Identifying Biomarkers of Aging and Metabolic Disease through Multi-Omics Analysis of Heterogeneous BXD Mouse Populations

Author: Moaraj Hasan
Direct Supervisor: Dr. Evan Graehl Williams,
IMSB Supervisors: Dr. Prof. Ruedi Aebersold & Dr. Nicola Zamboni
D-BSSE Supervisors: Dr. Prof. Karsten Borgwardt

Sept 1st 2017

My earliest memories of my parents are of seeing my mother writing poetry and my father working diligently at his desk, happy and busy in their work. I did not know it then, but it is one of the most precious gifts a parent can give to their child

---

This thesis is dedicated to my parents Hasan Afzal and Aneela Anjum

Contents

Abstract

A large-scale study of the BXD mouse genetic reference population metabolome, proteome, transcriptome, genome, and phenome was undertaken to determine factors involved in metabolic disease and aging. This investigation includes the use of statistical analyses to determine critical differences between metabolites, protein, and transcripts differentially expressing across diets and through the aging process in the mice.

To test the reproducibility of a metabolomics protocol used in previously published BXD mouse liver experiments a pilot study on 24 mouse livers was conducted. Additionally, the pilot study was used to optimize extraction parameters and determine whether all the homogenization and extracting steps were truly necessary. Once the procedure was optimized, 632 mice livers were subjected to metabolomics and analyzed. The mouse livers were also processed for proteomic and transcriptomics analysis using previously published protocols. As proteomics and transcriptomics sample require additional processing steps, a majority of these samples were not analyzed by the writing of this thesis.

Many differentially enriched metabolites, protein, and transcripts between the diet and age cohorts were discovered and added to a list of biomarker candidates. Next pathway analysis was performed. Firstly, the steady-state metabolite data faithfully reproduced known concentration ratios in mice. Next, all the metabolites were

plotted on their KEGG pathways and pathway in which we had high metabolic coverage and differences between age and diet cohorts was determined. Then a literature search was undertaken to determine a list of rate-limiting enzymes and metabolites in order to approximate the flux through certain pathways. The critical metabolites in the pathways were again appended to a growing list of possible aging and metabolic biomarkers.

QTL analysis found strongly regulated metabolites that had been previously identified and solved in a study with young BXD mice segregated in similar diet cohorts. The data from this experiment was used as a positive control for QTLs. Additionally, between diet fold changes of metabolites which were found in the previous study gave us a ground truth to benchmark our current sample preparation and metabolic analysis against. Additionally, three novel QTLs in central glucose metabolism, glutathione metabolism, and amino acid metabolism were found. Rudimentary analysis was performed on the proteomic and transcriptomic data as the datasets were largely incompletely at the writing of this thesis.

¡¡¡¡¡¡¡ HEAD Machine learning algorithms were used to determine the most important discriminating factors between diet and age cohorts. In the former case, a few metabolites with high bioavailability which were exclusively present in either the high fat or chow diet enabled a trivially easy determination of mouse diet. Training a classifier for age determination was not as easy using only the metabolites, however, some metabolites such as non-fully oxidized fats proved useful in discriminating between the young and old mice. This corroborates prior knowledge in which oxidation efficiency decreases as mice age. This analysis was also performed for the proteomic data as using fold change analysis between the proteomes of young and old mice did not yield many hits. ======= Machine Learning Algorithms were used to determine the most important discriminating factors between diet and age cohorts. In the former case, a few metabolites with high bioavailability which were

exclusive present in either the high fat or chow diet enabled a trivially easy determination of mouse diet. Training a classifier for age determination was not as easy using only the metabolites, however, some metabolites such as non-fully oxidized fats proved useful in discriminating between the young and old mice. This corroborates prior knowledge in which oxidation efficiency decreases as mice age. This analysis was also performed for the proteomic data as using fold change analysis between the proteomes of young and old mice did not yield many hits. ¿¿¿¿¿¿¿ f83ac35aa88ca1a19860f106f52e21e266dde4e7

From this analysis, a list of candidate genes and proteins that regulate these metabolites is made and a proof-of-concept integrated analysis is done in Linoleic acid and Tyrosine metabolism. For those pathways that have knock-out mice available or downstream enzymes, inhibitors are added to a list of metabolites to follow up with in validation experiments due to occur at the beginning of next year.

