PPCP concentrations are similar between theses sites (Tables B.2 and B.1, also Lazaro-Côté *et al.* [12]). This suggests that other, non-effluent contaminant sources may be impacted the caged fish, and proteome changes should not be directly ascribed to MWWE exposure.

While the large changes in CYP and oxidative stress protein abundance would suggest them as potential biomarkers, the wide variety of CYP proteins, and the variability of oxidative stress proteins limit their usefulness. Instead, examining lipid peroxidation or mitochondrial capacity of exposed fish may provide a more accurate biomarker of the impacts of exposure.[27,142] The effect on proteins involve in mRNA splicing suggest that investigation of the effects of MWWE exposure on transcript regulation may yield further insight.

Finally, gene sets involved in CDH1 cell cycle control, protein modification, and translation initiation were differentially enriched between fish at the Glen site and sites further downstream. Further study of post-translational modifications in the proteome would be beneficial to understand how regulatory

pathways are being altered.

Chapter 4

Conclusions and Future Directions

This thesis describes the use of proteomics to investigate changes in the fathead minnow (FHM) hepatic proteome downstream of wastewater treatment plants (WWTPs) in the Bow River in Calgary, Alberta. Different methods and tools for the identification of protein sequences from mass spectrometry (MS) data in a non-model organism were compared to determine the best method for accurate proteomics analysis. The method was then applied to fish liver samples generated from a caging experiment in the Bow River in the Calgary area. Changes in the liver proteome revealed increased xenobiotic metabolism, oxidative stress, and peroxisomal lipid metabolism, as well as changes in mitochondrial activity, mRNA splicing, and cadherin related cell cycle control in fish downstream of the WWTPs.

Chapter 2 investigates the selection of a protein database for the identification of proteins from FHM liver MS samples. Comparison between a protein collection (the UniRef90 Cyprinidae Proteome (U90CYP)), a reference proteome from another organism (the Uniprot Zebrafish Reference Proteome (ZRP)), and a genome derived database (the Fathead Minnow Predicted Proteome (FHMP)) revealed that more proteins could be identified when using a genome-derived database. This suggests two key points for the use of proteomics, particularly when studying non-model organisms:

1. The protein database requires careful selection. While proteins can be identified using a cross-species database, few identical peptides were found in both the FHM genome and the U90CYP. This will reduce the number proteins that can be identified, potentially preventing identification of biologically relevant proteins. It will also bias identified proteins toward more conserved proteins, potentially missing species-specific adaptations.

2. The integrations of omics data is beneficial for ecotoxicology. By building off previous genome sequencing work[59,60], several hundred additional proteins were identified, improving the ability to

identify relevant changes in the proteome. While genome sequence data is not available for many non-model organism, tools like ProteomeGenerator[67] make the integration of transcriptomic and proteomic data increasingly accessible.

An interesting possibility for the second point is the ability to iteratively improve genome annotations using mass spectrometry data. Proteomics has previously been used to improve and refine gene and exon models in genome sequence data (for example, [81]). By mapping identified peptides back to the genome, gene prediction models and exon boundaries can be improved. This could also use peptides from *de novo* sequencing tools such as PEAKS[50] to identify missing or novel genes. The improved gene database could then be used for future proteomics experiment with more accurate results. Transcriptomics data could also be included to further improve the database.

Chapter 3 compares protein expression in the fathead minnow liver across several sites in the Bow River. Expression changes were highest in the GLEN site, which is expected given the elevated temperatures and high concentration of effluent compared to other sites. Proteins involved in xenobiotic metabolism, oxidative stress response, and peroxisomal lipid metabolism were upregulated. Many of the proteins, such as glutathione S-transferases, have been previously studied as markers of xenobiotic exposures. The increase in peroxisomal lipid metabolism proteins strongly suggests exposure to a peroxisome proliferator-activated receptor (PPAR) inducer, such as fibrates or statins.[10] The downregulation of mitochondrial β-oxidation, and tricarboxylic acid cycle (TCA) proteins is an interesting find in the presence of increased peroxisomal lipid metabolism. The TCA is a key metabolic pathway, and downregulation in the nutrient-enriched WWTP outflow is unexpected. Additionally, the protein UCP1 was highly upregulated, which suggests 'decoupling' of mitochondrial oxidative phosphorylation from ATP generation and changes in mitochondrial redox status.[164] ROS management and lipid metabolism are key roles of peroxisomes[32], and there is significant regulatory interplay between peroxisomes and mitochondria.[24] Exposure to PPAR activators could be increasing lipid metabolism, which along with the effects of increased xenobiotic metabolism, lead sustained oxidative stress. The upregulation of UCP1 may be the effect of PPAR activation or increased ROS burden. The alteration to mitochondrial proteins in fish at the GLEN site suggests testing mitochondrial activity as a potentially useful biomarker, especially when PPAR activators such as statins or fibrates are expected in WWTP effluent.

While few proteins were differentially expressed in the sites further downstream, proteins that were differentially expressed showed a similar expression profile to the GLEN site. Additionally, Gene Set Enrichment Analysis (GSEA) showed a generally consistent response between the GLEN and the downstream sites. However, gene sets for cadherin-related cell cycle control, protein modification and translation initiation were downregulated in the GLEN site, but upregulated further downstream. A possible explanation

of this difference is that the increased reactive oxygen species (ROS) production in fish in the GLEN site induced sufficient oxidative damage to DNA, RNA, and cellular membranes[32] to impact cell cycle and growth. Further investigation of RNA splicing and protein modification after municipal wastewater effluent (MWWE) exposure could provide more insight into affected pathways for signaling pathways and cellular regulation.

References

1. Chambers, P. *et al.* The Impacts of Municipal Wastewater Effluents on Canadian Waters: A Review. *Water Quality Research Journal of Canada* **32,** 659–713 (1997).

2. Holeton, C., Chambers, P. A., Grace, L. & Kidd, K. Wastewater Release and Its Impacts on Canadian Waters. *Canadian Journal of Fisheries and Aquatic Sciences* **68,** 1836–1859 (Oct. 2011).

3. Metcalfe, C. D. *et al.* Antidepressants and Their Metabolites in Municipal Wastewater, and Downstream Exposure in an Urban Watershed. *Environmental Toxicology and Chemistry* **29,** 79–89 (Jan. 2010).

4. Krzeminski, P. *et al.* Performance of Secondary Wastewater Treatment Methods for the Removal of Contaminants of Emerging Concern Implicated in Crop Uptake and Antibiotic Resistance Spread: A Review. *Science of The Total Environment* **648,** 1052–1081 (Jan. 2019).

5. Carballa, M. *et al.* Behavior of Pharmaceuticals, Cosmetics and Hormones in a Sewage Treatment Plant. *Water Research* **38,** 2918–2926 (July 2004).

6. Kasprzyk-Hordern, B., Dinsdale, R. M. & Guwy, A. J. The Removal of Pharmaceuticals, Personal Care Products, Endocrine Disruptors and Illicit Drugs during Wastewater Treatment and Its Impact on the Quality of Receiving Waters. *Water Research* **43,** 363–380 (Feb. 2009).

7. McCance, W. *et al.* Contaminants of Emerging Concern as Novel Groundwater Tracers for Delineating Wastewater Impacts in Urban and Peri-Urban Areas. *Water Research* **146,** 118–133 (Dec. 2018).

8. Nelson, E. D., Do, H., Lewis, R. S. & Carr, S. A. Diurnal Variability of Pharmaceutical, Personal Care Product, Estrogen and Alkylphenol Concentrations in Effluent from a Tertiary Wastewater Treatment Facility. *Environmental Science & Technology* **45,** 1228–1234 (Feb. 2011).

9. Reinling, J., Houde, M. & Verreault, J. Environmental Exposure to a Major Urban Wastewater Effluent: Effects on the Energy Metabolism of Northern Pike. *Aquatic Toxicology* **191,** 131–140 (Oct. 2017).

10. Olivares-Rubio, H. F. & Vega-López, A. Fatty Acid Metabolism in Fish Species as a Biomarker for Environmental Monitoring. *Environmental Pollution* **218,** 297–312 (Nov. 2016).

11. Chen, M. *et al.* Pharmaceuticals and Endocrine Disruptors in Wastewater Treatment Effluents and in the Water Supply System of Calgary, Alberta, Canada. *Water Quality Research Journal of Canada* **41,** 351–364 (Nov. 2006).

12. Lazaro-Côté, A., Sadoul, B., Jackson, L. J. & Vijayan, M. M. Acute Stress Response of Fathead Minnows Caged Downstream of Municipal Wastewater Treatment Plants in the Bow River, Calgary. *PLOS ONE* **13** (ed Soengas, J. L.) e0198177 (June 2018).

13. Lacey, C., Basha, S., Morrissey, A. & Tobin, J. M. Occurrence of Pharmaceutical Compounds in Wastewater Process Streams in Dublin, Ireland. *Environmental Monitoring and Assessment* **184,** 1049–1062 (Feb. 2012).

14. Fuzzen, M. L. M. *et al.* An Assessment of the Spatial and Temporal Variability of Biological Responses to Municipal Wastewater Effluent in Rainbow Darter (*Etheostoma Caeruleum*) Collected along an Urban Gradient. *PLOS ONE* **11** (ed Meador, J. P.) e0164879 (Oct. 2016).

15. Sowers, A. D., Mills, M. A. & Klaine, S. J. The Developmental Effects of a Municipal Wastewater Effluent on the Northern Leopard Frog, *Rana Pipiens*. *Aquatic Toxicology* **94,** 145–152 (Aug. 2009).

16. Simmons, D. B. D. *et al.* Altered Expression of Metabolites and Proteins in Wild and Caged Fish Exposed to Wastewater Effluents in Situ. *Scientific Reports* **7** (Dec. 2017).

17. Bui-Nguyen, T. M. *et al.* Dichlorvos Exposure Results in Large Scale Disruption of Energy Metabolism in the Liver of the Zebrafish, *Danio Rerio*. *BMC Genomics* **16** (Dec. 2015).

18. Corcoran, J., Winter, M. J. & Tyler, C. R. Pharmaceuticals in the Aquatic Environment: A Critical Review of the Evidence for Health Effects in Fish. *Critical Reviews in Toxicology* **40,** 287–304 (Apr. 2010).

19. Martyniuk, C. J., Kroll, K. J., Doperalski, N. J., Barber, D. S. & Denslow, N. D. Environmentally Relevant Exposure to 17α-Ethinylestradiol Affects the Telencephalic Proteome of Male Fathead Minnows. *Aquatic Toxicology* **98,** 344–353 (July 2010).

20. Giang, P. T. *et al.* Biomarker Response, Health Indicators, and Intestinal Microbiome Composition in Wild Brown Trout (*Salmo Trutta* m. *Fario* L.) Exposed to a Sewage Treatment Plant Effluent-Dominated Stream. *Science of The Total Environment* **625,** 1494–1509 (June 2018).

21. Al-Habsi, A. A., Massarsky, A. & Moon, T. W. Exposure to Gemfibrozil and Atorvastatin Affects Cholesterol Metabolism and Steroid Production in Zebrafish (*Danio Rerio*). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **199,** 87–96 (Sept. 2016).

22. Poulsen, L. l. C., Siersbæk, M. & Mandrup, S. PPARs: Fatty Acid Sensors Controlling Metabolism. *Seminars in Cell & Developmental Biology* **23,** 631–639 (Aug. 2012).

23. Moon, T. W., Walsh, P. J. & Mommsen, T. P. Fish Hepatocytes: A Model Metabolic System. *Canadian Journal of Fisheries and Aquatic Sciences* **42,** 1772–1782 (Nov. 1985).

24. Fransen, M., Lismont, C. & Walton, P. The Peroxisome-Mitochondria Connection: How and Why? *International Journal of Molecular Sciences* **18,** 1126 (May 2017).

25. Poirier, Y., Antonenkov, V. D., Glumoff, T. & Hiltunen, J. K. Peroxisomal β-Oxidation—A Metabolic Pathway with Multiple Functions. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1763,** 1413–1426 (Dec. 2006).

26. Uno, T., Ishizuka, M. & Itakura, T. Cytochrome P450 (CYP) in Fish. *Environmental Toxicology and Pharmacology* **34,** 1–13 (July 2012).

27. Kroon, F., Streten, C. & Harries, S. A Protocol for Identifying Suitable Biomarkers to Assess Fish Health: A Systematic Review. *PLOS ONE* **12** (ed Meador, J. P.) e0174762 (Apr. 2017).

28. Lee, B.-Y. *et al.* Identification of 74 Cytochrome P450 Genes and Co-Localized Cytochrome P450 Genes of the CYP2K, CYP5A, and CYP46A Subfamilies in the Mangrove Killifish *Kryptolebias Marmoratus*. *BMC Genomics* **19** (Dec. 2018).

29. Goldstone, J. V. *et al.* Identification and Developmental Expression of the Full Complement of Cytochrome P450 Genes in Zebrafish. *BMC Genomics* **11,** 643 (2010).

30. Lepesheva, G. I. & Waterman, M. R. Sterol 14α-Demethylase Cytochrome P450 (CYP51), a P450 in All Biological Kingdoms. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1770,** 467–477 (Mar. 2007).

31. Puthumana, J. *et al.* Nine Co-Localized Cytochrome P450 Genes of the *CYP2N* , *CYP2AD*, and *CYP2P* Gene Families in the Mangrove Killifish *Kryptolebias Marmoratus* Genome: Identification and Expression in Response to B[α]P, BPA, OP, and NP. *Aquatic Toxicology* **187,** 132–140 (June 2017).

32. Walker, C. L., Pomatto, L. C. D., Tripathi, D. N. & Davies, K. J. A. Redox Regulation of Homeostasis and Proteostasis in Peroxisomes. *Physiological Reviews* **98,** 89–115 (Jan. 2018).

33. Mommsen, T. P., Vijayan, M. M. & Moon, T. W. Cortisol in Teleosts: Dynamics, Mechanisms of Action, and Metabolic Regulation. *Reviews in Fish Biology and Fisheries* **9,** 211–268 (1999).

34. Dindia, L. *et al.* Novel Nongenomic Signaling by Glucocorticoid May Involve Changes to Liver Membrane Order in Rainbow Trout. *PLOS ONE* **7** (ed Fuentes, J.) e46859 (Oct. 2012).

35. Ings, J. S., Servos, M. R. & Vijayan, M. M. Exposure to Municipal Wastewater Effluent Impacts Stress Performance in Rainbow Trout. *Aquatic Toxicology* **103,** 85–91 (May 2011).

36. Faught, E. & Vijayan, M. M. Mechanisms of Cortisol Action in Fish Hepatocytes. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **199,** 136–145 (Sept. 2016).

37. Alderman, S. L., McGuire, A., Bernier, N. J. & Vijayan, M. M. Central and Peripheral Glucocorticoid Receptors Are Involved in the Plasma Cortisol Response to an Acute Stressor in Rainbow Trout. *General and Comparative Endocrinology* **176,** 79–85 (Mar. 2012).

38. Karim, M., Puiseux-Dao, S. & Edery, M. Toxins and Stress in Fish: Proteomic Analyses and Response Network. *Toxicon* **57,** 959–969 (June 2011).

39. Philip, A. M. & Vijayan, M. M. Stress-Immune-Growth Interactions: Cortisol Modulates Suppressors of Cytokine Signaling and JAK/STAT Pathway in Rainbow Trout Liver. *PLOS ONE* **10** (ed Boudinot, P.) e0129299 (June 2015).

40. Dindia, L., Faught, E., Leonenko, Z., Thomas, R. & Vijayan, M. M. Rapid Cortisol Signaling in Response to Acute Stress Involves Changes in Plasma Membrane Order in Rainbow Trout Liver. *American Journal of Physiology - Endocrinology and Metabolism* **304,** E1157–E1166 (June 2013).

41. Kling, P., Norman, A., Andersson, P., Norrgren, L. & Förlin, L. Gender-Specific Proteomic Responses in Zebrafish Liver Following Exposure to a Selected Mixture of Brominated Flame Retardants, *Ecotoxicology and Environmental Safety* **71,** 319–327 (Oct. 2008).

42. Ings, J. S., Servos, M. R. & Vijayan, M. M. Hepatic Transcriptomics and Protein Expression in Rainbow Trout Exposed to Municipal Wastewater Effluent. *Environmental Science & Technology* **45,** 2368–2376 (Mar. 2011).

43. Bonga, S. W. The Stress Response in Fish. *Physiological reviews* **77,** 591–625 (1997).

44. Groh, K. J. & Suter, M. J. .-.-F. Stressor-Induced Proteome Alterations in Zebrafish: A Meta-Analysis of Response Patterns. *Aquatic Toxicology* **159,** 1–12 (Feb. 2015).

45. Aebersold, R. & Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **422,** 198–207 (Mar. 2003).

46. Aebersold, R. & Mann, M. Mass-Spectrometric Exploration of Proteome Structure and Function. *Nature* **537,** 347–355 (Sept. 2016).

47. Walther, T. C. & Mann, M. Mass Spectrometry–Based Proteomics in Cell Biology. *The Journal of Cell Biology* **190,** 491–500 (Aug. 2010).

48. Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of Proteins and Proteomes by Mass Spectrometry. *Annual review of biochemistry* **70,** 437–473 (2001).

49. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar Acquisition Method Enables Single-Shot Proteomics at a Depth of 10,000 Proteins in 100 Minutes. *Nature Methods* **15,** 440–448 (June 2018).

50. Zhang, J. *et al.* PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol Cell Proteomics* **11** (Apr. 2012).

51. Ma, K., Vitek, O. & Nesvizhskii, A. I. A Statistical Model-Building Perspective to Identification of MS/MS Spectra with PeptideProphet. *BMC Bioinformatics* **13,** S1 (2012).

52. Elias, J. E. & Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nature Methods* **4,** 207–214 (Mar. 2007).

53. Shteynberg, D. *et al.* iProphet: Multi-Level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Molecular & Cellular Proteomics* **10,** M111.007690 (Dec. 2011).

54. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Analytical Chemistry* **75,** 4646–4658 (Sept. 2003).

55. Eng, J. K. *et al.* A Deeper Look into Comet—Implementation and Features. *Journal of The American Society for Mass Spectrometry* **26,** 1865–1874 (Nov. 2015).

56. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nature Methods* **14,** 513–520 (Apr. 2017).

57. Li, H. *et al.* Evaluating the Effect of Database Inflation in Proteogenomic Search on Sensitive and Reliable Peptide Identification. *BMC Genomics* **17** (Dec. 2016).

58. Blakeley, P., Overton, I. M. & Hubbard, S. J. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *Journal of Proteome Research* **11,** 5221–5234 (Nov. 2012).

59. Burns, F. R. *et al.* Sequencing and de Novo Draft Assemblies of a Fathead Minnow (*Pimephales Promelas*) Reference Genome. *Environ Toxicol Chem* **35,** 212–217 (Jan. 2016).

60. Saari, T. W., Schroeder, A. L., Ankley, G. T. & Villeneuve, D. L. First-Generation Annotations for the Fathead Minnow (*Pimephales Promelas*) Genome. *Environmental Toxicology and Chemistry* **36,** 3436–3442 (Dec. 2017).

61. Rauniyar, N. & Yates, J. R. Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *Journal of Proteome Research* **13,** 5293–5309 (Dec. 2014).

62. Liu, J., Sweredoski, M. & Hess, S. Improved 6-Plex Tandem Mass Tags Quantification Throughput Using a Linear Ion Trap-High-Energy Collision Induced Dissociation MS3 Scan. *Analytical Chemistry* **88,** 7471–7475 (2016).

63. Martinez-Val, A. *et al.* On the Statistical Significance of Compressed Ratios in Isobaric Labeling: A Cross-Platform Comparison. *Journal of Proteome Research* **15,** 3029–3038 (Sept. 2016).

64. Pascovici, D. *et al.* Combining Protein Ratio *P*-Values as a Pragmatic Approach to the Analysis of Multirun iTRAQ Experiments. *Journal of Proteome Research* **14,** 738–746 (Feb. 2015).

65. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 Eliminates Ratio Distortion in Isobaric Multiplexed Quantitative Proteomics. *Nature Methods* **8,** 937–940 (Oct. 2011).

66. McAlister, G. C. *et al.* MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Analytical Chemistry* **86,** 7150–7158 (July 2014).

67. Cifani, P. *et al.* ProteomeGenerator: A Framework for Comprehensive Proteomics Based on de Novo Transcriptome Assembly and High-Accuracy Peptide Mass Spectral Matching. *J. Proteome Res.* **17,** 3681–3692 (Nov. 2018).

68. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165,** 535–550 (Apr. 2016).

69. Ruggles, K. V. *et al.* Methods, Tools and Current Perspectives in Proteogenomics. *Molecular & Cellular Proteomics* **16,** 959–981 (June 2017).

70. Armengaud, J. *et al.* Non-Model Organisms, a Species Endangered by Proteogenomics. *Journal of Proteomics* **105,** 5–18 (June 2014).

71. Ankley, G. T. & Villeneuve, D. L. The Fathead Minnow in Aquatic Toxicology: Past, Present and Future. *Aquatic Toxicology* **78,** 91–102 (June 2006).

72. The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Research* **47,** D506–D515 (Jan. 2019).

73. Menschaert, G. & Fenyö, D. Proteogenomics from a Bioinformatics Angle: A Growing Field. *Mass Spectrometry Reviews* **36,** 584–599 (Sept. 2017).

74. Zickmann, F. & Renard, B. Y. MSProGene: Integrative Proteogenomics beyond Six-Frames and Single Nucleotide Polymorphisms. *Bioinformatics* **31,** i106–i115 (June 2015).