Chapter 1

Introduction

The world post-industrial revolution, once riddled with infectious disease and appalling living conditions brought on by the influx of rural peoples into the poorly equipped cities caused a striking number of deaths from conditions that rarely occur in modern life(Joe Pinsker, 2013). The tremendous progress in health outcomes today can be attributed to the sanitary revolution; the successful implementation of health efforts such as large-scale water sanitation and sewage disposal and scientific efforts such as the increase in our understanding of germ theory, the development of antibiotics and vaccines (Mackenbach, 2007). Although we are protected from many conditions that would be fatal in the past, the result of living longer and having caloric surpluses in our diets has led to the manifestation of my other chronic conditions. As a result, the focus of biomedical research has shifted from infectious disease to diseases like cancer which are age-related degeneration is a risk factor and metabolic disorders which are associated with poor diet and sedentary lifestyles.

Despite large efforts using model organisms, the understanding of the pathways and molecules involved in the onset and progression of chronic diseases. How these pathways interact with aging remains unclear to due to the financial and logistical

challenge of large longitudinal studies in vertebrates (Williams and Auwerx, 2015). Additionally, there is a large difference in the chain of translatability between chronic illnesses challenging the biomedical research community today, and the infectious diseases that are more tractable to find cures for. In diseases like bacterial infections, *in vitro* mechanistic insights and efficacy of anti-biotic activity translates well from *in vitro* into in vivo models(Moffat et al., 2017). For example, antibiotics with strong plate clearing capacity in Petri dishes have a high chance of killing the bacteria in human tissue barring safety and bioavailability concerns. The chain of translatability is much weaker for a complex disease like age related reduction of hepatic functions which can be caused by mutation accumulated through the aging process and are more difficult to mechanistically study in *in vitro* models (Moffat et al., 2017). To address this issue, the use of multiple model organisms studied in different cohorts of diet, genetic background, sex and age over the lifetime of an organism can be used to deconvolute the mechanistic roles these factors and allow us to screen interventions that could be translated to longer healthier lifespans in humans (Armanios et al., 2015). The goal of the intervention is not simply to extend lifespan but to maintain the organ systems and cognitive faculties of a person as there in their youth, into old age (Armanios et al., 2015).

1.1 Project Outline

The design of this study is to have a large group of inbred mice segregated into two diet groups and sacrificed in four age cohorts of 8, 12, 16 and 20 months. This allows us to interpret the genetically driven factors that change between the aging and control for the environmental input (diet). This study is a follow up to a study by Williams et al. (2016) in which important regulators of the electron transport chain respiration and cholesterol synthesis were found. All of the mice in the study are the progeny of two founder strains and show genetic differences at 5 million loci.

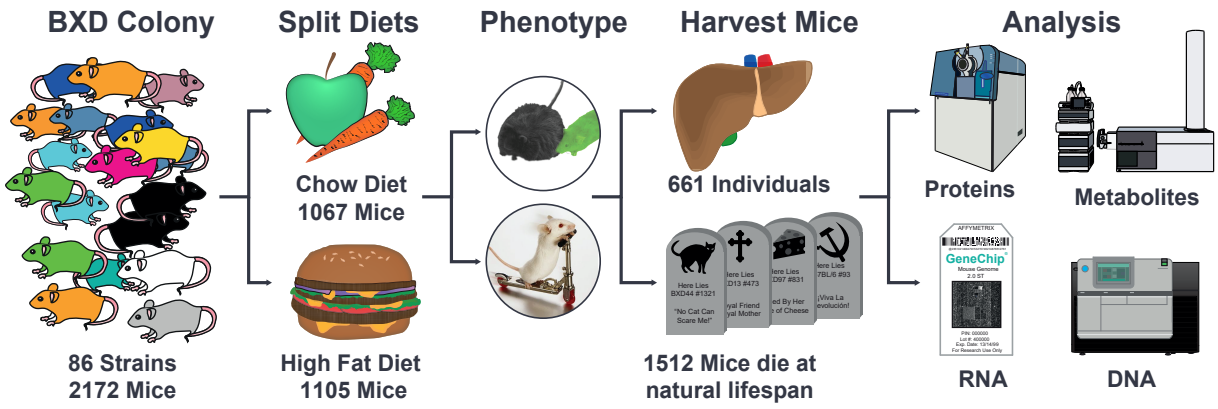


Figure 1.1: The time line of the BXD aging mouse population study. Over 2000 mouse in 83 strains are segregated into two diets and aging for eight, twelve, sixteen and twenty months. During their lifetimes all the mice are subjected to a range of phenotype examinations. Only the mice that make it to the exact time mark (ie. 4 months) are sacrificed. The mice are euthanized using a strong anesthetic and have all 26 organs harvested in a timely manner to reduce the post mortem changes in the physiology. In this study, only the frozen livers are pulverized and processing for transcriptomics, proteomics and metabolomics studies.