75. Omasits, U. *et al.* An Integrative Strategy to Identify the Entire Protein Coding Potential of Prokaryotic Genomes by Proteogenomics. *Genome Research* **27,** 2083–2095 (Dec. 2017).

76. Mitchell, N. M. *et al.* Proteogenomic Re-Annotation of *Coccidioides Posadasii* Strain Silveira. *Proteomics* **18,** 1700173 (Jan. 2018).

77. Nesvizhskii, A. I. Proteogenomics: Concepts, Applications and Computational Strategies. *Nat Methods* **11,** 1114–1125 (Nov. 2014).

78. Ivanov, M. V., Lobas, A. A., Levitsky, L. I., Moshkovskii, S. A. & Gorshkov, M. V. Brute-Force Approach for Mass Spectrometry-Based Variant Peptide Identification in Proteogenomics without Personalized Genomic Data. *Journal of The American Society for Mass Spectrometry* **29,** 435–438 (Feb. 2018).

79. Brosch, M. *et al.* Shotgun Proteomics Aids Discovery of Novel Protein-Coding Genes, Alternative Splicing, and "Resurrected" Pseudogenes in the Mouse Genome. *Genome Research* **21,** 756–767 (May 2011).

80. Low, T. Y. *et al.* Quantitative and Qualitative Proteome Characteristics Extracted from In-Depth Integrated Genomics and Proteomics Analysis. *Cell Reports* **5,** 1469–1478 (Dec. 2013).

81. Ye, X. *et al.* Improving Silkworm Genome Annotation Using a Proteogenomics Approach. *J. Proteome Res.,* acs.jproteome.8b00965 (July 2019).

82. Fröhlich, T., Arnold, G. J., Fritsch, R., Mayr, T. & Laforsch, C. LC-MS/MS-Based Proteome Profiling in *Daphnia Pulex* and *Daphnia Longicephala*: The *Daphnia Pulex* Genome Database as a Key for High Throughput Proteomics in Daphnia. *BMC Genomics* **10,** 171 (2009).

83. Diz, A. P., Dudley, E. & Skibinski, D. O. F. Identification and Characterization of Highly Expressed Proteins in Sperm Cells of the Marine Mussel *Mytilus Edulis*. *Proteomics* **12,** 1949–1956 (June 2012).

84. Suzek, B. E. *et al.* UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics* **31,** 926–932 (Mar. 2015).

85. *Fathead Minnow Genome Project* https://www.setac.org/page/fhmgenome.

86. Dale, R. *Gffutils*

87. Haas, B. & Papanicolaou, A. *TransDecoder* May 2019.

88. Camacho, C. *et al.* BLAST+: Architecture and Applications. *BMC Bioinformatics* **10,** 421 (2009).

89. El-Gebali, S. *et al.* The Pfam Protein Families Database in 2019. *Nucleic Acids Research* **47,** D427–D432 (Jan. 2019).

90. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions. *Nucleic Acids Research* **41,** e121–e121 (July 2013).

91. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* **35,** 543–548 (Mar. 2018).

92. Dohmen, E., Kremer, L. P., Bornberg-Bauer, E. & Kemena, C. DOGMA: Domain-Based Transcriptome and Proteome Quality Assessment. *Bioinformatics* **32,** 2577–2581 (Sept. 2016).

93. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **28,** 3150–3152 (Dec. 2012).

94. Madeira, F. *et al.* The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Research* **47,** W636–W641 (July 2019).

95. Deutsch, E. W. *et al.* Trans-Proteomic Pipeline, a Standardized Data Processing Pipeline for Large-Scale Reproducible Proteomics Informatics. *Prot. Clin. Appl.* **9,** 745–754 (Aug. 2015).

96. Craig, R. & Beavis, R. C. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics* **20,** 1466–1467 (June 2004).

97. *Proteome Bioinformatics* (eds Hubbard, S. J. & Jones, A. R.) *Springer Protocols* **604.** OCLC: ocn428029505. ISBN: 978-1-60761-443-2 978-1-60761-444-9 (Humana, New York, NY, 2010).

98. Moss, S. P., Joyce, D. A., Humphries, S., Tindall, K. J. & Lunt, D. H. Comparative Analysis of Teleost Genome Sequences Reveals an Ancient Intron Size Expansion in the Zebrafish Lineage. *Genome Biology and Evolution* **3,** 1187–1196 (Jan. 2011).

99. Howbert, J. J. & Noble, W. S. Computing Exact *P*-Values for a Cross-Correlation Shotgun Proteomics Score Function. *Molecular & Cellular Proteomics* **13,** 2467–2479 (Sept. 2014).

100. Fenyö, D., Eriksson, J. & Beavis, R. in *Computational Biology* (ed Fenyö, D.) 189–202 (Humana Press, Totowa, NJ, 2010). ISBN: 978-1-60761-841-6 978-1-60761-842-3.

101. Gillet, L. C., Leitner, A. & Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annual Review of Analytical Chemistry* **9,** 449–472 (2016).

102. Yin, L., Ma, R., Wang, B., Yuan, H. & Yu, G. The Degradation and Persistence of Five Pharmaceuticals in an Artificial Climate Incubator during a One Year Period. *RSC Advances* **7,** 8280–8287 (2017).

103. Tocher, D. R. Metabolism and Functions of Lipids and Fatty Acids in Teleost Fish. *Reviews in Fisheries Science* **11,** 107–184 (Apr. 2003).

104. Al-Habsi, A. A., Massarsky, A. & Moon, T. W. Atorvastatin Alters Gene Expression and Cholesterol Synthesis in Primary Rainbow Trout (*Oncorhynchus Mykiss*) Hepatocytes. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **224,** 262–269 (Oct. 2018).

105. Weston, A., Caminada, D., Galicia, H. & Fent, K. Effects of Lipid-lowering Pharmaceuticals Bezafibrate and Clofibric Acid on Lipid Metabolism in Fathead Minnow (*Pimephales Promelas*). *Environmental Toxicology and Chemistry* **28,** 2648 (2009).

106. Cajaraville, M. P., Cancio, I., Ibabe, A. & Orbea, A. Peroxisome Proliferation as a Biomarker in Environmental Pollution Assessment. *Microscopy Research and Technique* **61,** 191–202 (June 2003).

107. Buhler, D. R. & Wang-Buhler, J.-L. Rainbow Trout Cytochrome P450s: Purification, Molecular Aspects, Metabolic Activity, Induction and Role in Environmental Monitoring. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology* **121,** 107–137 (Nov. 1998).

108. Lazaro-Côté, A. The Effect of Municipal Wastewater Effluent on the Stress Response of Native Fish Species in the Bow River, Calgary, Alberta (2017).

109. Erde, J., Loo, R. R. O. & Loo, J. A. Enhanced FASP (eFASP) to Increase Proteome Coverage and Sample Recovery for Quantitative Proteomic Experiments. *J. Proteome Res.* **13,** 1885–1895 (Apr. 2014).

110. Chambers, M. C. *et al.* A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nature Biotechnology* **30,** 918–920 (Oct. 2012).

111. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2019).

112. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-3-319-24277-4 (Springer-Verlag New York, 2016).

113. Ritchie, M. E. *et al.* Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Research* **43,** e47–e47 (Apr. 2015).

114. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** (1995).

115. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package biomaRt. *Nature Protocols* **4,** 1184–1191 (2009).

116. ZFIN Staff* *et al.* in *Eukaryotic Genomic Databases* (ed Kollmar, M.) 307–347 (Springer New York, New York, NY, 2018). ISBN: 978-1-4939-7736-9 978-1-4939-7737-6.

117. Doncheva, N. T., Morris, J. H., Gorodkin, J. & Jensen, L. J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **18,** 623–632 (Feb. 2019).

118. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nature Protocols* **4,** 44 (Dec. 2008).

119. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Research* **37,** 1–13 (Jan. 2009).

120. Aibar, S., Fontanillo, C., Droste, C. & De Las Rivas, J. Functional Gene Networks: R/Bioc Package to Generate and Analyse Gene Networks Derived from Functional Enrichment and Clustering. *Bioinformatics* **31,** 1686–1688 (May 2015).

121. Subramanian, A. *et al.* Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (Oct. 2005).

122. The Gene Ontology Consortium. The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Research* **47,** D330–D338 (Jan. 2019).

123. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46,** D649–D655 (Jan. 2018).

124. Isserlin, R., Merico, D., Voisin, V. & Bader, G. D. Enrichment Map – a Cytoscape App to Visualize and Explore OMICs Pathway Enrichment Results. *F1000Res* **3,** 141 (July 2014).

125. van der Oost, R., Beyer, J. & Vermeulen, N. P. Fish Bioaccumulation and Biomarkers in Environmental Risk Assessment: A Review. *Environmental Toxicology and Pharmacology* **13,** 57–149 (Feb. 2003).

126. Kim, B.-M. *et al.* Effects of Benzo[a]Pyrene on Whole Cytochrome P450-Involved Molecular Responses in the Marine Medaka *Oryzias Melastigma*. *Aquatic Toxicology* **152,** 232–243 (July 2014).

127. Rhee, J.-S. *et al.* Whole Spectrum of Cytochrome P450 Genes and Molecular Responses to Water-Accommodated Fractions Exposure in the Marine Medaka. *Environmental Science & Technology* **47,** 4804–4812 (May 2013).

128. Kubota, A., Bainy, A. C., Woodin, B. R., Goldstone, J. V. & Stegeman, J. J. The Cytochrome P450 2AA Gene Cluster in Zebrafish (*Danio Rerio*): Expression of CYP2AA1 and CYP2AA2 and Response to Phenobarbital-Type Inducers. *Toxicology and Applied Pharmacology* **272,** 172–179 (Oct. 2013).

129. Mosadeghi, S., Furnes, B., Matsuo, A. Y. & Schlenk, D. Expression and Characterization of Cytochrome P450 2X1 in Channel Catfish (*Ictalurus Punctatus*). *Biochimica et Biophysica Acta (BBA) - General Subjects* **1770,** 1045–1052 (July 2007).

130. Wang, J., Wei, Y., Li, X., Xu, M. & Dai, J. Identification of Differentially Expressed Genes from Contaminant and Thermal Exposed Goldfish *Carassius Auratus* in Gaobeidian Lake in Beijing, China. *Ecotoxicology* **16,** 525–532 (Aug. 2007).

131. Oleksiak, M. F. *et al.* Identification, Functional Characterization, and Regulation of a New Cytochrome P450 Subfamily, the CYP2Ns. *Journal of Biological Chemistry* **275,** 2312–2321 (Jan. 2000).

132. Murray, M. CYP2J2 – Regulation, Function and Polymorphism. *Drug Metabolism Reviews* **48,** 351–368 (July 2016).

133. Dalvie, D. & Di, L. Aldehyde Oxidase and Its Role as a Drug Metabolizing Enzyme. *Pharmacology & Therapeutics,* 137–180 (May 2019).

134. Arimitsu, E. *et al.* Cloning and Sequencing of the cDNA Species for Mammalian Dimeric Dihydrodiol Dehydrogenases. *Biochemical Journal* **342,** 721–728 (1999).

135. Alazizi, A. *et al.* Identification, Characterization, and Ontogenic Study of a Catechol O-Methyltransferase from Zebrafish. *Aquatic Toxicology* **102,** 18–23 (Mar. 2011).

136. Glisic, B. *et al.* Characterization of Glutathione-S-Transferases in Zebrafish (*Danio Rerio*). *Aquatic Toxicology* **158,** 50–62 (Jan. 2015).

137. Brigelius-Flohé, R. & Maiorino, M. Glutathione Peroxidases. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1830,** 3289–3303 (May 2013).

138. Nevalainen, T. J. 1-Cysteine Peroxiredoxin: A Dual-Function Enzyme with Peroxidase and Acidic Ca2+-Independent Phospholipase A2 Activities. *Biochimie* **92,** 638–644 (June 2010).

139. Pajares, M. A. & Pérez-Sala, D. Betaine Homocysteine S-Methyltransferase: Just a Regulator of Homocysteine Metabolism? *Cell. Mol. Life Sci.* **63,** 2792–2803 (Dec. 2006).

140. Wang, Y.-P. *et al.* Regulation of G6PD Acetylation by KAT9/SIRT2 Modulates NADPH Homeostasis and Cell Survival during Oxidative Stress. *The EMBO Journal,* 1304–1320 (Apr. 2014).

141. Jasinska, E. J. *et al.* Assessment of Biomarkers for Contaminants of Emerging Concern on Aquatic Organisms Downstream of a Municipal Wastewater Discharge. *Science of The Total Environment* **530-531,** 140–153 (Oct. 2015).

142. Paskerová, H., Hilscherová, K. & Bláha, L. Oxidative Stress and Detoxification Biomarker Responses in Aquatic Freshwater Vertebrates Exposed to Microcystins and Cyanobacterial Biomass. *Environ Sci Pollut Res* **19,** 2024–2037 (July 2012).

143. Szklarczyk, D. *et al.* STRING V11: Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Research* **47,** D607–D613 (Jan. 2019).

144. Wanders, R. J. A., Waterham, H. R. & Ferdinandusse, S. Metabolic Interplay between Peroxisomes and Other Subcellular Organelles Including Mitochondria and the Endoplasmic Reticulum. *Front. Cell Dev. Biol.* **3** (Jan. 2016).

145. Drew, R. E. *et al.* Effect of Starvation on Transcriptomes of Brain and Liver in Adult Female Zebrafish (*Danio Rerio*). *Physiological Genomics* **35,** 283–295 (Nov. 2008).

146. Hu, C. *et al.* Retinol-Binding Protein 7 Is an Endothelium-Specific PPARγ Cofactor Mediating an Antioxidant Response through Adiponectin. *JCI Insight* **2** (Mar. 2017).

147. Lyche, J. L. *et al.* Natural Mixtures of Persistent Organic Pollutants (POP) Increase Weight Gain, Advance Puberty, and Induce Changes in Gene Expression Associated with Steroid Hormones and Obesity in Female Zebrafish. *Journal of Toxicology and Environmental Health, Part A* **73,** 1032–1057 (June 2010).

148. Hardwick, J. P. Cytochrome P450 Omega Hydroxylase (CYP4) Function in Fatty Acid Metabolism and Metabolic Diseases. *Biochemical Pharmacology* **75,** 2263–2275 (June 2008).

149. Jakobsson, A., Westerberg, R. & Jacobsson, A. Fatty Acid Elongases in Mammals: Their Regulation and Roles in Metabolism. *Progress in Lipid Research* **45,** 237–249 (May 2006).

150. Korbecki, J., Baranowska-Bosiacka, I., Gutowska, I. & Chlubek, D. The Effect of Reactive Oxygen Species on the Synthesis of Prostanoids from Arachidonic Acid. *Journal of Physiology and Pharmacology* **64,** 409–421 (2013).

151. Kriska, T., Pilat, A., Schmitt, J. C. & Girotti, A. W. Sterol Carrier Protein-2 (SCP-2) Involvement in Cholesterol Hydroperoxide Cytotoxicity as Revealed by SCP-2 Inhibitor Effects. *J. Lipid Res.* **51,** 3174–3184 (Nov. 2010).

152. Crumbley, C., Wang, Y., Banerjee, S. & Burris, T. P. Regulation of Expression of Citrate Synthase by the Retinoic Acid Receptor-Related Orphan Receptor α (RORα). *PLOS ONE* **7,** e33804 (Apr. 2012).

153. Schell, J. C. *et al.* A Role for the Mitochondrial Pyruvate Carrier as a Repressor of the Warburg Effect and Colon Cancer Cell Growth. *Molecular Cell* **56,** 400–413 (Nov. 2014).

154. Ellis, J. L. *et al.* Zebrafish *Abcb11b* Mutant Reveals Strategies to Restore Bile Excretion Impaired by Bile Salt Export Pump Deficiency. *Hepatology* **67,** 1531–1545 (Apr. 2018).

155. Pérez-Castrillón, J. L. *et al.* Polymorphisms of the Farnesyl Diphosphate Synthase Gene Modulate Bone Changes in Response to Atorvastatin. *Rheumatol Int* **34,** 1073–1077 (Aug. 2014).

156. Oettl, K., Höfler, G., Ness, G. C., Sattler, W. & Malle, E. An Apparent Decrease in Cholesterol Biosynthesis in Peroxisomal-Defective Chinese Hamster Ovary Cells Is Related to Impaired Mitochondrial Oxidation. *Biochemical and Biophysical Research Communications* **305,** 957–963 (June 2003).

157. Will, C. L. & Luhrmann, R. Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology* **3,** a003707–a003707 (July 2011).

158. Chanarat, S. & Sträßer, K. Splicing and beyond: The Many Faces of the Prp19 Complex. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1833,** 2126–2134 (Oct. 2013).

159. Morris, E. J. *et al.* Functional Identification of Api5 as a Suppressor of E2F-Dependent Apoptosis In Vivo. *PLOS Genetics* **2,** e196 (2006).

160. Zhang, D. *et al.* Identification of Annexin A4 as a Hepatopancreas Factor Involved in Liver Cell Survival. *Developmental Biology* **395,** 96–110 (Nov. 2014).

161. Mirsaeidi, M., Gidfar, S., Vu, A. & Schraufnagel, D. Annexins Family: Insights into Their Functions and Potential Role in Pathogenesis of Sarcoidosis. *Journal of Translational Medicine* **14** (Dec. 2016).

162. Fernandez-Capetillo, O., Allis, C. D. & Nussenzweig, A. Phosphorylation of Histone H2B at DNA Double-Strand Breaks. *The Journal of Experimental Medicine* **199,** 1671–1677 (June 2004).

163. Lacaze, E. *et al.* The Effects of Municipal Effluents on Oxidative Stress, Immunocompetence and DNA Integrity in Fathead Minnow Juveniles, 12 (2017).

164. Tseng, Y.-C. *et al.* Exploring Uncoupling Proteins and Antioxidant Mechanisms under Acute Cold Exposure in Brains of Fish. *PLoS ONE* **6** (ed Polymenis, M.) e18180 (Mar. 2011).

165. Wen, X. *et al.* iTRAQ-Based Quantitative Proteomic Analysis of *Takifugu Fasciatus* Liver in Response to Low-Temperature Stress. *Journal of Proteomics* **201,** 27–36 (June 2019).

166. Duffy, W. G. Population Dynamics, Production, and Prey Consumption of Fathead Minnows (*Pimephales Promelas*) in Prairie Wetlands: A Bioenergetics Approach. **55,** 13 (1998).

167. Jastroch, M., Wuertz, S., Kloas, W. & Klingenspor, M. Uncoupling Protein 1 in Fish Uncovers an Ancient Evolutionary History of Mammalian Nonshivering Thermogenesis. *Physiological Genomics* **22,** 150–156 (July 2005).

168. Shigetomi, K. & Ikenouchi, J. Cell Adhesion Structures in Epithelial Cells Are Formed in Dynamic and Cooperative Ways. *BioEssays* **41,** 1800227 (July 2019).

169. Heuberger, J. & Birchmeier, W. Interplay of Cadherin-Mediated Cell Adhesion and Canonical Wnt Signaling. *Cold Spring Harbor Perspectives in Biology* **2,** a002915–a002915 (Feb. 2010).

A.1 Code Used for Analysis

Listing A.1: Python code for converting gff3 annotations file to protein or transcript fasta files.

```
#!/usr/bin/env python
"""
Usage:
    gff_convert.py <gffutils_db_file> <ref_fasta> <output_directory>
"""

import os
import sys
import warnings

import argparse
import gffutils
from pyfaidx import Fasta

from Bio import SeqIO
from Bio.Seq import Seq
from Bio.SeqRecord import SeqRecord
from Bio.Data.CodonTable import TranslationError


def getargs(argv=None):
    parser = argparse.ArgumentParser(description="Read a gff3 file into a sqlite database using gffutils and produce "\
                                     "fasta formatted protein or transcript sequences from the "\
                                     "corresponding genome")

    parser.add_argument("--out_dir", metavar="output directory",
                        help="Output directory. Will be created if it doesn't already exist.")
    parser.add_argument("--out_name", metavar="output filename",
                        help="Name of the output fasta file.")
    parser.add_argument("--gff3",
                        help="gff3 file containing features to be translated to fasta format. If a db file with the "\
                             "same name is found in the output directory, that db will be used instead.")
    parser.add_argument("--ref",
                        help="Reference genome in fasta format")
    parser.add_argument("--in_mem", required=False, action="store_true",
                        help="Keep the feature database in memory. By default, it will be created on disk in the "\
                             "output directory.")
    parser.add_argument("--transcripts", action='store_true', required=False,
                        help="Output DNA transcript sequences instead of protein sequences.")
    parser.add_argument("--incomplete", action='store_true', required=False,
                        help="Allow incomplete protein sequences (missing start/stop codons). Has no effect on "\
                             "transcripts.")
    return parser.parse_args(argv)


def main(args):
    genome = Fasta(args.ref)
    out_dir = os.path.normpath(args.out_dir)
    gff3_base, ext = os.path.splitext(os.path.basename(args.gff3))
    db_path = os.path.join(out_dir, gff3_base + ".db")

    try:
        os.makedirs(out_dir, exist_ok=True)
    except OSError as err:
        print(err.strerror)
        print("Could not open output directory.")
        sys.exit(1)

    if os.path.isfile(db_path) and not args.in_mem:
        print("Existing database matching ", gff3_base, " found in output directory. Skipping database creation.")
        db=gffutils.FeatureDB(db_path)
    else:
        db = make_db(args.gff3, db_path, args.in_mem)

    if args.transcripts:
        out_type = "transcript"
    else:
        out_type = "protein"
    out_file = os.path.join(out_dir, args.out_name)
    print("Looking for ", out_type, "s in ", db_path, sep="")
    with open(out_file, "w") as out_handle:
        SeqIO.write(protein_recs(db, genome, not args.incomplete, out_type=out_type), out_handle, "fasta")


def make_db(input_gff, db_output_path, in_mem=False):
    if in_mem:
        print("Using in-memory database")
        db = gffutils.create_db(input_gff, ":memory:", verbose=True, keep_order=True)
    else:
        gffutils.create_db(input_gff, db_output_path, verbose=True, keep_order=True)
        db = gffutils.FeatureDB(db_output_path)
    return db


def protein_recs(db, genome, full_cds, out_type):
    """
```

```
    Generate protein records from GFF database.
    """
    if "transcript" in list(db.featuretypes()):
        return translate_feature(db=db, genome_obj=genome,
                                 parent_feature="transcript",
                                 child_name="CDS",
                                 out_type=out_type,
                                 full_cds=full_cds)
    else:
        return translate_feature(db=db, genome_obj=genome,
                                 parent_feature="gene",
                                 child_name="exon",
                                 out_type=out_type,
                                 full_cds=full_cds)

def translate_feature(db, genome_obj, parent_feature, child_name, out_type="protein", full_cds=True):
    """
    A generator returning BioPython Seq objects created from gffutils features.
    Given a parent feature name, a pyfaidx genome object, and the child feature
    name, iterate over all child features in the parent feature, combine their
    sequences, and create a seq object with a description from the gff
    annotations. The child feature can describe a complete CDS with start and
    stop codons, or a just a sequence in one complete reading frame.
    Keyword Arguments:
    db: the database to search
    parent_feature: name of the parent feature type to create the sequence for
    genome_obj: a Fasta genome object created by pyfaidx that contains sequence
                information
    child_name: the feature name of the child features
    full_cds: if True, the translation will only yield a sequence if in-frame
            stop and start codons are in the place, otherwise it just checks
            if the reading frame is complete (i.e. len(seq) % 3 == 0)

    """
    skipped = 0
    for feature in db.features_of_type(parent_feature):
        seq_exons = []
        for exon in db.children(feature, featuretype=child_name, order_by="start"):
            seq_exons.append(exon.sequence(genome_obj, use_strand=False))
        gene_seq = Seq("".join(seq_exons))
        # skip incorrect length CDS
        if feature.strand == "-":
            gene_seq = gene_seq.reverse_complement()

        if out_type == "protein":
            with warnings.catch_warnings():
                # We want to catch incomplete (not 3 frame) coding sequences)
                # if cds=True, this is raised as an error, and will also check if
                # the sequence begins and ends with correct start/stop codons
                # if cds=False, start/stop won't be checked, and partial sequences
                # will only raise a warning, so we need to upgrade it to an error
                # to properly skip it
                warnings.simplefilter("error")
                try:
                    protein_seq = gene_seq.translate(cds=full_cds, to_stop=True)
                    # if "*" in protein_seq: # if there's a stop codon in the sequence
                    # raise TranslationError
                    seq_description = make_seq_header(feature)
                    # translated += 1
                    yield SeqRecord(protein_seq, id=feature.id, description=seq_description)
                except (TranslationError, Warning):
                    skipped += 1

        elif out_type == "transcript":
            seq_description = make_seq_header(feature)
            yield SeqRecord(gene_seq, id=feature.id, description=seq_description)
        else:
            print(out_type, " is not a valid output type.")

    if skipped > 0:
        print("Skipped ", skipped, " features")

def make_seq_header(annotated_feature, key_list=None):
    """
    Given some annotated feature from a gffutils database, and a list of keys
    for the annotation, return an underscore separated string with the key values.
    Keyword Arguments:
        annotated_feature: an annotated feature from the gffutils db
        key_list: a list of keys to append, by default "Besthit", "Symbol", "Description", "Acc"
    Returns:
        underscore separated string of key values
    """

    if key_list is None:
        key_list = ["Besthit", "Symbol", "Description", "Acc"]

    seq_description = []
    for key in key_list:
        try:
            seq_description.append(annotated_feature[key][0])
        except KeyError:
            seq_description.append("")
```

```
    seq_description = "_".join(seq_description)
    return seq_description

if __name__ == '__main__':

    arguments = getargs()

    main(arguments)
```

Listing A.2: BASH script for creating complete protein database from genome sequence and annotions.

```
#! /bin/bash
################################################################################
# Mark Lubberts - 3 March 2019
# Script to create the fasta formatted protein database for LC-MS proteomics
# analysis of fathead minnow, using the SETAC fathead minnow genome and
# corresponding annotations as the input, along with a Uniprot protein database
# and PFAM domain hmms for homology based mapping.
################################################################################

source ~/anaconda3/etc/profile.d/conda.sh
conda activate database_create_env

# Set up input data and output locations
PROJECT_DIR="/home/mark1/projects/quantitative_proteomics/fhm/database_creation"
ZF_CDS="${PROJECT_DIR}/raw_data/gff_originals/ZF_CDS.gff3"
ZF_CDS_OUT="zf_cds-transcripts.fa"
AUGUSTUS_GENES="${PROJECT_DIR}/raw_data/gff_originals/Gene_Predictions.gff3"
AUGUSTUS_OUT="gene_predictions-proteins.fa"
REF_GENOME="${PROJECT_DIR}/raw_data/GCA_000700825.1_FHM_SOAPdenovo_genomic.fna"
REF_PROT_DB="${PROJECT_DIR}/raw_data/uniprot-cyprinidae.fasta"
OUT_DIR="${PROJECT_DIR}/final_protein_db"
TRD_DIR="/home/mark1/projects/quantitative_proteomics/transDecoder"
TRD_OUT="$OUT_DIR/transD_out"
PFAMA_FILE="${PROJECT_DIR}/raw_data/Pfam-A.hmm"
FILTER_SCRIPT="/home/mark1/projects/quantitative_proteomics/fhm/database_creation/filter_out_fasta_sequences.py"
FILTER_LIST="/home/mark1/projects/quantitative_proteomics/fhm/database_creation/bad_proteins.lst"
OUT_DB="$OUT_DIR/final_protein_db.fa"
REDUNDANCY_LVLS=( 1.00 0.95 0.90 0.85 )

# Create initial protein and transcript files
"$PROJECT_DIR"/gff3_to_fasta.py --out_dir "$OUT_DIR" \
                            --out_name "$AUGUSTUS_OUT" \
                            --gff3 "$AUGUSTUS_GENES" \
                            --ref "$REF_GENOME"
"$PROJECT_DIR"/gff3_to_fasta.py --out_dir "$OUT_DIR" \
                            --out_name "$ZF_CDS_OUT" \
                            --gff3 "$ZF_CDS" \
                            --ref "$REF_GENOME" \
                            --transcripts

# Create directory for TransDecoder output files and blastdb
mkdir -p "$TRD_OUT"
CWD="`pwd`"
cd "$OUT_DIR"

echo "  -----Creating Blast database-----  "
makeblastdb -dbtype prot \
          -title "$REF_PROT_DB DB" \
          -out "${TRD_OUT}/ref_prot_db" \
          -in "$REF_PROT_DB"

echo "  -----Running TransDecoder-----  "
# we already have strand specific input
"$TRD_DIR"/TransDecoder.LongOrfs -S \
               -m 80 \
                           -t "${OUT_DIR}/${ZF_CDS_OUT}" \
                           -O "$TRD_OUT"

echo "  -----Running BLAST Search-----  "
blastp -query "${TRD_OUT}/longest_orfs.pep" \
      -db "${TRD_OUT}/ref_prot_db" \
      -max_target_seqs 1 \
      -outfmt 6 \
      -evalue 1e-5 \
      -num_threads 8\
      > "${TRD_OUT}/blastp.outfmt6"

echo "  -----Running HMMScan Search-----  "
tempfile=$(mktemp)
hmmscan --cpu 8 \
       -o "$tempfile" \
       --domtblout "${TRD_OUT}/pfam.domtblout" \
       "$PFAMA_FILE" \
       "${TRD_OUT}/longest_orfs.pep"
rm "$tempfile"

# Run Transdecoder using homology information
echo "  -----TransDecoder Prediction-----  "
```

```
"$TRD_DIR"/TransDecoder.Predict -t "${OUT_DIR}/${ZF_CDS_OUT}" \
                           --retain_pfam_hits "${TRD_OUT}/pfam.domtblout" \
                           --retain_blastp_hits "${TRD_OUT}/blastp.outfmt6" \
                           --single_best_only \
                           -O "$TRD_OUT"

echo "  -----Concatenating Proteins Files----- "
cat "$AUGUSTUS_OUT" "${ZF_CDS_OUT}.transdecoder.pep" > "$OUT_DB"

echo "  -----Filtering Protein Fasta----- "
tempfile=$(mktemp)
cp "$OUT_DB" "$tempfile"
"$FILTER_SCRIPT" "$tempfile" "$FILTER_LIST" > "$OUT_DB"

echo "  -----Creating Non-redundant Files----- "
IN_DB="${OUT_DB##*/}"
for LEVEL in ${REDUNDANCY_LVLS[@]}; do
      DB_NAME="${LEVEL}_${IN_DB}"
      cdhit -c ${LEVEL} -M 102400 -g 1 -d 1000 -i "$OUT_DB" -o "$DB_NAME"
done
cd ${CWD}
```

Listing A.3: Python code for performing in-silico tryptic digests on multiple databases and comparing the results.

```
#! /usr/bin/env python3
##############################################################################
# Mark Lubberts - 3 March 2019
# Program perform in-silico digestion of protein or all proteins in a fasta file
# then plots distribution of peptides. If more than one fasta file is given,
# distributions will be plotted one per graph. Multicore variant.
##############################################################################

import argparse
import sys
import os
from Bio import SeqIO
from tryptic_digest_funcs import get_peptide_list, get_length_freq, digest_protein, get_unique_peptides
from pandas import DataFrame, Series
import matplotlib
# prevent complaints about display when headless
matplotlib.use("Agg")
from matplotlib import pyplot as plt
import numpy as np
from matplotlib_venn import venn2_unweighted, venn3_unweighted
from textwrap import wrap


if __name__ == "__main__":

    parser = argparse.ArgumentParser(description="Check for and score unique peptides in a protein sequence")
    parser.add_argument("sequence", type=str, nargs='+',
                        help="The protein sequence or fasta file")
    parser.add_argument("--misses", type=int, default=1,
                        help="The number of missed cleavages. (default: %(default)s")
    parser.add_argument("--min", type=int, dest="min_length", default=7,
                        help="The minimum peptide length to check. (default: %(default)s)")
    parser.add_argument("--max", type=int, dest="max_length", default=30,
                        help="the maximum peptide length to check. (default: %(default)s)")
    parser.add_argument("--keep_iso", action="store_true", default=False,
                        help="Don't convert isoleucine to leucine. (default: convert)")
    parser.add_argument("--low_mem", action="store_true", default=False,
                        help="Use low memory mode, slower, but will decrease memory usage by ~50%%. Will always use 3 "
                             "cores. (default: off)")
    parser.add_argument("--n_core", type=int, default=4,
                        help="The number of cores to use when processing fasta files. More than 4 cores has minimal "
                             "effect on speed. (default: %(default)s, minimum: 1)")
    parser.add_argument("--fasta", action="store_true",
                        help="Read protein sequence from a fasta file. (default: from string)")
    parser.add_argument("--ambiguous", action="store_true",
                        help="Retain peptides with ambiguous ('X') amino acids (default: discard)")
    parser.add_argument("--out_dir", type=str, default=".",
                        help="Output directory for graphs. Applicable for fasta file input. (default: current "
                        "directory)")
    parser.add_argument("--cutoff", type=int, dest="cutoff_length", nargs="?", const=7,
                        help="Draw a vertical line on output graphs at the peptide length specified. The default length"
                        " is 7 if no integer argument is given, or no line if the flag is not used.")

    args = parser.parse_args()

    if args.fasta:

        if args.out_dir == ".":
            out_dir = os.getcwd()
        else:
            out_dir = os.path.normpath(args.out_dir)
            try:
                os.makedirs(out_dir, exist_ok=True)
```

```
    except OSError as err:
        print(err.strerror)
        print("Specify alternate directory.")
        sys.exit(1)

n_core = args.n_core
if args.low_mem:
    n_core = 4

length_qty_columns = list(range(args.min_length, args.max_length + 1))
uniq_qty_columns = ["uniq" + str(col) for col in length_qty_columns]
results = DataFrame()
peptides_dict = {}

bar_plot_index = []
plot_index = 0
bar_tick_labels = []
bar_peptide_value = []
bar_unique_value = []
print(args.cutoff_length)
possible_peptides_fig, possible_peptides_ax = plt.subplots()
uniq_fraction_fig, unique_fraction_ax = plt.subplots()

for file in args.sequence:
    db_size = len([1 for line in open(file) if line.startswith(">")])
    seq_generator = SeqIO.parse(file, "fasta")
    peptides = get_peptide_list(seq_generator,
                            n_core,
                            args.low_mem,
                            misses=args.misses,
                            l_min=args.min_length,
                            l_max=args.max_length,
                            convert_isoleucine=not args.keep_iso,
                            redundant=True,
                            ambiguous_peptides=args.ambiguous)
    unique_peptides, unique_length_freq = get_unique_peptides(peptides)
    peptides = set(peptides)
    pep_length_freq = get_length_freq(peptides)
    db_name, ext = os.path.splitext(os.path.basename(file))
    db_name = db_name.replace("_", " ")
    plt.figure("num_peptides")
    pep_length_freq.plot(kind='line', label=db_name, ax=possible_peptides_ax)
    plt.figure("frac_unique")
    frac_uniqueness = unique_length_freq/pep_length_freq
    frac_uniqueness.plot(kind="line", label=db_name, ax=unique_fraction_ax)

    results[db_name+"_potential"] = Series({"db_size": db_size}).append(pep_length_freq)
    results[db_name + "_unique"] = Series({"db_size": db_size}).append(unique_length_freq)
    peptides_dict[db_name] = peptides
    bar_plot_index.append(plot_index)
    plot_index += 1
    bar_peptide_value.append(sum(pep_length_freq))
    bar_unique_value.append(sum(unique_length_freq))
    bar_tick_labels.append("\n".join(wrap(db_name, 26)))

# Do we need to log scale data?
use_log_scale = False
if max(bar_peptide_value) > 20*min(bar_peptide_value):
    use_log_scale = True
    print("log scaling graph")

# set up the x-axis limits
x_limits = [args.min_length, args.max_length + 1]

# plt.figure("num_peptides")
if args.cutoff_length is not None:
    possible_peptides_ax.axvline(args.cutoff_length, color='r', linestyle=":")
possible_peptides_ax.set_xlim(x_limits)
if use_log_scale:
    possible_peptides_ax.set_yscale('log')
possible_peptides_ax.set_title("Number of possible tryptic peptides in the database")
possible_peptides_ax.legend()
possible_peptides_ax.set_ylabel("Number of peptides")
possible_peptides_ax.set_xlabel("Peptide length (residues)")
out_file = os.path.join(out_dir, "peptide_length_distribution.png")
possible_peptides_fig.savefig(out_file)
plt.close()

# plt.figure("frac_unique")
unique_fraction_ax.set_xlim(x_limits)
if args.cutoff_length is not None:
    unique_fraction_ax.axvline(args.cutoff_length, color='r', linestyle=':')
unique_fraction_ax.set_title("Fraction of unique tryptic peptides versus peptide length")
unique_fraction_ax.legend()
unique_fraction_ax.set_ylabel("Fraction of Unique peptides")
unique_fraction_ax.set_xlabel("Peptide length (residues)")
out_file = os.path.join(out_dir, "frac_unique.png")
uniq_fraction_fig.savefig(out_file)
plt.close()

# plt.figure("total_peptides")
fig, ax = plt.subplots()
bar_width = 0.35
```

```
    bar_plot_index = np.array(bar_plot_index)
    bar1 = ax.bar(bar_plot_index, bar_peptide_value, bar_width, label="Possible")
    bar2 = ax.bar(bar_plot_index+bar_width, bar_unique_value, bar_width, label="Unique")

    ax.set_xlabel("Database")
    ax.set_ylabel("Number of Peptides")
    ax.set_title("Total tryptic peptides by database")
    ax.set_xticks(bar_plot_index + bar_width/2)
    ax.set_xticklabels(bar_tick_labels)
    if use_log_scale:
       ax.set_yscale('log')
    ax.legend()
    plt.setp(ax.get_xticklabels(), wrap=True)
    fig.tight_layout()
    out_file = os.path.join(out_dir, "total_peptides.png")
    plt.savefig(out_file)
    plt.close()
    # if len(peptides_dict) < 4:
    # if len(peptides_dict) == 2:
    # venn2_unweighted([peptides_dict[db] for db in peptides_dict.keys()],
    # set_labels=[db for db in peptides_dict.keys()])
    # elif len(peptides_dict) == 3:
    # venn3_unweighted([peptides_dict[db] for db in peptides_dict.keys()],
    # set_labels=[db for db in peptides_dict.keys()])
    # plt.title("Tryptic Peptide Overlap")
    # plt.savefig("overlap_venn.png")
    # plt.close()
    # else:
    # print("skipping drawing venn, too many items")

    results.to_csv("tryptic_digest_results.csv", index_label="db_name")

else:
    # sequence = args.sequence[0]
    for seq in args.sequence:
        peptides = list(digest_protein(seq, args.misses, args.min_length, args.max_length, args.keep_iso))
        pep_length_freq = get_length_freq(peptides)
        print(pep_length_freq)
```