As the mice are kept in the same housing environment, allowed to eat and exercise *Ad libitum*, the study is designed to uncover gene interaction with age and diet that would otherwise be confounded by environmental variables. An assumption in the study is that mice from the same strain are considered biologically identical despite the stochastic nature of the expression of certain traits(Czyz et al., 2012).

At the beginning of this masters thesis, all the mice had been phenotyped, sacrificed with their liver extracted and ready for processing for the various omics analysis. Of the over 2000 mice that were enrolled in the study, 621 mice survived to their sacrifice date ensuring there at least 2 biological replicates in every age and diet cohort which are subjected to metabolomics, proteomics and transcriptomics analysis. Significantly enriched metabolites, proteins, and transcripts were all found. Metabolite QTLs found in previous studies were detected alongside two novel QTLs. Network analysis indicated many significantly enriched pathways, with respect to diet and age. Although the metabolomics analyses were completed for all the mice, pro-

teomics and transcriptomics data are incomplete, and the findings in this thesis are subject to change as new data are introduced. Machine learning techniques were used to determine age-related proteins, in a data set where it was difficult to discern between mice of different age cohorts with simply using fold changes. It was thought, the use of an ensemble of proteins, similar to the increase in power of using a haplotype rather than specific SNPs, would give us higher discrimination power between the age cohorts(Lorenz et al., 2010). For transcript data, the highly expressed genes were plotted alongside proteins on integrated maps in order to determine causal graphs for specific metabolites.

1.2 Gene-Function Paradigms

The reductionist approach to genetics involves identifying a gene of interest and altering it through means of random or directed mutagenesis, followed by observing the manifested effect of the gene's absence or super over-expression (Williams and Auwerx, 2015). In reverse genetics, a diverse panel of animals is used and phenotypes variants are probed at the genetic, proteomic, and metabolic levels to determine the sources of the varied physical characteristic. In reductive study design, when a single gene is knocked-out or augmented to show a 10-fold greater expression than the normal physiological range, the intervention is not addressing the question of what the physiological function, but rather illustrates the outcome of large changes in the physiology that are either fatal or rare in real biological systems. Moreover, reductive study design usually observe the effect of interventions on a single loci and are not highly instructive for complex diseases like diabetes (Williams and Auwerx, 2015).

In contrast, forward genetics involves determining polymorphisms that modulate certain phenotypes and disease risks. This is the generalization of medial inheritance

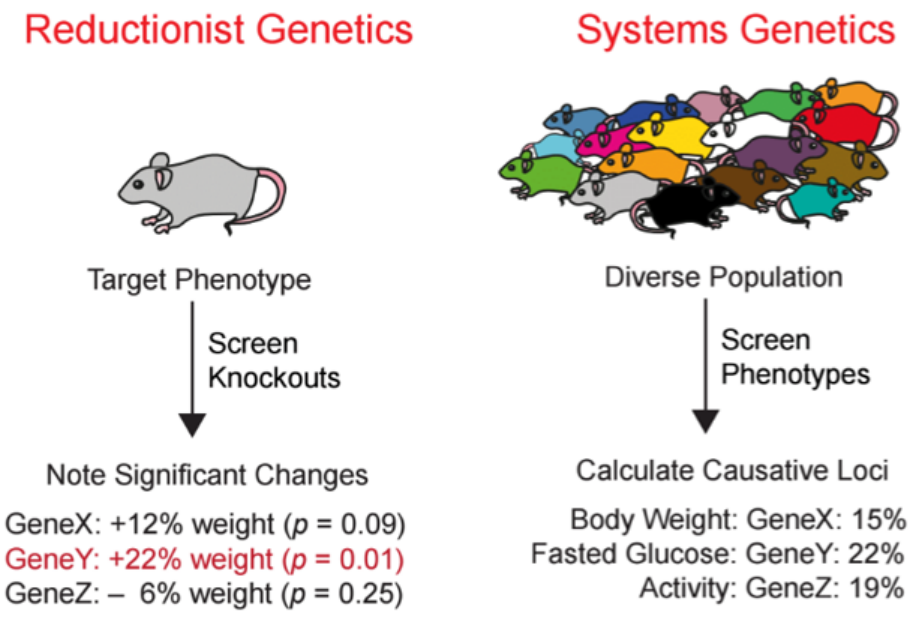


Figure 1.2: Adapted with permission from (Williams, 2014). An illustrated example of the two school of thought concerning the regulation of body weight by a single loci. **Left** In the case of reductionist genetics, genes are targeted by knock-down experiments to determine if any have a significant effect on the target phenotypes (body weight). This is done with the assumption that there is a strong direct link between the loci's expression and the phenotype.**Right** In the case of systems genetics, it is assume that genetic variants contribute small effect sizes which are pooled to determine the final phenotype. In the systems approach, natural or inbred genetically heterogeneous populations are screened in a consortium of molecular factors which can explain the variation of the phenotype in the population.

analysis applied to complex traits (Williams and Auwerx, 2015). Certain genes are altered through direct editing, silencing or mutagenesis with the commensurate phenotype used to induct the function of the gene.