Listing A.4: Python code for performing in-silico tryptic digests.

```
#! /usr/bin/env python3
################################################################################
# Mark Lubberts - 3 March 2019
# functions for in-silico tryptic digest of a protein or protein fasta files
################################################################################

from pandas import value_counts, unique
import multiprocessing as mp
from functools import partial


def digest_protein(protein_sequence, misses=1, l_min=7, l_max=30,
                   convert_isoleucine=True, as_list=False, ambiguous_peptides=False):
    """Perform in-silico digestion of a protein sequence with a specified enzyme

    Loop through a string specifying a protein sequence and return a set of strings of peptides that would be produced
    by digesting the protein with trypsin with up to the specified number of misses cleavages

    Params:
    prot_sequence: string: amino acid sequence of the protein
    misses: integer: with the number of allowed missed cleavages by the enzyme
    min_length: integer: of the minimum peptide length to include in the list, default value is 7
    max_length: integer: of the maximum peptide length to include in the list, default is 30
    convert_isoleucine: bool: Convert isoleucine to leucine before digesting
    as_list: bool: Output the peptides as a list instead of a set. This will keep redundant peptides in the output.
    ambiguous_peptides: Keep peptides with ambiguous ('X', by IUPAC standards) amino acids

    Returns:
    peptides: set or list of strings of peptides produced, with complete sequence if no cut sites are found
    """

    peptides = []
    enzyme_sites = {"K", "R"}
    # split_sites is a list of positions the protein sequences will be split at. If there are no cut sites, the list
    # should be [0,len(seq()], since python is zero-indexed and specifies intervals half open (e.g. a list length five
    # has indexes 0,1,2,3,4)

    split_sites = [0]

    if convert_isoleucine:
        protein_sequence = protein_sequence.replace("I", "L")

    # find all the enzyme cut sites in the protein sequence. If the last peptide is K/R, the 'cut site' is skipped,
    # because it's outside of the protein length
    for i in range(len(protein_sequence)):
        if protein_sequence[i] is "*":
            split_sites.extend([i-1, i])
        elif protein_sequence[i - 1] in enzyme_sites and protein_sequence[i] != 'P':
            split_sites.append(i)
```

```
    split_sites.append(len(protein_sequence))

    # loop to create position pairs from split_sites that define new peptides and grab them from the protein sequence
    for i in range(misses+1):
        # Loop through number of missed cleavages. For 1 missed cleavages, the protein would be cut at index 0 and 0+1,
        # and 0 and 0+2
        i += 1
        if i < len(split_sites): # Can't have more missed cut sites than cut sites
            for j in range(len(split_sites)-i):
                # j and j+i are the index in split_sites of the positions we want to cut, so split_sites[j] and
                # split_sites[j+i] give the index position in protein_sequence of the new peptide
                start = split_sites[j]
                stop = split_sites[j+i]
                if stop-start in range(l_min, l_max+1):
                    # ignore peptides outside of min/max length
                    new_peptide = protein_sequence[start:stop]
                    if ("X" not in new_peptide or ambiguous_peptides) and ("*" not in new_peptide):
                        # If the peptide sequence contains an 'X' we don't want it, unless specifically asked.
                        peptides.append(protein_sequence[start:stop])
        else: # if the we have more missed cut sites that cut sites, stop loop
            break

    if as_list is False:
        peptides = set(peptides)

    return peptides


def seq_generator(seq_obj_iter):
    """Convert the BioPython Seq object to a string in a generator"""
    for record in seq_obj_iter:
        yield(str(record.seq))


def chunkify_list(the_list, k):
    """divide the_list into a list of k sublists. From MaPePeR on stackoverflow"""
    from math import log10
    n = len(the_list)
    k = k*int(max((1, log10(n) - 2))) # Try splitting large lists into more chunks to struct size error
    chunkified_list = [the_list[i*(n//k) + min(i, n % k):(i+1) * (n//k)+min(i+1, n % k)] for i in range(k)]

    return chunkified_list


def list_worker(protein_list, as_list=False, **kwargs):
    """Wrapper function for digest_protein to produce a peptide set from a protein list"""
    if as_list is True:
        peptide_set = list()
        for prot in protein_list:
            peptide_set.append(digest_protein(prot, as_list=as_list, **kwargs))

    else:
        peptide_set = set()
        for prot in protein_list:
            peptide_set.update(digest_protein(prot, as_list=as_list, **kwargs))
        peptide_set = list(peptide_set)

    return peptide_set


def get_peptide_list(seqobject_iterator, n_core=4, low_mem=False, redundant=False, **kwargs):
    """Perform in-silico digest of all protein SeqObject in an iterable, and returns a list of unique peptides

    Given an iterable or generator of Biopython SeqObjects, will loop through all items in the iterable and find their
    tryptic peptides, using one of three methods. If n_core == 1, will loop through and add peptides to the set one
    protein at a time, checking for redundant peptides each time. If n_core > 1 AND low_mem == True, will convert
    SeqObjects using a generator, and pass them to n_core-1 process for digest. Results from the worker processes will
    be asynchronously and immediately added to the peptide set. This keeps memory usage ~equal to the final peptide list
    size, but since one process is handling the generator and peptide set updating at the same time, does not scale well
    with n_core. If n_core > 1 and low_mem !=True, the iterable will be converted to a string list, split n_core ways,
    sent to n_core processes, then combined and deduplicated afterwards. Note that each worker process will still remove
    redundant peptides, as this has significant effects on memory usage and not much on speed, as the very short, highly
    repetitive proteins can be removed on a per core basis, rather than afterwards. This scales better with n_cores, but
    increases memory usage on large protein sets.

    Params:
    seq_obj_iter = iterable of Biopython SeqObjects
    n_core = number of processes to use
    low_mem = work in (slightly) slower low memory mode
    redundant: Bool: Keep redundant peptides
    **kwargs = arguments to pass to digest_protein

    Returns:
    all_peptides = list of unique peptides
    """

    if redundant is True:
        all_peptides = list()
    else:
        all_peptides = set()

    if n_core == 1:
        for prot in seq_generator(seqobject_iterator):
            if redundant is True:
```

```
            all_peptides += digest_protein(prot, as_list=redundant, **kwargs)
        else:
            all_peptides.update(digest_protein(prot, as_list=redundant, **kwargs))

    elif low_mem is True:
        # use a generator to convert seqObjects to strings, pass to digest_protein then combine resulting peptides sets
        # The imap function is asynchronous, so the main process will start updating as soon as it
        # gets results from the first protein. This means the memory used is ~equal to the final
        # size of the peptides set, however, the generator and set update are performed on the same
        # process, limiting multiprocessing to ~3 cores.
        pool = mp.Pool(max(n_core-1, 1))
        worker_partial = partial(digest_protein, as_list=redundant, **kwargs)
        for result in pool.imap_unordered(worker_partial, seq_generator(seqobject_iterator), 7):
            if redundant is True:
                all_peptides += result
            else:
                all_peptides |= result
        pool.close()
        pool.join()

    else:
        all_peptides = []
        # break the list in n_core equal lengths. This is *much* faster than passing the whole list
        # to pool.map() and letting it hand things out.
        prot_list = chunkify_list([str(prot.seq) for prot in seqobject_iterator], n_core)
        pool = mp.Pool(n_core)
        # Create a partial from the digest_protein wrapper function to handle the list chunk
        worker_partial = partial(list_worker, as_list=redundant, **kwargs)
        results = pool.map(worker_partial, prot_list)
        # append to list and then remove redundant with pandas unique()
        for result in results: # This is one list per individual 'chunk' of the protein list
            for peptide_list in result:
                all_peptides += peptide_list
        if redundant is not True:
            all_peptides = unique(all_peptides)
        pool.close()
        pool.join()

    all_peptides = list(all_peptides)

    return all_peptides


def get_length_freq(string_list):
    """function to convert a list of strings (ie. peptides) to a Series with frequency of each observed string length"""
    str_length_frequency = value_counts([len(string) for string in string_list], sort=False, dropna=False).sort_index()
    return str_length_frequency


def seq_parse_filter(seqobject_iterator, filter_string, keep=False):
    """Parse a fasta file and filter by sequence description/name"""
    for seq in seqobject_iterator:
        if filter_string in seq.description and keep:
            yield(seq)
        elif filter_string not in seq.description and not keep:
            yield(seq)


def get_unique_peptides(redundant_peptide_list):
    """Find all peptide sequences with only one occurrence in the list.

    :param redundant_peptide_list: a list of peptide strings with redundancy
    :return: a list single occurrence strings and a Pandas series of frequency of string lengths.
    """
    unique_peptides = value_counts(redundant_peptide_list)
    unique_peptides = unique_peptides[unique_peptides == 1]
    unique_peptides = list(unique_peptides.index)
    unique_length_frequencies = get_length_freq(unique_peptides)
    return unique_peptides, unique_length_frequencies


if __name__ == "__main__":
    print("this is just functions to use")
```

A.2 Additional Figures

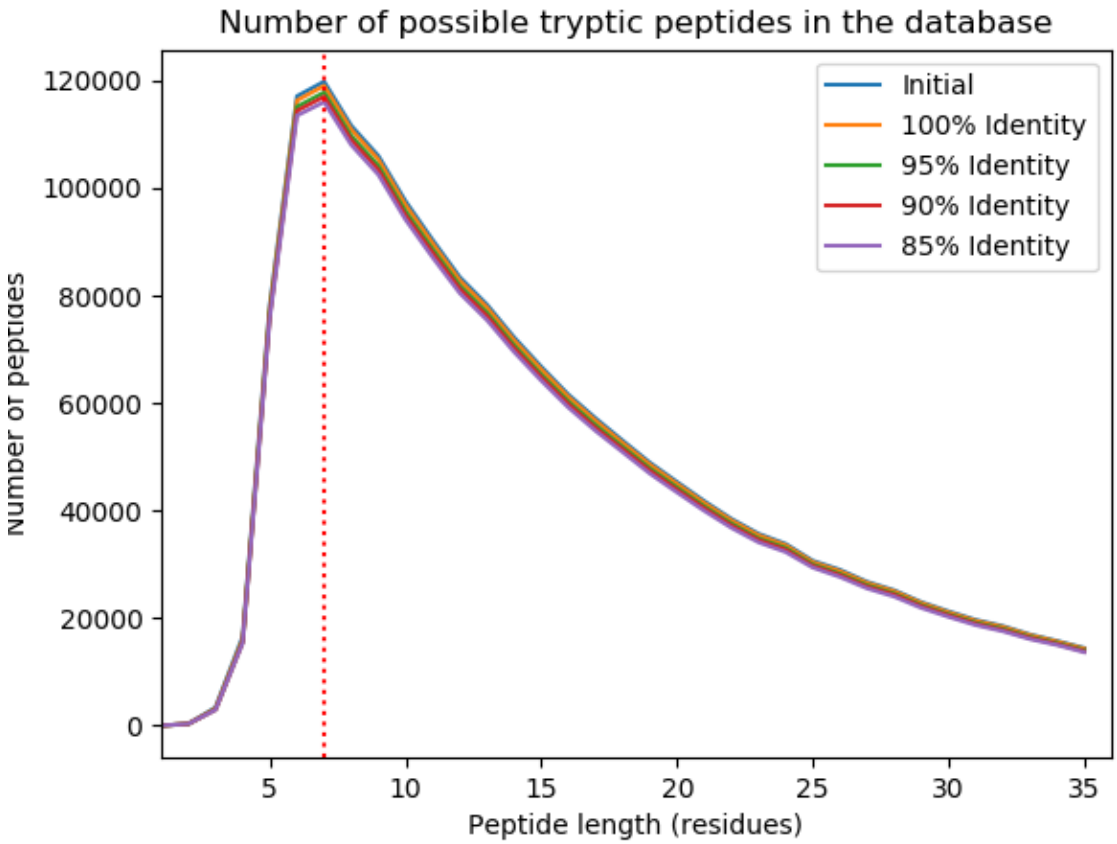


Figure A.1: Distribution of peptide lengths in the FHMP at various levels of clustering.

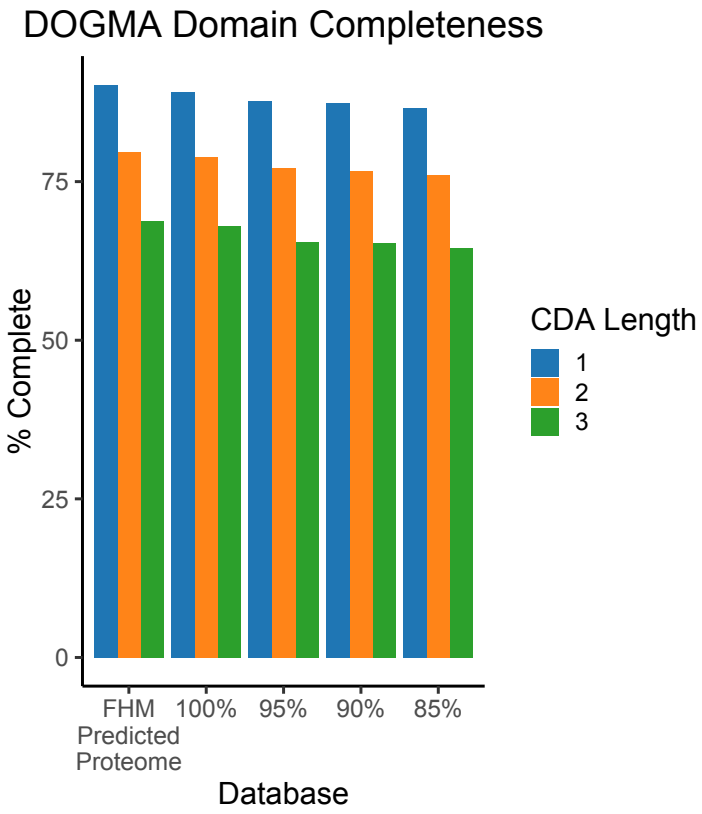


Figure A.2: Distribution of conserved domain arrangements (CDAs) in the FHMP at various levels of clustering.

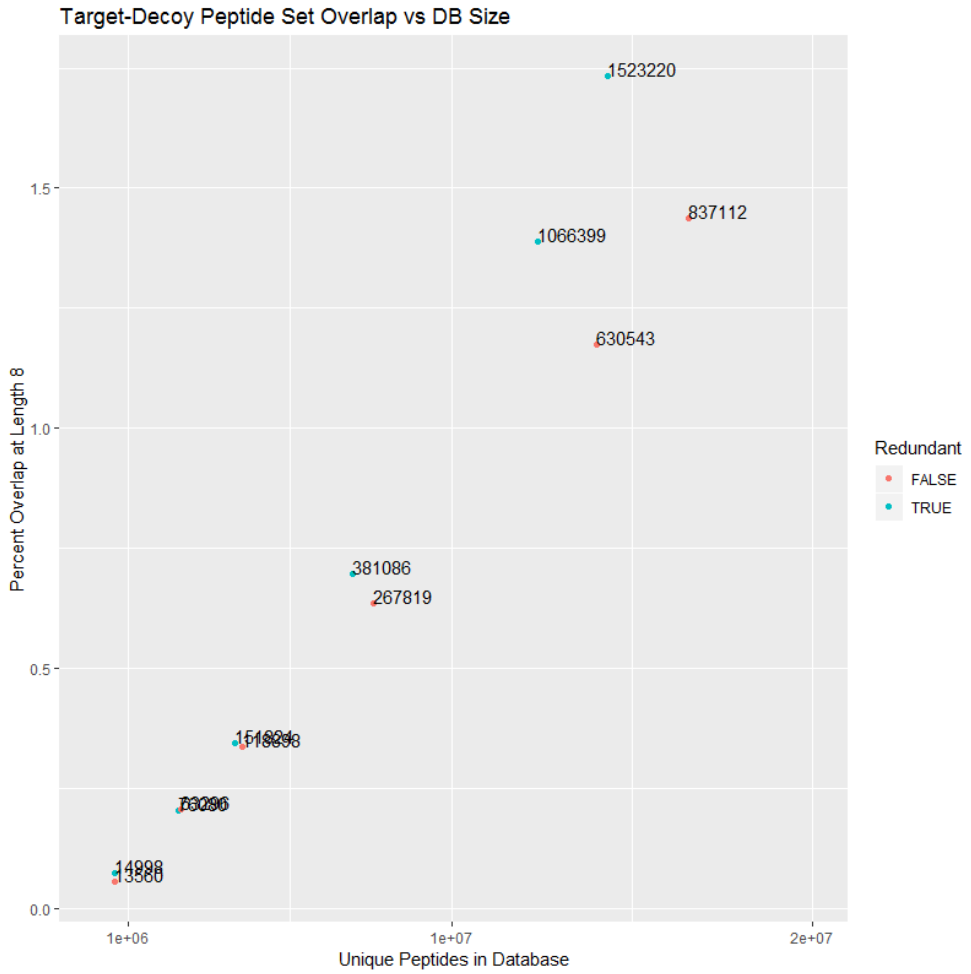


Figure A.3: Percent overlap between the target and decoy database versus db size in tryptic peptides. The number of proteins in the database is indicated next to each point.

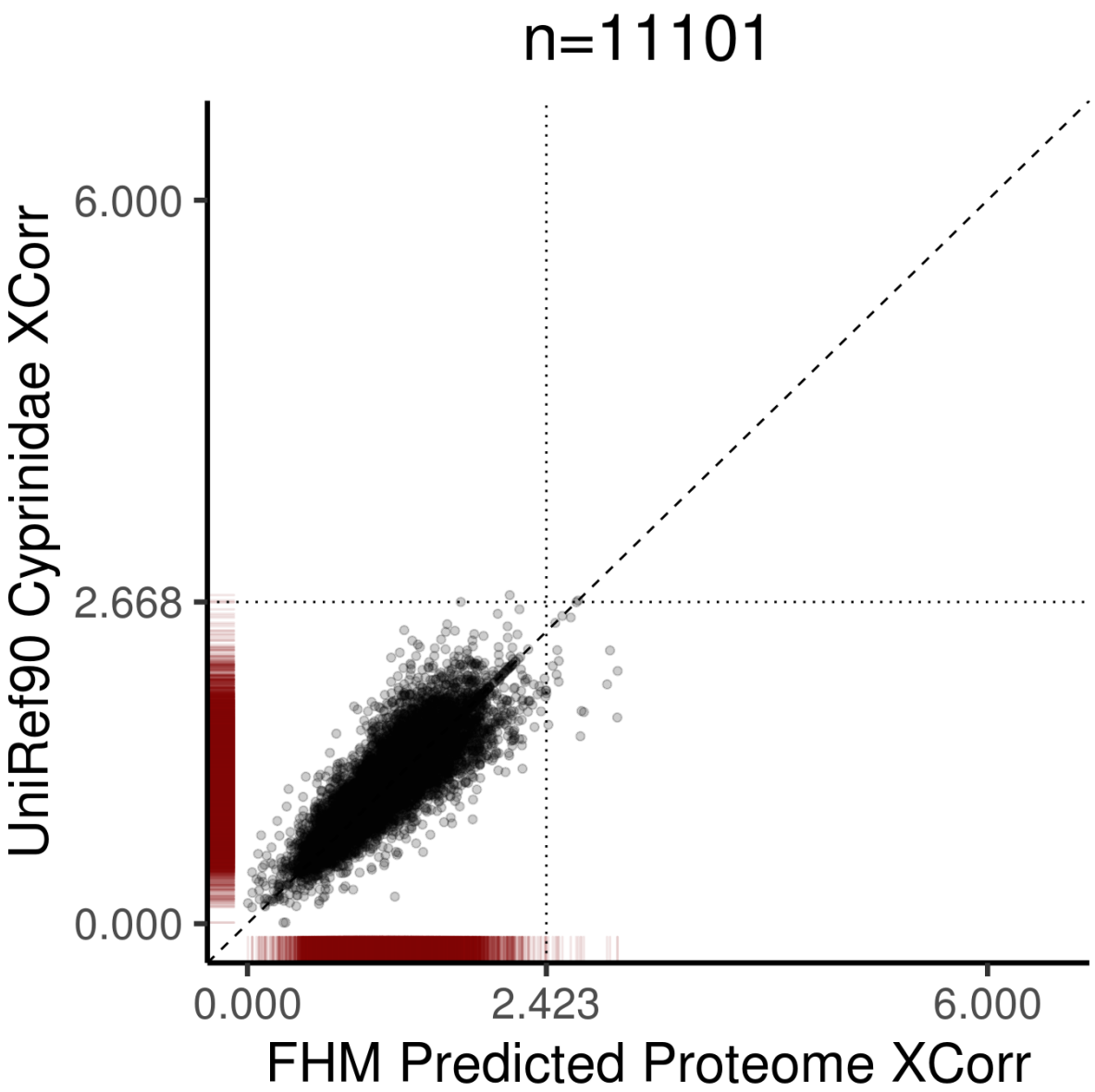


Figure A.4: Comparison of PSM scores between the FHMP and U90CYP databases for PSMs which matched to decoy peptides in both databases. The total number of PSMs is indicated above the plot.

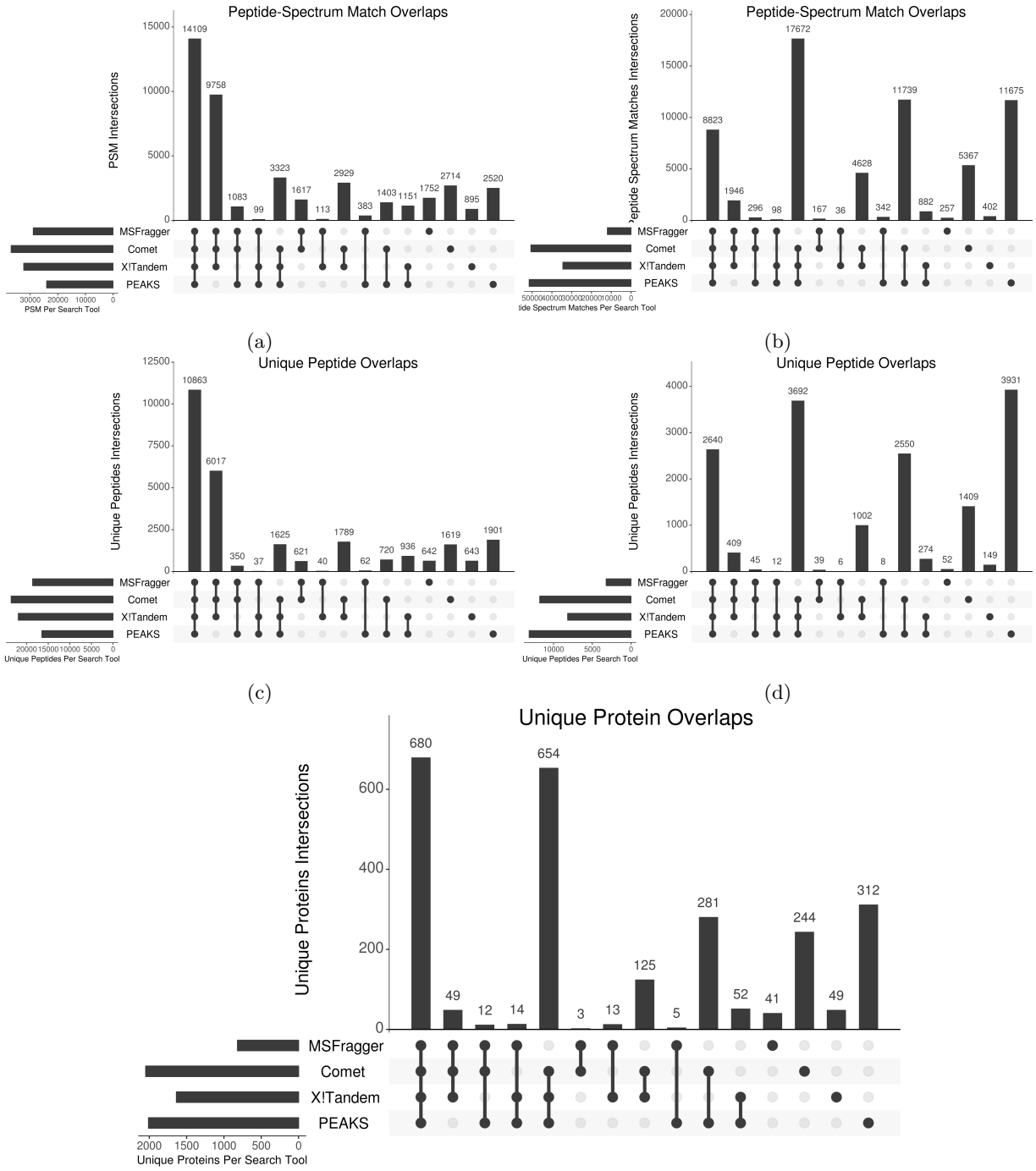


Figure A.5: Overlap in peptide-spectrum matches (PSMs) and peptide and protein identifications between search engines in the trout and human datasets. a) and b) PSM overlap in human and trout, respectively. c) andd) peptide identification overlap for human and trout. e) protein identification overlap in the trout dataset.

A.3 Database comparison

Table A.1: Decoy and Target PSM scoring values. Ion Count refers to the number of ions per mass spectra matched to the theoretical spectra

Database	Source	Mean xCorr	Mean ion count	Mean mass difference
Fathead Minnow Predicted Proteome (FHMP)	Target	1.65	12.43	0.307
	Decoy	1.14	10.07	0.517
UniRef90 Cyprinidae Proteome (U90CYP)	Target	1.60	12.12	0.322
	Decoy	1.17	10.17	0.490

A.4 Search Tool Comparison Parameters

Table A.2: Parameters for the Search Engine Comparison.

Tool	Parameter	Sample		
		Human	Trout	FHM
All	Fixed Mods	TMT, Carbo		
	Variable Mods	Oxidation of Methionine		
Comet	Peptide Mass Tolerance	10 ppm	0.1 Da	15 ppm
	Fragment Bin Tolerance	1.0005	0.02	1.0005
	Fragment Bin Offset	0.4	0.0	0.4
MSFragger	Precursor Mass Tolerance		±500	
	Fragment Mass Tolerance	0.6 Da	0.6 Da	0.6 Da
X!Tandem	Parent Monoisotopic Mass Error	10 ppm	0.1 Da	15 ppm
	Fragment Monoisotopic Mass Error	0.6 Da	0.2 Da	0.1 Da
PEAKS 8.5	Parent Mass Error Tolerance	10 ppm	0.1 Da	10 ppm
	Fragment Mass Error Tolerance	0.6 Da	0.2 Da	0.1 Da

B.1 Water Quality and Contaminants Data

Table B.1: Average concentrations of select contaminants in the three groups. MDL, minimum detection limit. Concentrations in undiluted effluent from the Pine Creek WWTP are provided for comparison.

Compound	MDL	Upstream	Outflow	Downstream	Pine Creek*
Neutral PPCPs					
Cotinine	0.005	ND	0.034	0.005	0.034
Caffeine	0.010	0.010	0.127	0.022	0.022
Trimethoprim	0.004	ND	0.227	0.014	0.227
Carbamazepine	0.002	ND	0.393	0.026	0.482
Acidic PPCPs					
Naproxen	0.004	ND	0.086	0.007	0.029
Diclofenac	0.004	ND	0.480	0.051	0.507
Ibuprofen	0.007	ND	0.060	ND	ND
Gemfibrozil	0.007	ND	0.017	ND	ND
Antidepressants					
Fluoxetine	n/a	ND	0.002	ND	0.005
Norfluoxetine	n/a	ND	0.014	ND	0.013
Venlafaxine	n/a	ND	0.472	0.023	0.538
O-Desmethylvenlafaxine	n/a	ND	1.514	0.082	1.735

Table B.2: Water quality data at the five sites. Superscripts indicate statistically significant groups after Bonferroni's correction for multiple hypothesis testing.

Site	Temperature (°C)	Dissolved Oxygen (mg/L)	Conductivity (µS/cm)	pH	Turbidity (NTU)
BEAR	12.0 ± 0.3^{a}	9.63 ± 0.51^{b}	298.3 ± 1.5^{a}	8.25 ± 0.03^{b}	4.28 ± 1.37
CUSH	12.0 ± 0.5^{a}	10.45 ± 0.14^{b}	325.8 ± 2.1^{b}	8.18 ± 0.08^{b}	3.27 ± 0.36
GLEN	18.4 ± 0.2^{b}	6.06 ± 0.25^{a}	928.5 ± 3.6^{e}	7.23 ± 0.08^{a}	4.32 ± 0.53
H22X	12.5 ± 0.4^{a}	11.04 ± 0.56^{b}	376.0 ± 2.8^{c}	8.19 ± 0.09^{b}	5.06 ± 0.60
UPHI	12.3 ± 0.3^{a}	9.91 ± 0.53^{b}	397.4 ± 2.3^{d}	8.00 ± 0.08^{b}	4.33 ± 1.14

Table B.3: Length(mm), weight(g), and condition factor of fathead minnows used for proteome analysis after caging at the listed sites in the Bow River. Values are listed as mean ± SD. Letters within columns indicate statistically significant differences between sites after Bonferroni correction for multiple tests.

Site	Sex Ratio (M/F)	Length (mm)	Weight (g)	Condition Factor
BEAR	2/4	48 ± 8	1.22 ± 0.71	1.049 ± 0.035^{a}
CUSH	3/3	48 ± 8	1.55 ± 0.61	1.247 ± 0.200^{b}
GLEN	3/3	51 ± 6	1.70 ± 0.64	1.198 ± 0.083^{b}
H22X	3/3	46 ± 6	1.10 ± 0.47	1.104 ± 0.085^{ab}
UPHI	1/5	45 ± 9	0.99 ± 0.54	1.079 ± 0.189^{ab}

Table B.4: Characteristics and morphometrics of individal fathead minnows used for the proteomics study. Fish ID is the same as used in Lazaro-Côté *et al.* [12], run indicates which of the three mass spectrometry runs the sample was run in, and TMT label indicates the TMT reporter tag used for that sample.

Fish ID	Run	TMT Label	Site	Cage	Sex	Fish mass (g)	Liver tissue mass (mg)	Fork length (mm)	Condition Factor
49	A	126	BEAR	1	M	1.03	10.8	46	1.058
51	B	130C	BEAR	1	F	0.98	8.7	45	1.075
52	C	129C	BEAR	1	M	0.96	8.1	46	0.986
42	A	131	BEAR	2	F	0.82	7.5	43	1.031
41	B	127C	BEAR	2	M	2.65	7.9	63	1.060
48	C	127C	BEAR	2	M	0.86	9.2	43	1.082
33	A	130N	CUSH	1	M	2.06	7.2	55	1.238
37	B	129N	CUSH	1	M	1.87	8.7	51	1.410
38	C	128C	CUSH	1	F	0.43	8.0	34	1.094
26	A	127C	CUSH	2	F	1.39	12.5	44	1.632
28	B	128N	CUSH	2	F	1.55	7.9	47	1.493
32	C	126	CUSH	2	M	2.02	10.4	55	1.214
59	A	129C	GLEN	1	F	1.25	7.8	48	1.130
61	B	131	GLEN	1	F	1.59	11.8	50	1.272
62	C	128N	GLEN	1	M	2.18	12.4	55	1.310
71	A	130C	GLEN	2	M	2.60	13.5	60	1.204
70	B	126	GLEN	2	M	1.75	9.0	53	1.175
68	C	130C	GLEN	2	F	0.81	7.9	42	1.093
20	A	128C	H22X	1	F	0.67	12.9	38	1.221
21	B	129C	H22X	1	M	0.96	7.0	45	1.053
23	C	131	H22X	1	F	1.93	11.1	56	1.099
14	A	129N	H22X	2	M	0.67	11.6	41	0.972
16	B	127N	H22X	2	F	1.25	7.9	48	1.130
19	C	130N	H22X	2	M	1.12	7.5	46	1.151
2	A	127N	UPHI	1	M	0.73	12.1	42	0.985
4	B	128C	UPHI	1	M	2.06	10.3	61	0.908
5	C	129N	UPHI	1	M	0.65	10.7	36	1.393
10	A	128N	UPHI	2	F	0.98	13.4	45	1.075
9	B	130N	UPHI	2	M	0.63	8.8	41	0.914
8	C	127N	UPHI	2	M	0.89	11.6	42	1.201

B.2 Code Used for Analysis

Listing B.1: R code for analysis of differentially expressed proteins in the mass spectrometry data.

```
#!/usr/bin/env Rscript
# Analysis pipeline
library(tidyr)
library(reticulate)
library(FGNet)
#library(Normalyzer)
library(optparse)
library(stringr)
library(data.table)
library(ggplot2)
library(ggrepel)
library(Cairo)
source("norm_and_collate_protein_quant.R")
source("go_annotations.R")
source("additional_functions.R")
use_condaenv("database_create_env", required=TRUE)

### Create output directory
save_data <- TRUE
if(save_data){
   out_dir <- make_dir("final_thesis_db_90_outlier_drop_final")
   img_dir <- make_dir("img", out_dir)
}else{
   out_dir <- NULL
   img_dir <- NULL
}

#### Load Raw Data ####
# file containing mapping of sample to label in each run
design_table <- fread("input_data/experiment_layout.txt")
# peptide file exported from TPP with the experiment label added and matching runs in the design_table file
pep_file <- "/mnt/storage/tpp_data/fhm/final_thesis_db/final_thesis_run/all.interact.ipro.pep.xls"

# protein IDs from the combined runs
prot_file <- "/mnt/storage/tpp_data/fhm/final_thesis_db/final_thesis_run/interact.ipro.prot.tsv"

# annotation database locations
augustus_db <- "../database_creation/final_protein_db/Gene_Predictions.db"
zf_db <- "../database_creation/final_protein_db/ZF_CDS.db"

#pep_file <- "/mnt/storage/tpp_data/fhm/new_db/augustus_and_transdecoder_90/all.interact.ipro.pep.xls"
#prot_file <- "/mnt/storage/tpp_data/fhm/new_db/augustus_and_transdecoder_90/interact.ipro.prot.tsv"
#augustus_db <- "../database_creation/final_protein_db_bad/Gene_Predictions.db"
#zf_db <- "../database_creation/final_protein_db_bad/ZF_CDS.db"

housekeeping_list <- fread("input_data/housekeeping_list.txt")

# map peptides to good protein IDs
source_python("prot_id_parser.py")
cleaned_data <- clean_prot_list(prot_file,
                          pep_file,
                          augustus_db,
                          zf_db, contaminant_strings=list("CRAP"))
# read the remapped peptides and filter out low quality peptides
quant_data_raw <- data.table(cleaned_data[[1]])

# PSMs from the same scan are duplicates detected by comet/X!Tandem
quant_data_raw <- unique(quant_data_raw, by=c("experiment_label", "start_scan", "peptide"))

prot_full_data <- data.table(cleaned_data[[2]])
quant_data_raw <- quant_data_raw[iprobability > 0.9,]

# Make MDS and bar plot of spectra intensities across all runs before normalization
peptide_intensity <- renamed_intensity_from_raw(quant_data_raw, design_table, keep_unlabelled = TRUE)
raw_mds_plots <- make_mds_plots(peptide_intensity)
output_plot(raw_mds_plots[[1]], img_dir, "raw_spectra_mds.png")

spectra_intensity <- renamed_intensity_from_raw(quant_data_raw, design_table, keep_unlabelled = FALSE, sum_peptide =
     FALSE)
spectrum_boxplot <- plot_logfc_boxplot(spectra_intensity)
output_plot(spectrum_boxplot, img_dir,"raw_spectra_boxplot.png", width=12, height = 8)

#### run normalization and collate protein quantities ####
quant_a <- run.quant(quant_data_raw[experiment_label == "McConkey-TMT-A",],
                 design_table,
                 minimum_reporters = 2,
                 housekeeper_ids = housekeeping_list$zf_symbol,
                 img_dir = img_dir,
                 save_plots = save_data)
quant_b <- run.quant(quant_data_raw[experiment_label == "McConkey-TMT-B",],
                 design_table,
                 minimum_reporters = 2,
                 housekeeper_ids = housekeeping_list$zf_symbol,
                 img_dir = img_dir,
                 save_plots = save_data)
```

```
quant_c <- run.quant(quant_data_raw[experiment_label == "McConkey-TMT-C",],
                 design_table,
                 minimum_reporters = 2,
                 drop_channels = c("H22X.C.2.M"),
                 housekeeper_ids = housekeeping_list$zf_symbol,
                 img_dir = img_dir,
                 save_plots = save_data)

# merge dataset by protein
quant_data <- merge(quant_a, quant_b, by=1, all.x = TRUE, all.y = TRUE)
quant_data <- merge(quant_data, quant_c, by=1, all.x = TRUE, all.y = TRUE)

# Remove NAs and unnecessary columns
quant_data <- quant_data[, !(names(quant_data) %in% colnameslike(quant_data, "n."))]
quant_data <- na.omit(quant_data)

# Make some plots
quant_data_melt <- melt(data.table(quant_data), id="protein", variable.name = "sample_id", value.name = "log2FC")
quant_data_melt[, c("site", "run", "cage", "sex") := tstrsplit(sample_id, split="\\.")]
run_labels <- c("Run A", "Run B", "Run C")
names(run_labels) <- c("A", "B", "C")
norm_fc_plot <- ggplot(quant_data_melt, aes(x=sample_id, y=log2FC, fill=site, linetype=sex)) +
   geom_boxplot(notch=TRUE) +
   theme_classic() +
   facet_grid(.~run, scales = "free_x", labeller = labeller(run = run_labels)) +
   theme(text = element_text(size=20), axis.text.x = element_text(angle = 90, vjust=0.5, hjust=0)) +
   scale_fill_brewer(type = "qual", palette = "Set1") +
   labs(fill="Caging Site", x="Sample", y="Normalized Log2 Fold-Change")

output_plot(norm_fc_plot, img_dir, "normalized_logFC.png", height=8, width=12)

mds_plots <- make_mds_plots(quant_data)

output_plot(mds_plots[[1]], img_dir, "normalized_mds_plot.png")

output_plot(mds_plots[[2]], img_dir, "mds_variance_by_coordinate.png")

if (save_data){
   plot_file <- file.path(img_dir, "heatmap.png")
   png(type = "cairo", filename = plot_file, width=10, height=10, units = "in", res = 300)
   coolmap(as.matrix(as.data.table(quant_data), rownames = "protein"),
          cluster.by = "de pattern",
          linkage.col = "complete",
          margins = c(6,2),
          labRow=NA)
   dev.off()
}else{
   coolmap(as.matrix(as.data.table(quant_data), rownames = "protein"),
          cluster.by = "de pattern",
          linkage.col = "complete",
          margins = c(6,2),
          labRow=NA)
}

protein_boxplot <- plot_logfc_boxplot(quant_data, sort_columns = TRUE)
output_plot(protein_boxplot, img_dir,"normalized_protein_boxplot.png", height=8, width=12)

#### Set up limma experiment design ####
#quant_data <- quant_data[,c(1:28,30:31)]
# sample group of each column
# tr <- as.factor(rep(c("BEAR","BEAR","CUSH","CUSH","GLEN","GLEN","H22X","H22X","UPHI","UPHI"), 3))
tr <- as.factor(c(rep(c("BEAR","BEAR","CUSH","CUSH","GLEN","GLEN","H22X","H22X","UPHI","UPHI"), 2),
             "BEAR","BEAR","CUSH","CUSH","GLEN","GLEN","H22X","UPHI","UPHI"))
# sex of each biological sample in the run
sex <- as.factor(c("M","F","M","F","F","M","F","M","M","F",
                "F","M","M","F","F","M","M","F","M","M",
                "M","M","F","M","M","F","F","F","M"))
exp <- as.factor(c(rep(1,10), rep(2,10), rep(3,9)))
design <- model.matrix(~0+tr+sex)

print(min(svd(design)$d))

# do Limma stuff
corfit <- duplicateCorrelation(quant_data[2:30],design, block=exp)
fit <- lmFit(quant_data[2:30], design, block=exp, correlation = corfit$consensus)
fit.eb <- eBayes(fit)
# Compare Glen vs upstream and other two vs upstream
contrastmatrix <- makeContrasts(Outflow_vs_Upstream=trGLEN-(trBEAR+trCUSH)/2,
                          Downstream_vs_Upstream=(trH22X+trUPHI)/2-(trBEAR+trCUSH)/2,
                          Outflow_vs_Downstream=trGLEN-(trH22X+trUPHI)/2,
                          levels=design)
fit2 <- contrasts.fit(fit, contrastmatrix)
fit2.eb <- eBayes(fit2)
sig_changes <- decideTests(fit2.eb, p.value = 0.05, method="global", adjust.method="BH")
summary(sig_changes)

# Create Venn Diagram of the overlap between analysis groups
if(save_data){
   png(filename = file.path(img_dir, "de_venn.png"), width=8, height=8, units="in", res=300)
   vennDiagram(sig_changes, circle.col = c("red", "yellow"))
   title(main= "Number of differentially expressed proteins")
   dev.off()
}else{
```

```
   vennDiagram(sig_changes, circle.col = c("red", "yellow"))
   title(main= "Number of differentially expressed proteins")
}

# get significantly differentially expressed proteins
sig_proteins <- data.frame(cbind(as.character(quant_data[,1]), sig_changes[,1:3]), stringsAsFactors = FALSE)
colnames(sig_proteins) <- c("protein", "Out_vs_Up","Down_vs_Up", "Out_vs_Down")

#### Get Protein IDs and symbols ####
# Grab all ID'd proteins from TPP dataset
prot_id_list <- prot_full_data[,c("protein","protein description")]
prot_id_list$protein <- lapply(prot_id_list$protein, trimws, which = "right")

# get list with description
sig_proteins <- merge(sig_proteins, prot_id_list, by=1, all.x=TRUE)
sig_proteins$ensembl_ID <- str_extract(sig_proteins$`protein description`,"ENSDAR[TP][0-9.]+")
sig_proteins$zfin_ID <- str_extract(sig_proteins$`protein description`, "ZDB-GENE-[0-9]+-[0-9]+")
sig_proteins$symbol <- lapply(sig_proteins$protein,
                        function(x) unlist(str_split(str_extract(x, "ENSDAR[TP][0-9.]+_+.*_"), "_+"))[2])

# separate out DE and background proteins
#sig_proteins <- sig_proteins[sig_proteins$Glen != 0 | sig_proteins$Downstream != 0,]

# get p-value and fold change for differentially expressed proteins
de_prots <- data.frame(cbind(fit2.eb$p.value, fit2.eb$coefficients, sig_changes))
#colnames(de_prots) <- c("p_value_out","p_value_down","logfc_out","logfc_down","out_sig","down_sig")
colnames(de_prots) <- c("p_value_out","p_value_down", "p_value_btwn","logfc_glen","logfc_down", "logfc_btwn",
     "glen_sig","down_sig", "btwn_sig")
de_prots$protein <- quant_data$protein
de_prots <- merge(de_prots, sig_proteins[,c("protein","ensembl_ID","zfin_ID")], by.x = "protein", by.y = "protein",
     all.x=TRUE)
de_prots <- data.table(de_prots)

#### Get consistent ensembl IDs for easier analysis and functional enrichment ####
de_prots[ensembl_ID %like% "ENSDAR", ensembl_ID := str_extract(ensembl_ID, "ENSDAR[PT][0-9]+")]


# Grab protein descriptions from ensembl
desired_attributes <-c("external_gene_name",
                   "ensembl_transcript_id",
                   "ensembl_peptide_id",
                   "ensembl_gene_id",
                   "description",
                   "name_1006",
                   "zfin_id_id")

ensembl_data <- rbindlist(list(get_biomart_attributes(de_prots[ensembl_ID %like% "ENSDARP",]$ensembl_ID,
                                       desired_attributes,
                                       "ensembl_peptide_id"),
                         get_biomart_attributes(de_prots[ensembl_ID %like% "ENSDART",]$ensembl_ID,
                                       desired_attributes,
                                       "ensembl_transcript_id")))

de_prots <- ensembl_data[de_prots, on=.(ensembl_ID), nomatch = NA]

de_prots[external_gene_name == "",]$external_gene_name <- NA

# Use an older archive to fill in proteins/genes with missing data

desired_attributes <-c("external_gene_name",
                   "ensembl_transcript_id",
                   "ensembl_peptide_id",
                   "ensembl_gene_id",
                   "description",
                   "name_1006",
                   "zfin_id")

missing_transcripts <- get_biomart_attributes(de_prots[is.na(external_gene_name) & ensembl_ID %like%
     "ENSDART"]$ensembl_ID,
                                      desired_attributes,
                                      "ensembl_transcript_id",
                                      host = "Jul2016.archive.ensembl.org")
colnames(missing_transcripts)[colnames(missing_transcripts) %like% "zfin"] <- "zfin_id_id"
missing_proteins <- get_biomart_attributes(de_prots[is.na(external_gene_name) & ensembl_ID %like%
     "ENSDARP"]$ensembl_ID,
                                    desired_attributes,
                                    "ensembl_peptide_id",
                                    host = "Jul2016.archive.ensembl.org")
colnames(missing_proteins)[colnames(missing_proteins) %like% "zfin"] <- "zfin_id_id"

desired_attributes <-c("external_gene_name",
                   "ensembl_transcript_id",
                   "ensembl_peptide_id",
                   "ensembl_gene_id",
                   "description",
                   "name_1006",
                   "zfin_id_id")
de_prots[missing_transcripts, on=.(ensembl_ID), (desired_attributes) := mget(paste0("i.", desired_attributes))]
de_prots[missing_proteins, on=.(ensembl_ID), (desired_attributes) := mget(paste0("i.", desired_attributes))]
de_prots[!is.na(zfin_ID), zfin_id_id := zfin_ID]
de_prots[, zfin_ID := NULL]
de_prots[, zfin_id_id := as.character(zfin_id_id)]
```

```
de_prots[external_gene_name == "",]$external_gene_name <- NA
de_prots[zfin_id_id == "",]$zfin_id_id <- NA

# Do some cleanup for broken links
de_prots[external_gene_name == "cyp2k19",]$zfin_ID <- "ZDB-GENE-091211-1"
de_prots[external_gene_name == "cyp2k19",]$gene_id <- "ZDB-GENE-091211-1"

# This doesn't seem to get pulled, despite be linked in zfin
de_prots[ensembl_ID == "ENSDARP00000140915",]$external_gene_name <- "zgc:172341"

# This protein/transcript has been removed from current ensembl database
de_prots[ensembl_ID == "ENSDART00000158091",]$external_gene_name <- "zgc:101540"
# rename the 'ensembl_gene_id column so that if we have to update things it still makes sense
# colnames(de_prots)[colnames(de_prots) == "ensembl_gene_id"] <- "gene_id"
setnames(de_prots, c("ensembl_gene_id", "zfin_id_id"), c("gene_id", "zfin_ID"))

#### Save differentially expressed proteins ####
# Write out DE proteins
de_prots$name_1006 <- sapply(de_prots$name_1006, FUN=paste,collapse=";")

de_prots <- de_prots[, c("ensembl_ID", "protein", "description",
                    "p_value_out", "p_value_down", "p_value_btwn",
                    "logfc_glen", "logfc_down", "logfc_btwn",
                    "glen_sig", "down_sig", "btwn_sig",
                    "zfin_ID", "external_gene_name",
                    "gene_id", "ensembl_transcript_id", "ensembl_peptide_id",
                    "name_1006")]

full_expression_data <- data.table(quant_data)
full_expression_data[de_prots, on=.(protein), gene_id := gene_id]
setcolorder(full_expression_data, c("gene_id", "protein",
                                colnameslike(full_expression_data, "BEAR"),
                                colnameslike(full_expression_data, "CUSH"),
                                colnameslike(full_expression_data, "GLEN"),
                                colnameslike(full_expression_data, "H22X"),
                                colnameslike(full_expression_data, "UPHI")))
setnames(full_expression_data, c("gene_id", "protein"), c("NAME", "DESCRIPTION"))

if(save_data){
   write.table(de_prots[glen_sig != 0 | down_sig != 0 | btwn_sig !=0,
        ],file=file.path(out_dir,"sig_de_proteins_with_descriptions.tsv"), sep="\t")
   write.table(full_expression_data, file=file.path(out_dir, "expression_data_with_ens_gene_id.txt"), sep="\t",
        row.names = FALSE, quote = FALSE)
}
# #### Write out file for DAVID and FGnet analysis ####
# # create an output directory for all the fgnet files
# fgnet_dir <- make_dir("fgnet_output", out_dir)
# print(paste("FGNet output will be stored in", fgnet_dir))
#
# # GLEN site
# fgnet_glen_aux <- fgnet_data_setup(de_prots[de_prots$glen_sig !=0,
# c("gene_id","external_gene_name", "glen_sig")], "glen_site", fgnet_dir)
#
# # FGNET doesn't understand how file paths work
# current_dir <- getwd()
# setwd(fgnet_dir)
# # format david results
# glen_david <- format_david(normalizePath("glen_site_david_clusters.txt"), geneLabels = fgnet_glen_aux$labels)
# # Generate FGNET stuff
# FGNet_report(glen_david, geneExpr = fgnet_glen_aux$expr, plotKeggPw=TRUE)
# setwd(current_dir)
#
# # Downstream site
# fgnet_downstream_aux <- fgnet_data_setup(de_prots[de_prots$down_sig !=0,
# c("gene_id","external_gene_name", "down_sig")], "downstream", fgnet_dir)
#
# # FGNET doesn't understand how file paths work
# current_dir <- getwd()
# setwd(fgnet_dir)
# # Upload david_expressed_proteins_down file to DAVID and save results of the analysis clustering
# # format david results
# down_david <- format_david(normalizePath("downstream_david_clusters.txt"), geneLabels = fgnet_downstream_aux$labels)
# # Generate FGNET stuff
# FGNet_report(down_david, geneExpr = fgnet_downstream_aux$expr, plotKeggPw=TRUE)
# setwd(current_dir)
```