As age-related degeneration is a multi-factorial and highly complex disease, systems genetics paradigm is used in this study. Many genes, proteins and metabolites may be differentially expressed as mice age, finding a genetic link between many factors that have large variances between strains in different ages can allow us to identify biomolecules that arise from age-related degeneration.

1.3 Reverse and Forward genetics

There are two different study designs that can be used to determine the function of a gene as shown in figure 1.3. With a reverse genetics design, a specific gene is disrupted usually through a knockout or by inserting a viral promoter that causes the super over-expression of genes to determine the down-stream effects. The disruption of the gene can be done in a targeted manner using siRNA, CRISPR-Cas9 or homologous recombination or can be done in a non-targeted manner such chemical or transposon-mediated mutagenesis followed by screening the library of individuals(Melinda B. Tierney and Kurt H. Lamour, 2005). For smaller organisms such as *C. elegans* or *Drosophila melanogaster*, random mutagenesis is a tenable strategy, however, for larger more complex organism targeted techniques are preferred (Melinda B. Tierney and Kurt H. Lamour, 2005).

Forward genetics refers to an experimental design where studies are initiated to determine the genetic underpinnings of observable phenotype variation between a heterogeneous population of an organism. Variants that deviate from the average trait presentation in heterogeneous population can be measured as physiological traits such as body weight and morphology to molecular variation or protein profiles, transcript abundance and DNA sequence variants (Melinda B. Tierney and Kurt H. Lamour, 2005). In the case of genetic reference populations such as the BXD mice, there is a limit to the variation that comes from the allele differences in the parental strain (Williams and Williams, 2017). Additional trait variation is introduced through controlled out breeding with members of wild populations(Melinda B. Tierney and Kurt H. Lamour, 2005).

The widespread availability of sequence data gives us many advantages, previous researchers did not have. Once individuals in populations with traits of interest are identified, having access to all gene sequences for the organism allows us to