Listing B.2: Python code for parsing high quality protein IDs from the full protein ID list in peptide output data.

```
#! /usr/bin/env python3

import pandas as pd
import sys
from itertools import chain
import gffutils

def get_gene_feature(protein_id, predictions_db, transcript_db):
   try:
```

```
        feature = predictions_db[protein_id]
    except gffutils.FeatureNotFoundError:
        # TransDecoder appends ".pX" to the end of each gene name, need to remove for the ID to match
        feature = transcript_db[protein_id.rsplit(".",1)[0]]

    return feature


def find_overlapping_genes(protein_list, predictions_db, transcript_db):

    if len(protein_list) == 1:
        return protein_list

    if any([protein_id.startswith("DECOY") for protein_id in protein_list]):
        return []

    features_list = [get_gene_feature(protein, predictions_db, transcript_db) for protein in protein_list]

    # if any of the gene features are on separate chromosomes assume they don't overlap
    if len(set([feature.chrom for feature in features_list])) > 1:
        return protein_list

    feature_positions = [range(feature.start, feature.end) for feature in features_list]
    feature_lengths = [len(coords) for coords in feature_positions]
    largest_feature = feature_lengths.index(max(feature_lengths))
    largest_coords = set(feature_positions[largest_feature])

    if min([len(position)/len(largest_coords.intersection(position)) if len(largest_coords.intersection(position)) > 0
          else 0 for position in feature_positions]) > 0.5:
        protein_list = [protein_list[largest_feature]]

    return(protein_list)


def clean_prot_list(prot_file, pep_file, augustus_db_location, zfin_db_location, contaminant_strings=None):

    augustus_db = gffutils.FeatureDB(augustus_db_location)
    zfin_db = gffutils.FeatureDB(zfin_db_location)

    # Get the set of proteins passing the quality filter from the TPP output tsv
    prot_dt = pd.read_csv(prot_file, sep="\t")
    # split protein ids into a list by the delimiter and discard the last entry (empty string)
    prot_dt["protein"] = prot_dt["protein"].str.split(" ").str[:-1]
    prot_dt["protein"] = prot_dt["protein"].apply(find_overlapping_genes, args=[augustus_db, zfin_db])
    proteins = set(chain(*prot_dt.protein))

    # find all peptide that match to proteins from the filtered set
    pep_dt = pd.read_csv(pep_file, sep="\t")
    if contaminant_strings is not None:
        for contaminant in contaminant_strings:
            pep_dt = pep_dt[pep_dt["protein"].str.contains(contaminant)==False]
    pep_dt['protein'] = pep_dt['protein'].apply(lambda protein:
          list(set(protein.split(sep=","))).intersection(proteins)))

    # Find all peptides that match to only one protein from the filtered set
    uniq_prot_rows = pep_dt["protein"].str.len() == 1
    uniq_pep_indx = pep_dt[uniq_prot_rows].index
    uniq_pep_dt = pep_dt.loc[uniq_pep_indx, :]

    uniq_pep_dt['protein'] = uniq_pep_dt['protein'].apply(lambda x: "".join(map(str,x)))

    return [uniq_pep_dt, prot_dt]


if __name__ == '__main__':

    if len(sys.argv) == 6:
        out_dt = clean_prot_list(sys.argv[1], sys.argv[2], sys.argv[3], sys.argv[4])
        out_dt.to_csv(path_or_buf=sys.argv[5], sep="\t")

    else:
        print("Usage: %s protein_id.tsv peptide_id.tsv, augustus_db, zf_db, output_file_name" % sys.argv[0])
```

Listing B.3: R code for normalizing and collating report ion data for different peptides.

```
library(ggplot2)
library(data.table)
library(limma)
library(qvalue)
library(vsn)
library(statmod)
library(corrplot)
library(Cairo)


make_dir <- function(dir_name, location="."){
    full_path <- file.path(location, dir_name)
    if(!file_test("-d", full_path)){
        if(file_test("-f", full_path)){
            stop(paste("can't create folder", full_path, "because a file with that name alread exists."))
        } else {
```

```
            dir.create(full_path)
            return(full_path)
        }
    }
}


quantify.proteins.vsn <- function(dat, cha, housekeeper_ids=NULL){
  ### Taken from http://www.biostat.jhsph.edu/~kkammers/software/CVproteomics/R_guide.html ###
  e.function <- function(x, seq) tapply(x, seq, median)
  output <- NULL

  dat$Sequence <- toupper(dat$Sequence) # Capital letters
  accessions <- as.character(unique(dat$Protein.Group.Accessions))
  n.proteins <- length(accessions)
  n.cha <- length(cha)

  for(k in 1:n.proteins){
    id <- accessions[k]
    sdat <- subset(dat, Protein.Group.Accessions==id)[c("Sequence", cha)]
    pdat <- sdat[, -1]
    n.spectra <- ifelse(is.integer(dim(pdat)), nrow(pdat), 1)
    temp <- apply(sdat[,-1], 2, e.function,seq=sdat[, 1])
    n.peptides <- ifelse(is.integer(dim(temp)), nrow(temp), 1)
    pdat <- apply(sdat[cha], 2, sum)
    pdat <- c(pdat, n.peptides=n.peptides, n.spectra=n.spectra)
    output <- rbind(output, pdat)
  }

  row.names(output) <- accessions

  # test_dt <- data.table(output, keep.rownames = TRUE)
  # setnames(test_dt, c("rn"), c("protein"))
  # test_dt[,ionSum := rowSums(.SD), .SDcols = 2:11]
  # unique_list <- lapply(colnames(test_dt)[2:11], function(sample_id){unique(test_dt[get(sample_id) == ionSum &
      ionSum > 0,]$protein)})
  # names(unique_list) <- colnames(test_dt)[2:11]
  # print(unique_list)
  old_data <- output[,1:n.cha]
  output[,1:n.cha] <- justvsn(as.matrix(output[,1:n.cha]))

  return(output)
}


plot_sum_intensity <- function(pep_dt, plot_title=NULL){
    # plot total ion intensity of each report ions by sample name

    sum_intensity=data.frame(value=apply(pep_dt[3:ncol(pep_dt)],2,sum))
    sum_intensity$reporter=rownames(sum_intensity)
    sum_intensity <- data.table(sum_intensity)
    sum_intensity[, c("site", "run", "cage", "sex") := tstrsplit(reporter, "\\.")]
    raw_intensity <- ggplot(sum_intensity, aes(x=reporter, y=value, fill=site)) +
        geom_bar(stat="identity") +
        theme_classic() +
        ylab("Reporter Ion Intensity") +
        xlab("Sample") +
        ylim(c(0, 1.8e9)) +
        labs(fill = "Caging Site") +
        ggtitle(plot_title) +
        scale_fill_brewer(type = "qual", palette = "Set1") +
        theme(text = element_text(size=20), axis.text.x = element_text(angle = 90, hjust=0, vjust=0.5))
    return(raw_intensity)
}

plot_logfc_boxplot <- function(intensity_dt, plot_title=NULL, column_list=NULL, sort_columns=FALSE){
    if(is.null(column_list)){
        column_list <- c(1:ncol(intensity_dt))
    }
    intensity_dt <- data.table(intensity_dt)
    if(sort_columns){
        column_names <- as.character(colnames(intensity_dt[, ..column_list]))
        setcolorder(intensity_dt, sort(column_names))
    }
    melt_data <- melt(intensity_dt[,..column_list], variable.name = "sample_id", value.name = "log_intensity")
    melt_data[, c("site", "run", "cage", "sex") := tstrsplit(sample_id, "\\.")]
    normalized_plot <- ggplot(melt_data, aes(x=sample_id, y=log_intensity, fill=site, linetype=sex)) +
        geom_boxplot(notch = TRUE) +
        theme_classic() +
        ylab("Log2 Fold Change") +
        xlab("Sample") +
        ggtitle(plot_title) +
        xlab("Sample ID") +
        ylab("Log2 Intensity") +
        labs(fill = "Caging Site", sex="Sex") +
        theme(text = element_text(size=20), axis.text.x = element_text(angle = 90, vjust=0.5, hjust=0)) +
        scale_fill_brewer(type = "qual", palette = "Set1")
    return(normalized_plot)
}

plot_logfc_density <- function(normalized_matrix, column_list, plot_title=NULL){
    melt_data <- melt(normalized_matrix[,column_list], variable.name = "Sample", value.name = "log2FC")
```

```
    p <- ggplot(data=melt_data, aes(x=log2FC, color=Sample, fill=Sample)) +
        geom_density(alpha=0.01, size=1) +
        scale_color_brewer(type="qual", palette = "Paired") +
        scale_fill_brewer(type="qual", palette = "Paired") +
        ggtitle(plot_title) +
        theme_classic() +
        ylab("Density") +
        geom_vline(xintercept = 0, linetype="dotted") +
        theme(text = element_text(size=20))

    return(p)
}

output_plot <- function(plot, image_dir=NULL, image_name = NULL, width=8, height=8, units="in", res=300, ...){
    if (is.null(image_dir)){
        plot(plot)
    } else{
        image_path <- file.path(image_dir, image_name)
        png(type="cairo", filename = image_path, width=width, height = height, units = units, res = res)
        plot(plot)
        dev.off()
    }
}

run.quant <- function(quant_dt_raw, design_table, minimum_reporters, drop_channels = NULL,
                      housekeeper_ids = NULL, img_dir = ".", save_plots = FALSE){

    # Get the experiment label from the raw data - should only be one id
    exp_label <- unique(quant_dt_raw$experiment_label)

    if(!save_plots){
        img_dir=NULL
    }

    # rename TPP columns to something sensible - this need to be made flexible for tmt/itraq
    quant_dt <- quant_dt_raw[,list(Protein.Group.Accessions = protein, Sequence = peptide,
                                   tag126N = libra1, tag127N = libra2, tag127C = libra3,
                                   tag128N = libra4, tag128C = libra5, tag129N = libra6,
                                   tag129C = libra7, tag130N = libra8, tag130C = libra9,
                                   tag131N = libra10)]
    print(paste0("All: ", nrow(quant_dt)))
    # for some reason data table breaks something below
    quant_dt <- as.data.frame(quant_dt)
    # get the last column so we can set the range for plotting/normalization
    last_col <- ncol(quant_dt)

    # convert columns from isobaric tag to sample name
    sample_ids <- unlist(design_table[run==exp_label, 3:ncol(design_table)], use.names = FALSE)
    colnames(quant_dt)[3:last_col] <- sample_ids
    quant_dt <- quant_dt[,c(1,2,order(sample_ids)+2)]
    if(!(is.null(drop_channels))){
        quant_dt <- quant_dt[ , !(names(quant_dt) %in% drop_channels)]
        sample_ids <- sample_ids[!(sample_ids %in% drop_channels)]
    }
    # get the total intensity for each sample/channel and plot it
    img_1 <- paste(exp_label, "total_ion_inten.png", sep="_")
    output_plot(plot_sum_intensity(quant_dt, NULL), img_dir, img_1)

    # Remove peptides with zero intensity ions in more than x channels\
    max_missing <- length(sample_ids) - minimum_reporters
    quant_dt <- quant_dt[rowSums(quant_dt[,3:ncol(quant_dt)] == 0) <= max_missing,]
    # Plot new intensity for each channel/sample
    img_2 <- paste(exp_label, "filtered_ion_intensity.png", sep="_")
    output_plot(plot_sum_intensity(quant_dt, NULL), img_dir, img_2)

    cha <- sample_ids[order(sample_ids)]
    number_channels <- length(cha)
    print(paste0("Min Reporters: ", nrow(quant_dt)))

    if ( !(is.null(housekeeper_ids)) ) {
        peptide_correlation <- cor(quant_dt[quant_dt$Protein.Group.Accessions %in% housekeeper_ids,cha])
        #peptide_correlation_title <- paste(exp_label, "Raw Peptide Level Correlation between Housekeeping Proteins")
        peptide_correlation_title <- NULL
        if (save_plots){
            plot_file <- file.path(img_dir, paste(exp_label,"peptide_level_correlation.png", sep="_"))
            png(filename = plot_file, width=8, height=8, units="in", res=300)
            corrplot(peptide_correlation, type="upper", method = "number", title=peptide_correlation_title)
            dev.off()
        }else{
            corrplot(peptide_correlation, type="upper", method = "number", title=peptide_correlation_title)
        }
    }

    normalized_prot <- quantify.proteins.vsn(as.data.frame(quant_dt), cha, housekeeper_ids)

    if ( !(is.null(housekeeper_ids)) ) {
        protein_correlation <- cor(normalized_prot[rownames(normalized_prot) %in% housekeeper_ids, cha])
        #protein_correlation_title <- paste(exp_label, "Normalized Protein Level Correlation between Housekeeping
             Proteins")
        protein_correlation_title <- NULL
        if (save_plots){
            plot_file <- file.path(img_dir, paste(exp_label,"protein_level_correlation.png", sep="_"))
```

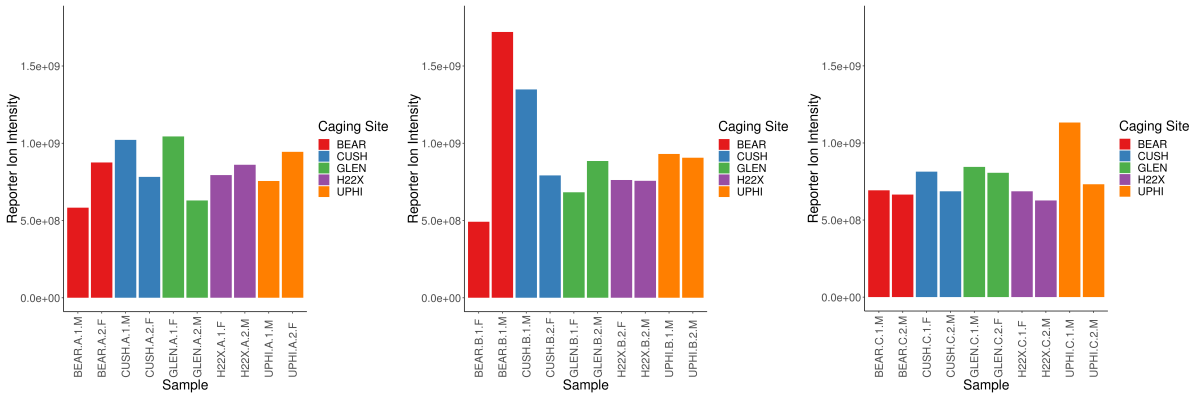
```
        png(filename = plot_file, width=8, height=8, units="in", res=300)
        my.chart.Correlation(normalized_prot[rownames(normalized_prot) %in% housekeeper_ids, cha], histogram=TRUE,
            sizing=1.05)
        dev.off()
      }else{
        my.chart.Correlation(normalized_prot[rownames(normalized_prot) %in% housekeeper_ids, cha], histogram=TRUE,
            sizing=2)
      }
   }

   normalized_prot[,1:number_channels] <- normalized_prot[,1:number_channels] -
        apply(normalized_prot[,1:number_channels], 1, median)
   quantified_prot <- as.data.frame(normalized_prot)
   quantified_prot <- subset(quantified_prot, quantified_prot$n.peptides > 1)

   #quantified_prot[quantified_prot == 0] <- NA
   quantified_prot <- na.omit(quantified_prot)

   img_6 <- paste(exp_label, "normalized_logFC_density.png", sep="_")
   output_plot(plot_logfc_density(quantified_prot, column_list = cha), img_dir, img_6)

   quantified_prot <- data.frame(protein = row.names(quantified_prot), quantified_prot)

   return(quantified_prot)
}
```

Listing B.4: R code for annotating differentially expressed proteins from Ensembl.

```
library(biomaRt)
library(reshape2)


unlist_ensembl_attributes <- function(ensembl_id_list){
  uniq_id <- unique(unlist(as.character(ensembl_id_list)))[[1]][1]
  uniq_id <- gsub(" [[].*", "", uniq_id)
  if (length(uniq_id) ==0){
    uniq_id <- NA
  }
  return(uniq_id)
}


get_biomart_attributes <- function(ensembl_transcript_id_list, attribute_list, filter, host =
     "Dec2017.archive.ensembl.org"){
   ensembl = useMart('ENSEMBL_MART_ENSEMBL', dataset="drerio_gene_ensembl", host)
   prot_go_mapping <- getBM(mart=ensembl,attributes = c(filter, attribute_list),
                     filters=filter, values=ensembl_transcript_id_list)
   colnames(prot_go_mapping) <- c("ensembl_ID", attribute_list)
   out_dt <- data.frame()
   if(length(unique(prot_go_mapping[,1])) < dim(prot_go_mapping)[1]){
      out_dt <- aggregate(.~ensembl_ID, data=prot_go_mapping, FUN=list, na.action = na.pass)
   }
   else {
      out_dt <- prot_go_mapping
   }
   if("description" %in% colnames(out_dt)){
   out_dt$description <- sapply(out_dt$description, unlist_ensembl_attributes)
   }
   if("external_gene_name" %in% colnames(out_dt)){
   out_dt$external_gene_name <- sapply(out_dt$external_gene_name, unlist_ensembl_attributes)
   }
   if("ensembl_transcript_id" %in% colnames(out_dt)){
   out_dt$ensembl_transcript_id <- sapply(out_dt$ensembl_transcript_id, unlist_ensembl_attributes)
   }
   if("ensembl_peptide_id" %in% colnames(out_dt)){
   out_dt$ensembl_peptide_id <- sapply(out_dt$ensembl_peptide_id, unlist_ensembl_attributes)
   }
   if("ensembl_gene_id" %in% colnames(out_dt)){
      out_dt$ensembl_gene_id <- sapply(out_dt$ensembl_gene_id, unlist_ensembl_attributes)
   }
   if("zfin_id_id" %in% colnames(out_dt)){
      out_dt$zfin_id_id <- sapply(out_dt$zfin_id_id, unlist_ensembl_attributes)
   }
   if("zfin_id" %in% colnames(out_dt)){
      out_dt$zfin_id <- sapply(out_dt$zfin_id, unlist_ensembl_attributes)
   }
   return(out_dt)
}

fgnet_data_setup <- function(prot_table, file_label, out_dir="fgnet_out"){

   gene_ids <- prot_table[,1][[1]]
   gene_labels <- prot_table[,2][[1]]
   significance <- prot_table[,3][[1]]

   file_prefix <- file.path(out_dir, file_label)
   de_gene_id_file <- paste(file_prefix, "de_protein_list.csv", sep="_")

   write.table(na.omit(gene_ids), file = de_gene_id_file, row.names = FALSE,
```

```
            col.names = FALSE, quote = FALSE)

    print(paste("Upload", de_gene_id_file, "to DAVID for functional analysis"))

    expression_pattern <- setNames(significance, gene_labels)
    write.table(expression_pattern, file = paste(file_prefix, "expr_pattern.csv", sep="_"))

    label_to_id_map <- setNames(gene_labels, gene_ids)
    write.table(label_to_id_map, file=paste(file_prefix, "label_id_map.csv", sep="_"))

    fgnet_aux_data <- list(expr = expression_pattern, labels = label_to_id_map)
    return(fgnet_aux_data)
}
```

B.3 Additional Figures



(a) a) (b) b) (c) c)

Figure B.1: Sum of reporter ion intensity for the 10 reporter tags in each of the 3 mass spectrometry runs.

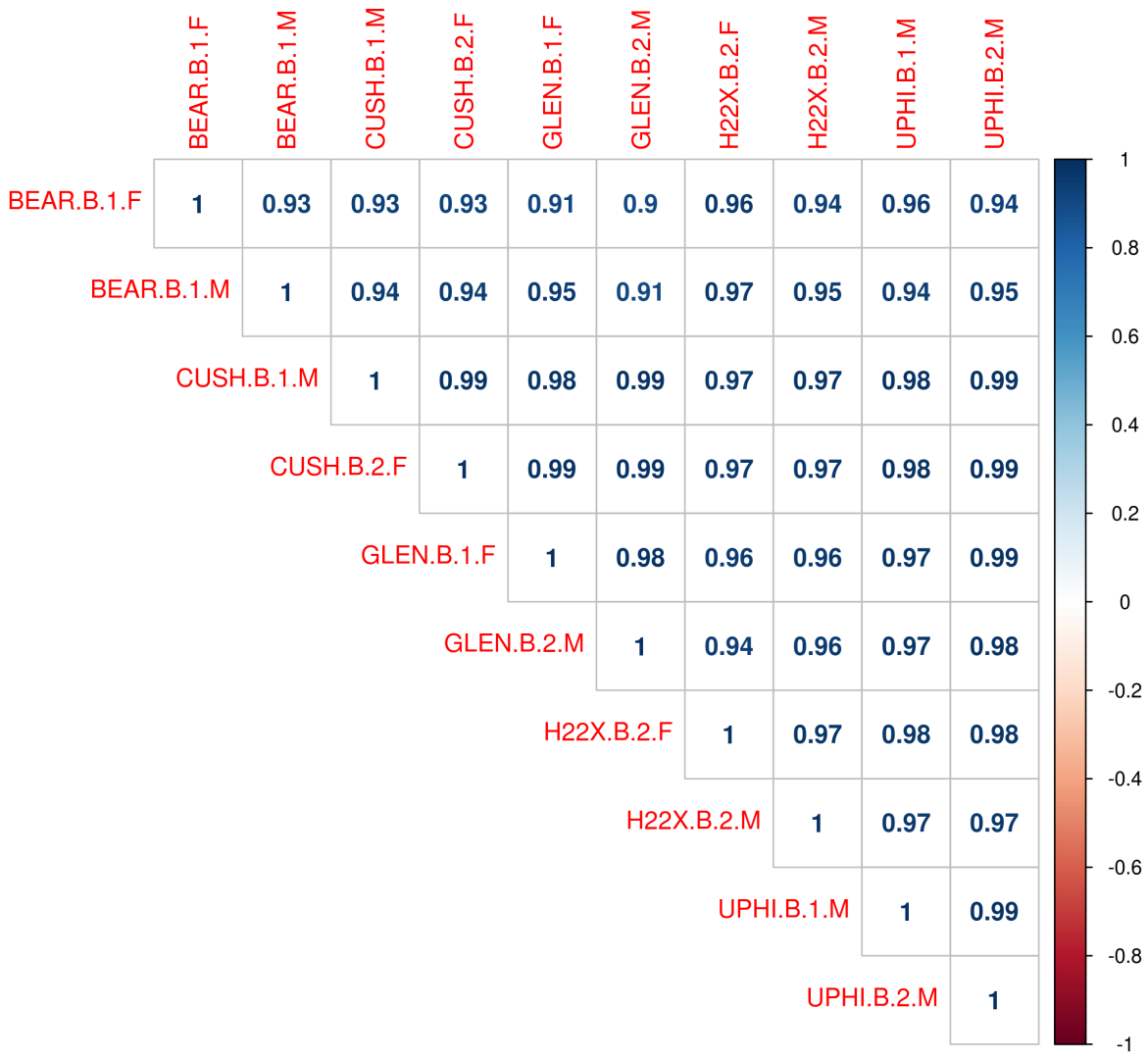


Figure B.2: Pearson correlation of PSM intensities between samples in mass spectrometry run B.

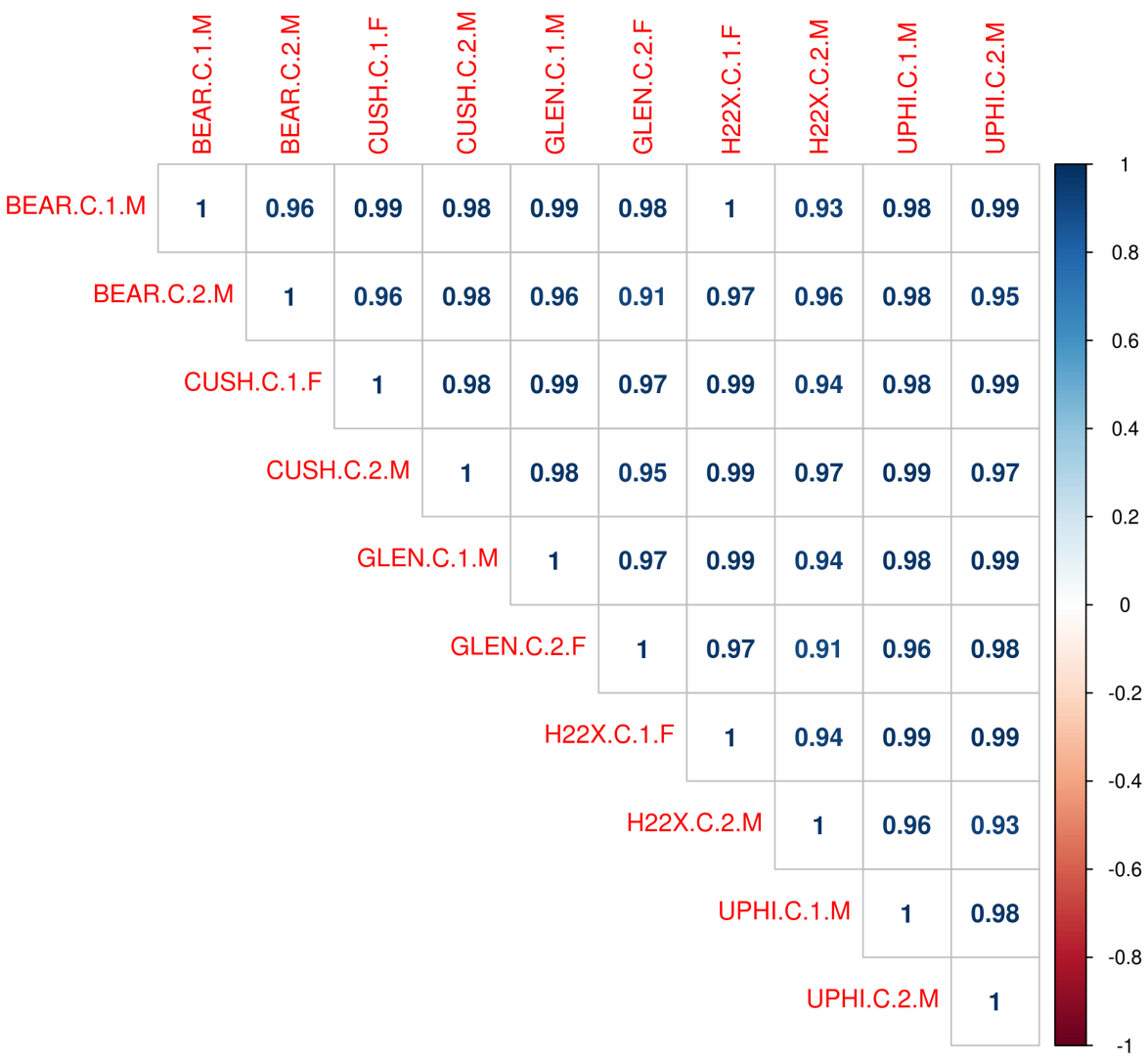


Figure B.3: Pearson correlation of PSM intensities between samples in mass spectrometry run C.

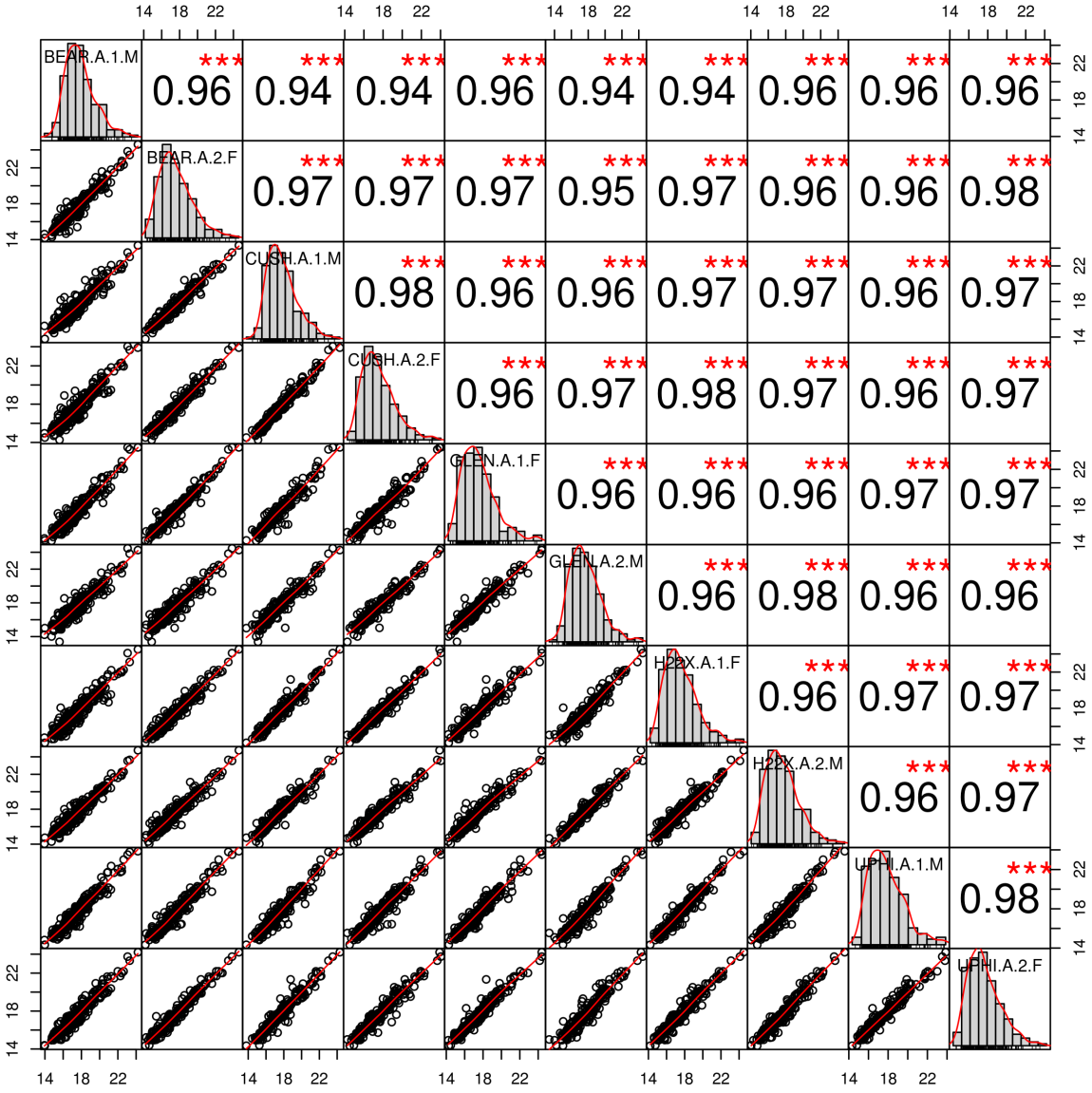


Figure B.4: Pearson correlation of housekeeping-protein expression between samples in mass spectrometry run A.

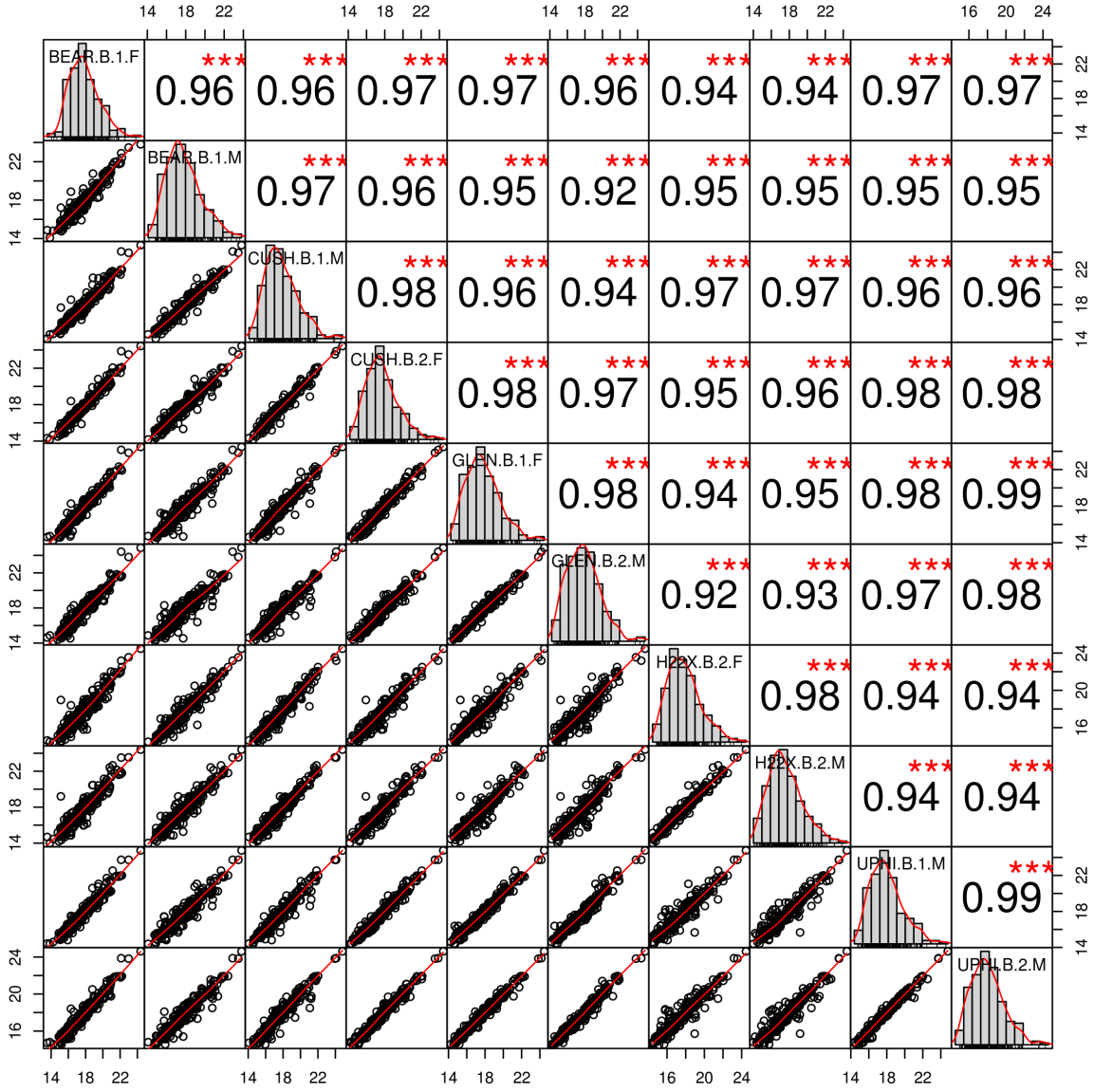


Figure B.5: Pearson correlation of housekeeping-protein expression between samples in mass spectrometry run B.

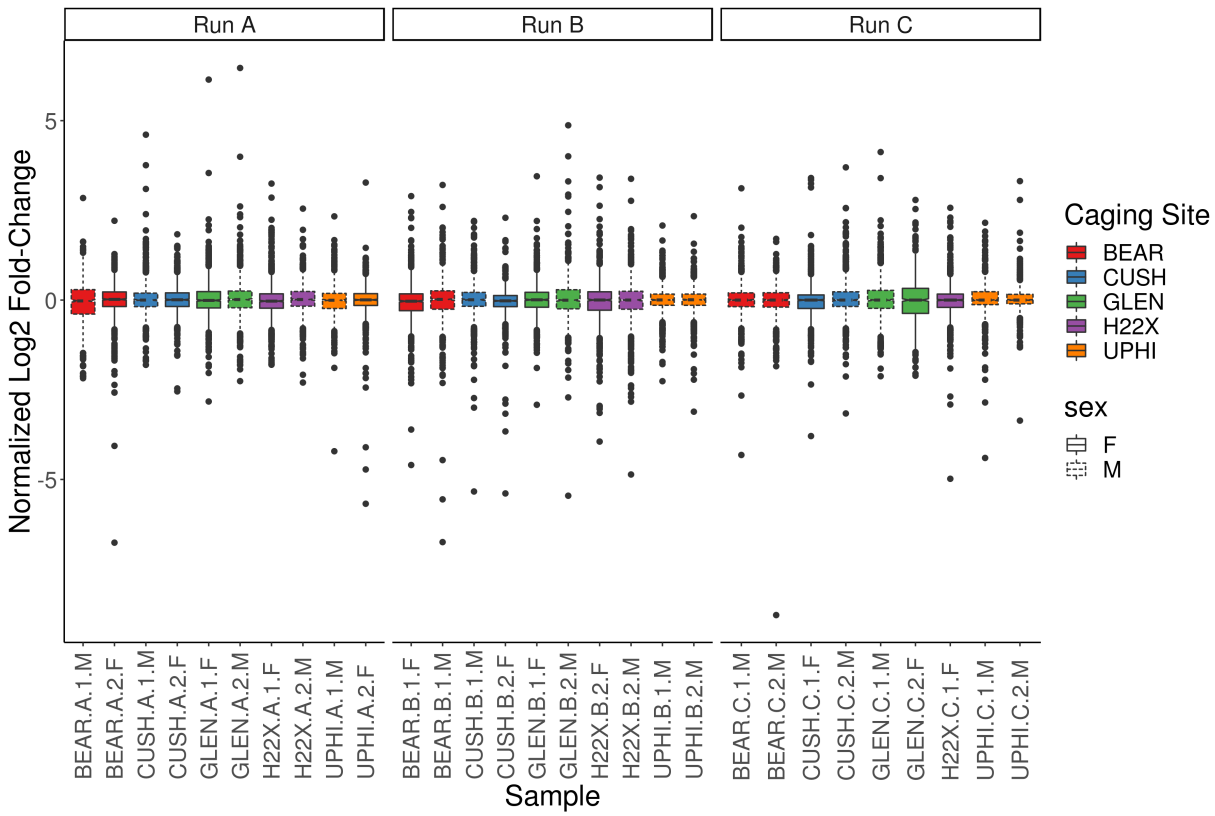


Figure B.6: Box plot of the Log_2 fold-change for each sample in the three mass spectrometry runs.

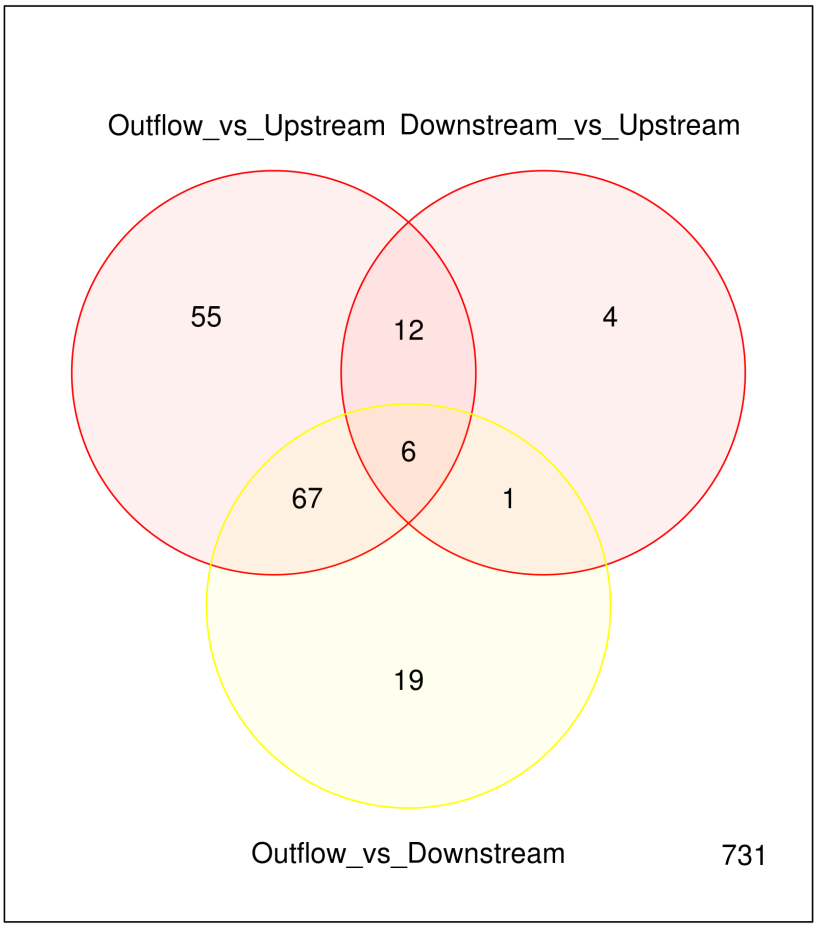


Figure B.7: Overlap between differentially expressed proteins in the different contrasts.