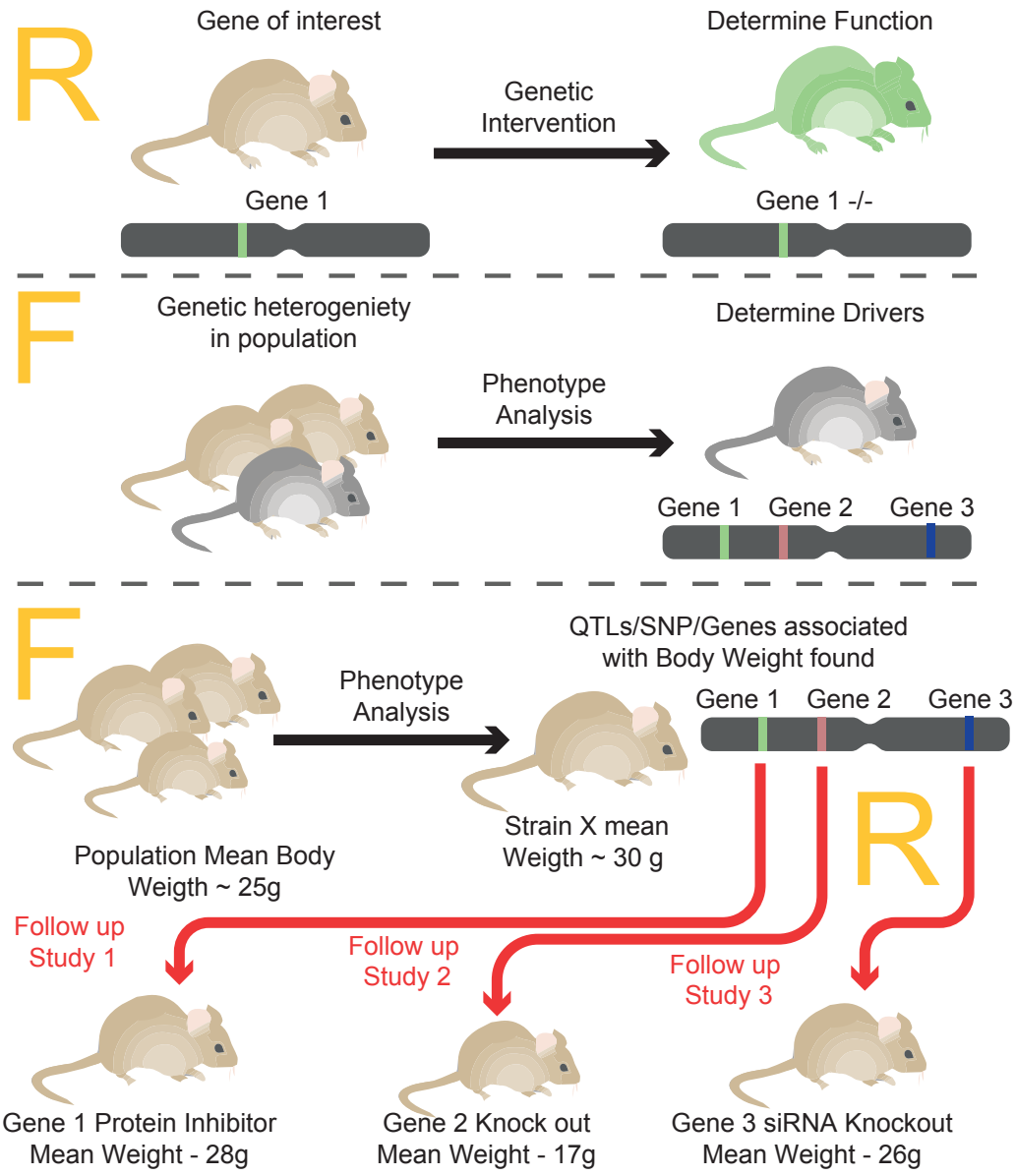


Figure 1.3: Summary of forward and reverse genetics experiments which can be used to determine the genetic drivers of a phenotype.**Top:** Schematic of the reverse genetics approach. **Middle:** Schematic of the forward genetics approach**Bottom:** Combination of both study types in which forward genetics techniques are used to find candidate loci that may regulate a phenotype, and reverse genetic techniques to validate these hypotheses.

determine a list of candidate loci that influence the phenotype of intrest and thier interaction with age, sex and diet. Once there is a finalized list of hypothesized

genes through to be driving aphenotype of intrest, the exact function of the list of these genes is elucidating using reverse genetics techniques(Melinda B. Tierney and Kurt H. Lamour, 2005). For example, some BXD mice strains exhibit high body weight despite being on a lean chow diet, genes that show hyper or hypo-expression can be or modified in a batch of mice in a follow up experiment to measure their effect on the phenotype (Figure 1.3 Bottom).

1.4 Systems Approach to Complex Trait Analysis

In this study the use of multiple biomolecule readouts for every mouse allows us to take a Systems Genetics approach to complex trait analysis. Systems Genetics is an approach that tries to understand the communication between of multiple layers of biological information and determine how the sum of their contributions results in the presentation of a complex trait (Civelek and Lusis, 2014). A systems approach requires many experimental techniques to oberserve and collect data on a range of phenotypes and statistical methods to quantify and organize molecular phenotypes like proteomes, transcriptomes, and genomes into interacting circuits with common inputs and phenotypic outputs. The output which which is the oberserved phenotype is thus generated by multiple functional molecules in the cellular milieu on the level of genes, transcripts, proteins, metabolites, and their interactions within themselves (Civelek and Lusis, 2014). Additional inputs from an organism's environment are manifested at the cellular level through changes in chemical concentrations and signaling states around cells. These environmental factors also interact with all the aforementioned molecules within a cell in complex ways (Civelek and Lusis, 2014). Figure 1.4 illustrates the complicated reciprocal nature of the interaction between genes, transcripts, metabolites and metagenomes that may lead to a specific phenotype presentation.

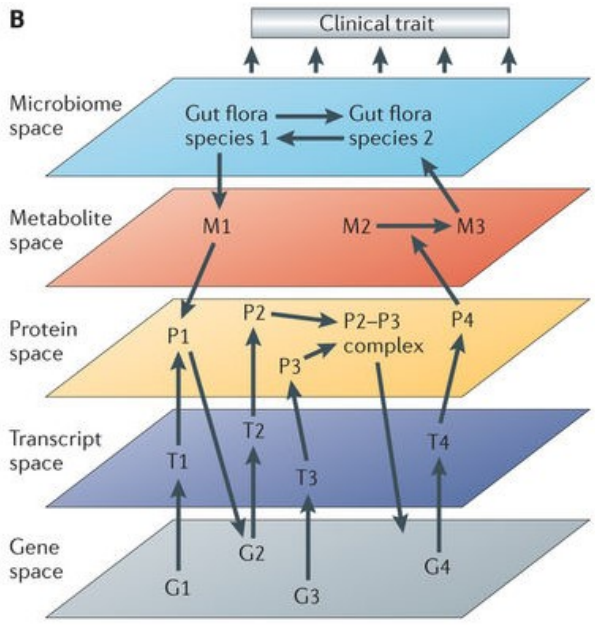


Figure 1.4: Taken from Civelek and Lusis, this figure illustrates the complex determination of a clinical by multiple interacting layers. The gene, transcript layers reciprocally interact with proteins in an or organism that bias the metabolome and increase the production of a metabolite. This metabolite causes changes in the microbiome that propagates an effect back down to the genetic level, illistrating the complex causal structures that produce a single observable trait.