B.4 Outlier Removal

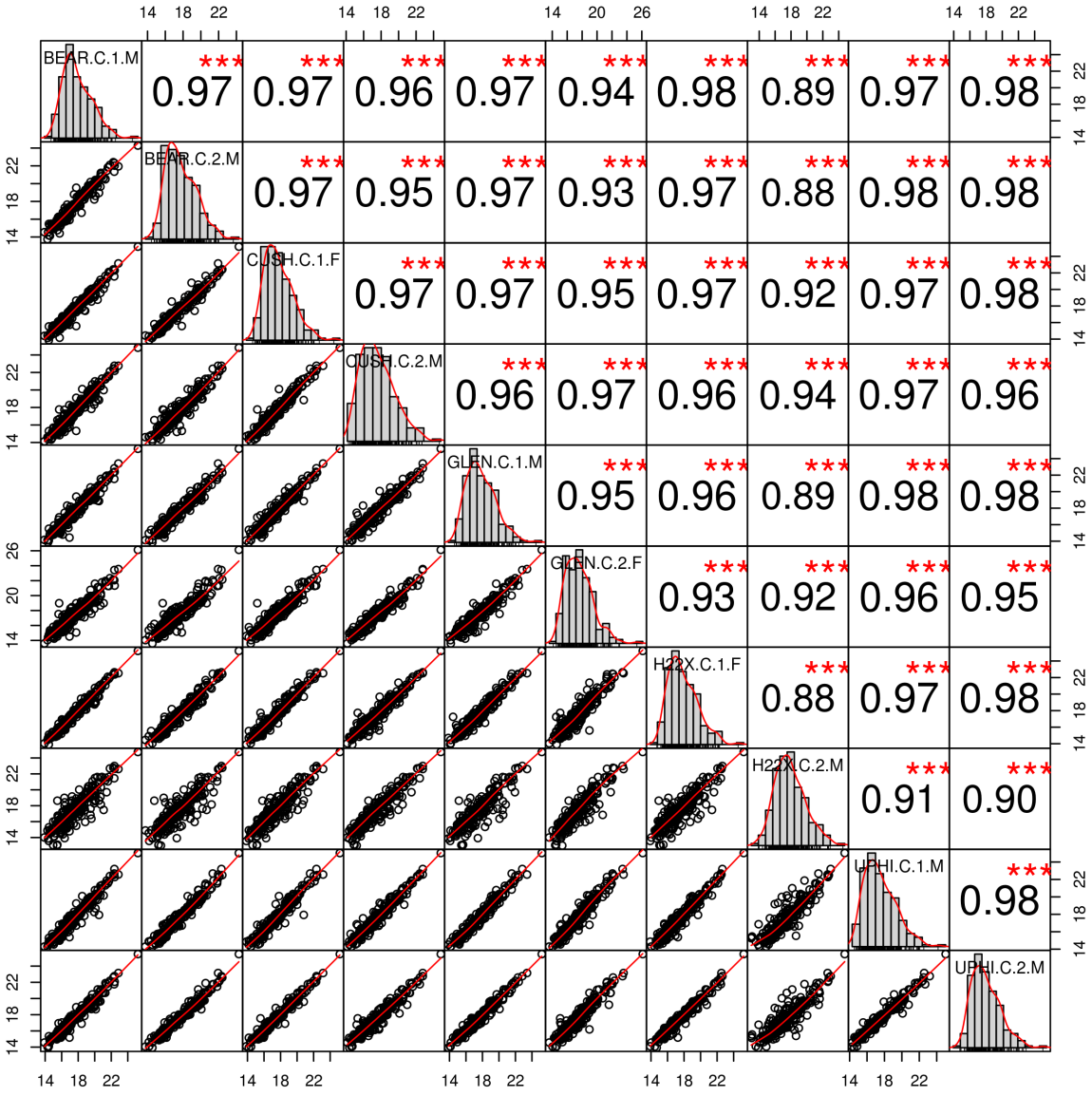


Figure B.8: Intra-run correlation of housekeeping-protein intensities aftervariance stabilized normalization (VSN) for mass spectrometry run C, including sample H22X.C.2.M. The correlation between sample H22X.C.2.M and all other samples is noticeably lower, with a maximum correlation score, 0.94, comparable to the lowest of all other sample correlations, 0.93.

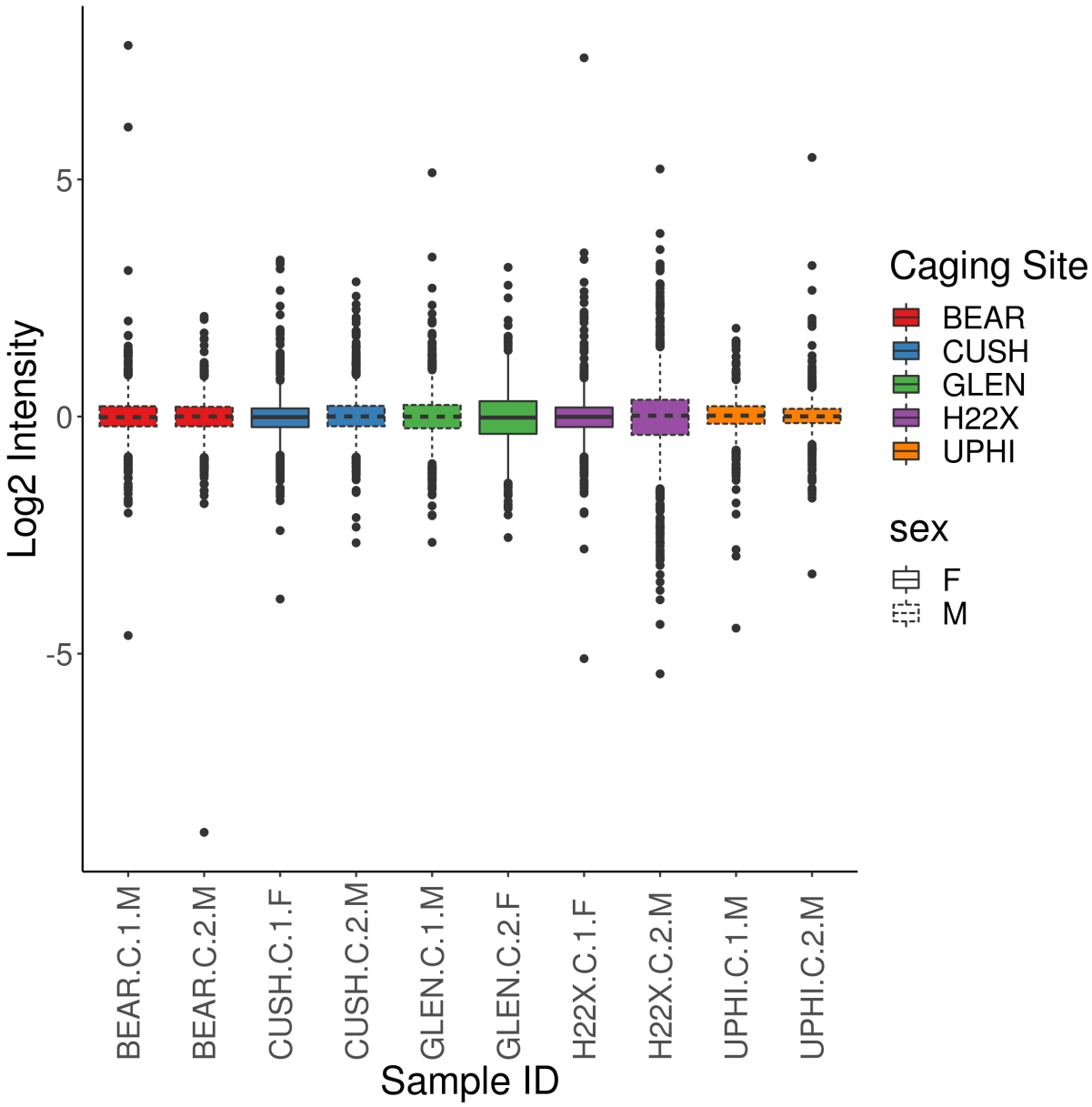


Figure B.9: Normalized Log_2 fold-change (Log_2 fold-change) of samples in run C, including the outlier sample H22X.C.2.M. The interquartile range is larger than for other samples in the same sample.

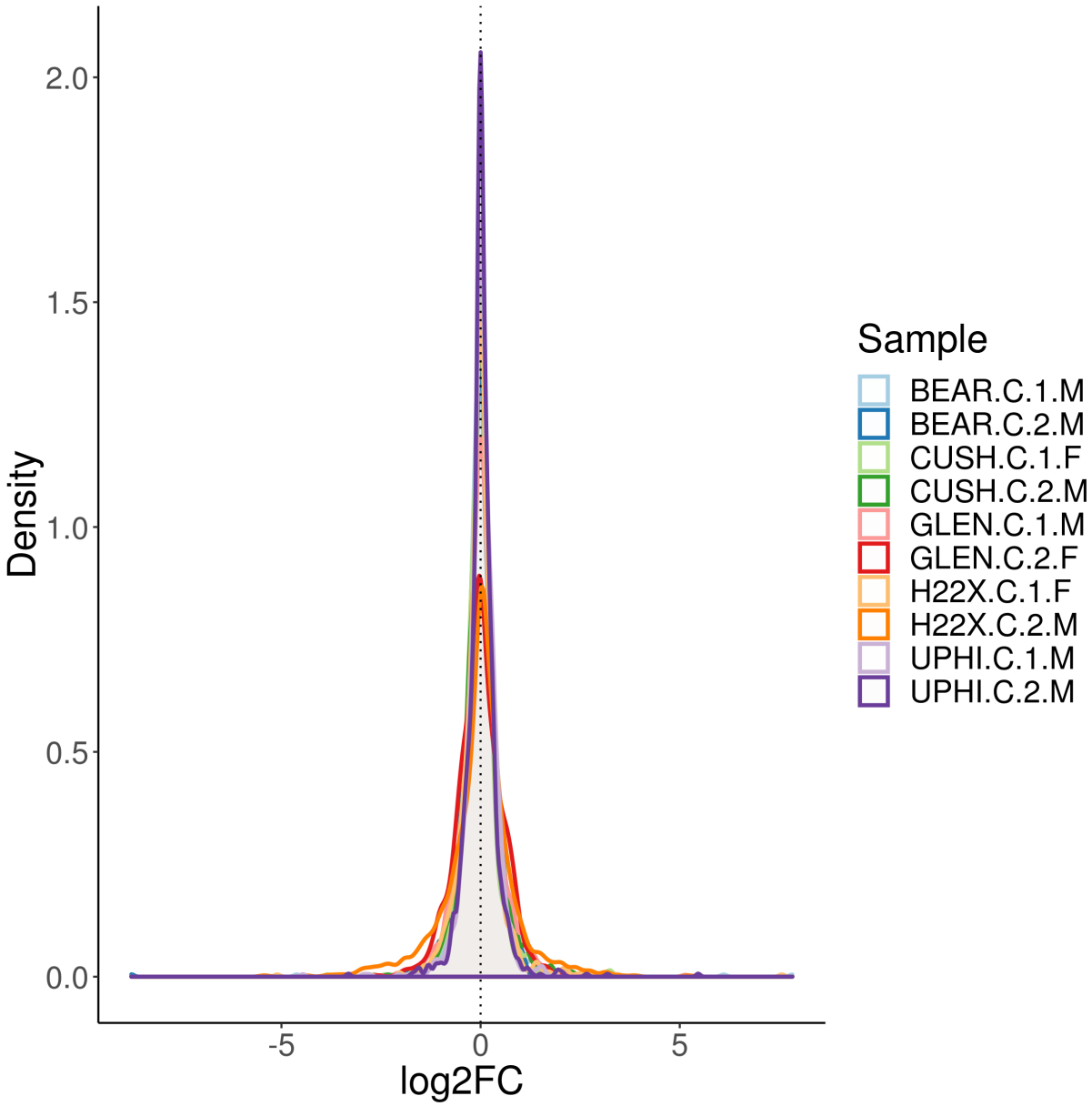


Figure B.10: Density plot of normalized Log_2 fold-change of samples in run C, including the outlier sample H22X.C.2.M. Distributions of different samples are coloured separately, with the outlying sample (H22X.C.2.M) in dark orange. The greater variance is visible at the base of the distribution, particularly for proteins with negative log-fold changes.

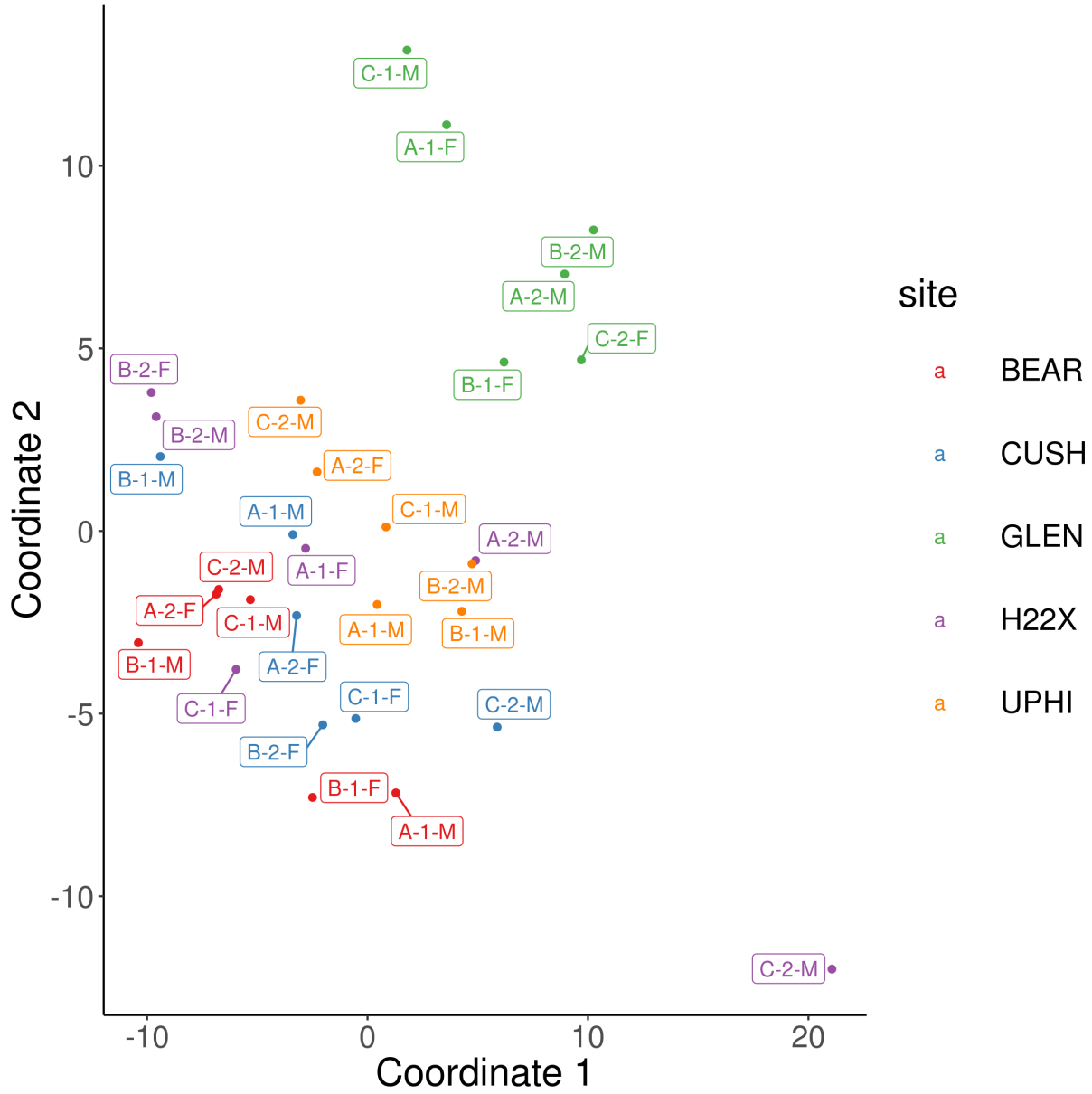


Figure B.11: multidimensional scaling (MDS) plot of normalized protein Log_2 fold-change. The label next to each point indicates which of the three mass spectrometry runs (A-C), sample cage (1 or 2), and the sex (M or F) of the sample. The outlying sample, H22X.C.2.M can be found in the lower right of the figure.

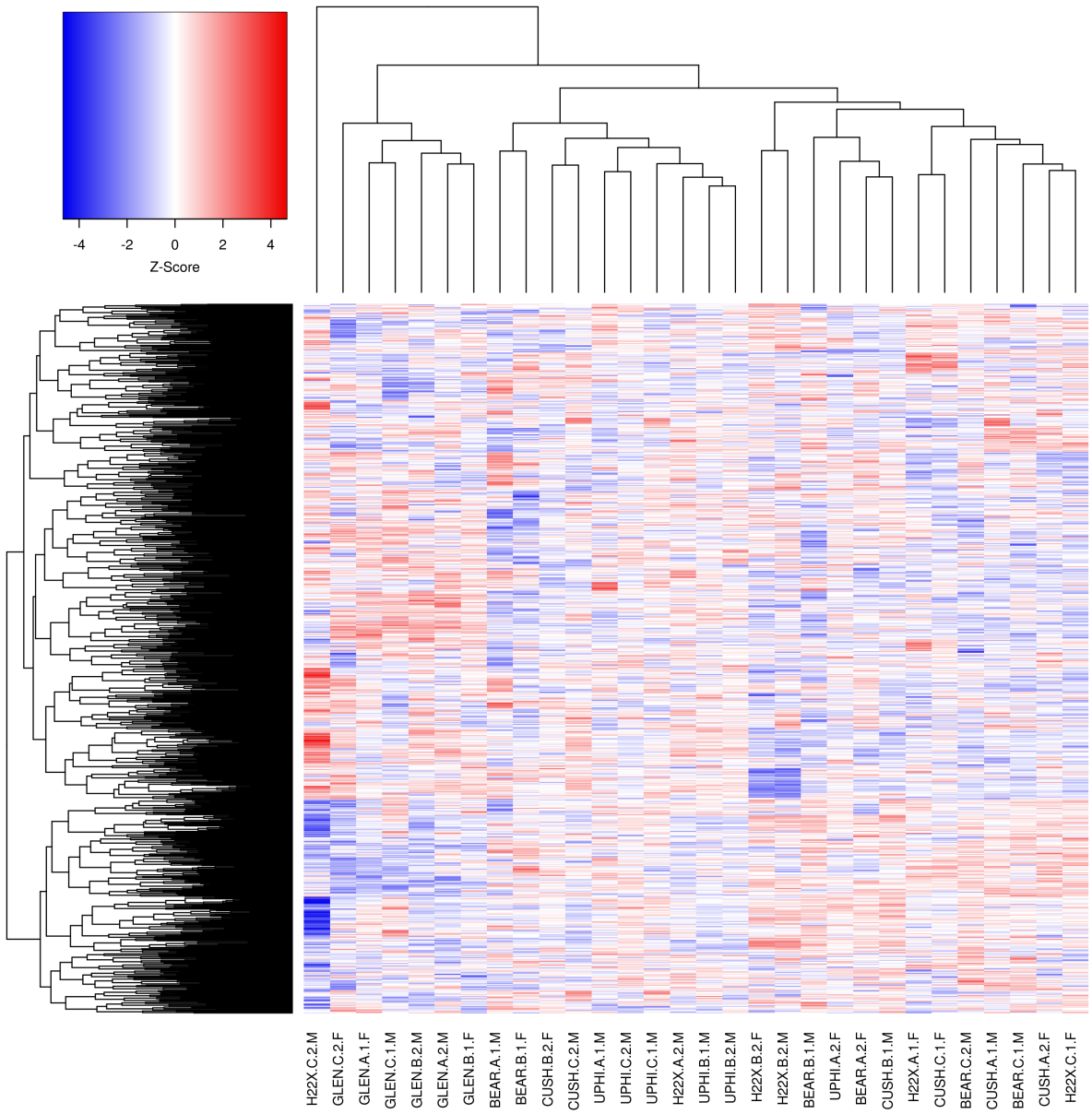


Figure B.12: Heatmap of samples. Samples with similar expression patterns are grouped together.

B.5 KEGG Diagrams

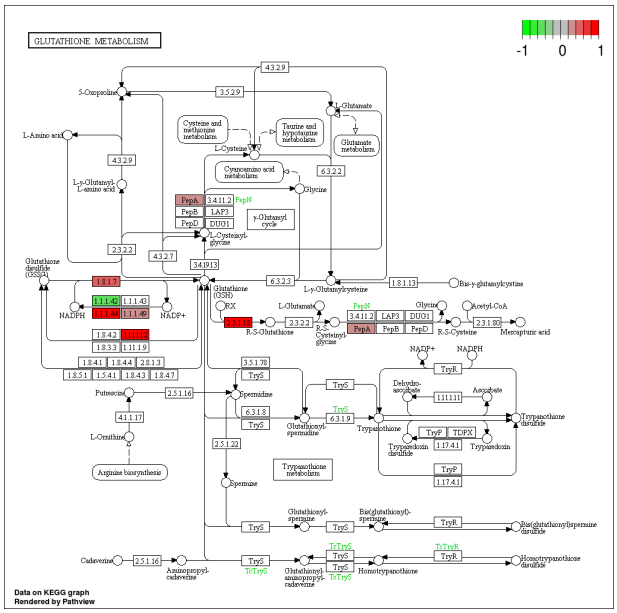


Figure B.13: Differential expression of proteins involved in the Glutathione Metabolism pathway. PGD (EC 1.1.1.44), GPX4a (EC 1.11.1.12) and multiple GST proteins (EC 2.5.1.18) were significantly increased, G6PD (EC 1.1.1.49), GSR (EC 1.8.1.7) and zgc:152830 (PepA) were increased, but not significant. IDH1 (EC1.1.1.42) was downregulated, but not significantly.