Unlike Mendelian traits, a constellation of changes sum to generate a complex trait. Therefore, finding mechanisms for complex trait presentation are difficult to determine from a single layer of molecular information. Instead, there is a requirement to understand the pathways and networks that underly traits which can be symptoms of a metabolic and age-related disease. A node in these networks has a small contribution to the overall phenotype and may have stochastic components to its manifestation. As a result, the complete mechanistic decomposition of complex traits remains challenging even with large amounts of molecular data available on humans and model organism (Civelek and Lusis, 2014).

With vast amount of molecular data available to the public, it is often just as important to know which complex traits are tractable for genetic analysis.A key criteria to identifying whether the causal mechanism of a trait can be determined is

if a trait shows heritablity in the test population.

1.5 Heritability

A fundamental question in biology is whether a presenting trait is inherited from parents or the result of environmental factors. The Heritablity (H^2) is an attempt to quantify the relationship between genetics and the environment in the determination of a complex trait phenotype (Visscher et al., 2008). In order to calculate heritability, it is assumed the phenotype measured is determined by a combination of environmental and genetic factors.

$$Phenotype(P) = Genotype(G) + Environment(E) \tag{1.1}$$

All of the variation that is observed in the trait is thus the sum of variation in genetic and environmental factors (Visscher et al., 2008).

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 \tag{1.2}$$

The genetic variance in the formula can be partitioned into additive breeding or strain-related effects, effects of the dominance, and effects of epistatic interactions (Visscher et al., 2008). The narrow sense of the concept h^2 quantifies the portion of variation within an observed population that can be explained by only breeding value differences, as opposed to the broad sense values H^2 used in this which includes dominance and epistatic interactions and is defined below.

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} \tag{1.3}$$

H^2 ranges from [0.0 , 1.0], with the low heritability meaning there is a low probability phenotypes in an offspring with a specific genotype will show a particular phenotype, and high heritability values [0.7, 1.0] stating, the phenotype observed in the offspring will be similar to parents with the same allele.

In order to determine heritability, an ideal experiment involves creating an environment that is identical in every aspect for a particular population of organisms. As the population grows and develops, differences that manifest in different traits can be observed. In a well-controlled experimental study, the population is subject to the identical environmental conditions and unless the individuals in the population are genetically identical observed differences can be attributed to genetics. A constant environment is not necessary for heredity calculations if all the variations in the environments are known and can be controlled for. Thus determining the heritability of a metabolite, protein, and transcript allows us to estimate which traits have a high likelihood of segmenting between populations according to their genotype resulting in strong QTLs. Conversely, if we do not see a strong QTL where we should, a heritability measurement before and after controlling for environmental factors, if significantly higher can reveal why a strong QTL is not observed.

Caveats of Heritability Calculations

In this framework for studying heritability, if either there is no variation in the genotypes, theoretically there can be no variation in the phenotypes due to genes and heritability is zero. However, even in mono-zygotic twins or inbred mice of the exact same genotype and rearing, slightly variations in physical features can be seen. There are stochastic manifestation of environment factors, especially over a long life time that can lead to the divergance of a physical trait(Czyz et al., 2012). For example, in a pair of twins that both smoke, cancer risk does not increase in a discrete, constant level with every cigarette smoked. The incipient metastatic

event occurs through random interactions between the polyaromatic hydrocarbons from the smoke and the replication machinery in the nucleus of a lung cells. One hydrocarbon may cause a critical p53 mutation by chance in one twin leading to cancer while the other twin might smoke more yet remain healthy.

Illustrated in the example with BXD mice in the following section, heredity is difficult to deduce without environmental controls. The lack of conditional controls make studies of complex traits and their heritability in humans significantly complicated. As such, complex trait analysis in mammals are most frequently performed using a recombinant inbred population of mice (Williams and Auwerx, 2015), as is done in this project. When outbred and wild strains are used, additional complicating factors for the experimental design need to be considered and the studies have lower statistical power for the same number of mice. Additionally, heritability is hard to measure robustly for behavioral and complex traits such as intelligence but is strongly predictive in core physiological traits across large phylogenetic distances (Falconer and Mackay, 1996b). Thus highly heritable core physiological traits we may discover in mice also have a good chance of being heritable in humans.

1.6 Studying Heredity in Model Organism Populations

For consideration, BXD mouse weight has an observed heritability of 0.74. This shows that while the nutritional intake of the mice, which is the only evironemntal factors changing between the animals, has an effect on body weight, a major portion of the variation in the mouse body edihts can be explained by genetics (Gerhard Adam, 2012). An important distinciton is that hertibality explains the between mouse variance in a specifc mouse population but does not actually provide any mechanistic insights into the apidpose tissue anabolism or fatty acid metbolism that may explain the combintioan of alleles that could predict a particular weight

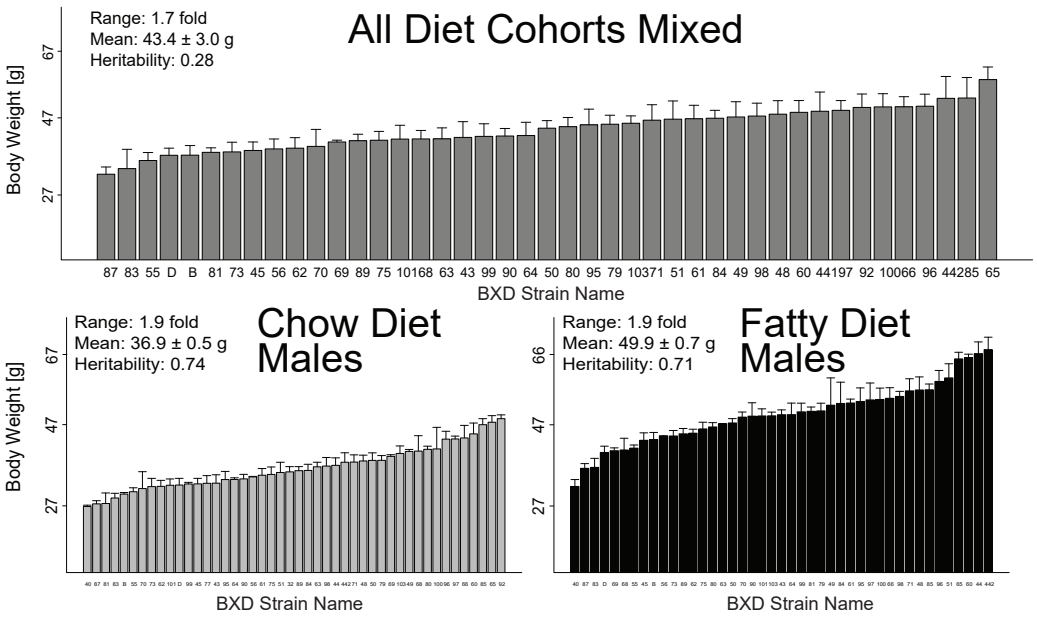


Figure 1.5: **Top:** Average body weight of BXD mouse strains in a mixed cohort of CD and HF mice. **Bottom:** Average body weight of BXD mouse strain in CD cohorts and HF cohorts separately Heritability of Body Weight and fixed and mixed Diet population

range for a mouse having a certain haplotype (Gerhard Adam, 2012).

In figure 1.5 the panel across the top shows the body weight of mice from CD and HF mice mixed in a single population. There is a mean of 43.4g and a 1.7 fold difference between the lowest body weight mouse strain BXD 87 and the highest, BXD 65. The diets have not been adequately controlled for in the mixed population so heritability is only 0.28. There is a low calculated heritability because the diets play a large role in the mean weight of the mice and must be controlled for in order to determine how of the variation in the mouse weight can be explained solely by genetics(Gerhard Adam, 2012).

The panels across the bottom of figure 1.5 show the body weights of mice segregated into CD and HF in separate cohorts. When the mouse population is segregated there is a large difference in the environment the mice a reared in but within the diet groups, the environmental conditions (dietary intake) is constant. However,

the genetic variation between the mouse strains within each of the diet cohorts still remains (Gerhard Adam, 2012). As a result, a heritability of 71% and 74% is calculated for the CD and HF cohorts respectively because the variation of the mouse body weight in each diet cohort can be mostly attributed to genetics.

This is why model organism populations are used for complex trait analysis in this study. The inbred stocks are invaluable to the study of genetic drivers and is the reason it so complex trait analysis is so difficult in human and wild populations. Humans already have several additional layers of redundancy and regulation in their biology as compared to mice(Kafri et al., 2006). If one was to compound the effect of mixed diets, family and a myriad of unknown contributing factors on top of this biological redundancy, finding QTLs and genetic drivers for phenotypes becomes quite difficult.

1.7 What is QTL Analysis?

Quantitative Trait Locus (QTL) analysis is a statistical method which permits determining the probabilty of linkage between DNA sequence variance with observed phenotypic variants (Falconer and Mackay, 1996a). In order to perform QTL analysis, populations of individuals that vary in the trait of interest are required (Mackay et al., 2009). In figure 1.6 the distribution of a phenotype parental strain along side genotype are shown. Through successive recombination events in the F2 and recombinant inbred(RI) mice, the population phenotype converges to an average of the parental phenotype in most traits.

Depending on how many loci determine a trait, a range of values can be seen in the continuous trait in offspring that have different combinations of the parental alleles. In this case (shown in figure 1.6 H0), If there is no effect from certain loci, there will be no differences in the trait with respect to the genetic background of the

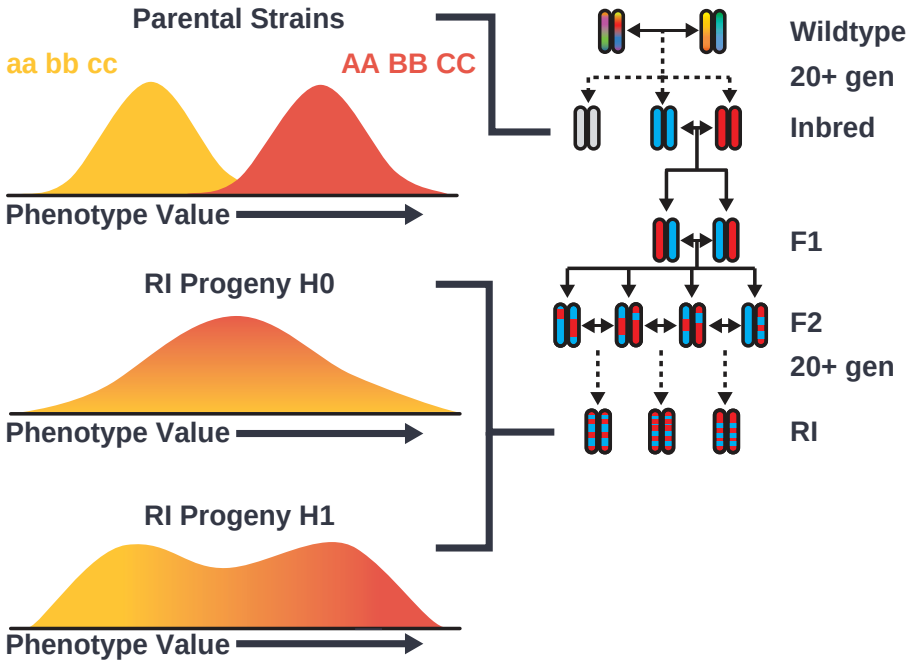


Figure 1.6: Adapted from (Williams and Williams, 2017) This figure shows the principles of QTL mapping in an inbred population derived from two parents. Distinct phenotypic trains in the parent's strains combine to form a continuum of the traits in the recombinant inbred progeny. The purpose of QTL mapping is to determine which of the inherited markers explain the variance in the trait

mice. If a locus regulates a quantitative trait, then one would see two populations of mice when all of the strains are ranked with respect to the trait and a bimodal distribution emerges (figure 1.6, H1). Strains with one of the parent's alleles in that position and one with high values for the trait with the other parental loci(Mackay et al., 2009). As one performs this test for a single trait across the entire genome, loci that segregate the strains into two populations are thought to regulate the trait. In the BXD mice, all of the strains have been sequenced and each of the marker blocks has annotated with the parental origins (Mulligan et al., 2017). To perform QTL analysis on the BXD mice, a quantitative mapping is done with a trait against an array of molecular markers at approximately 3Mb marker intervals.

In order to perform a QTL screen one tries to determine the parameters of the

following relationship:

$$P_i = f(g_i) + e_i$$

Where P_i is the individual's phenotype g_i is the genotypes and e_i is an environmental contribution. Moreover, is is assume that covariance of genotype and environment are 0 (Broman and Sen, 2009)

$$\sigma^2_{phenotype} = \sigma^2_{genotype} + \sigma^2_{envir} + \cancelto{0}{2\sigma^2_{GxE}}$$

An appropriate additive model for mapping phenotype to genotype is described as such:

$$f_a(g_i) = \sum_{j \in QTL} \beta_j g_{ij} + \beta_0$$

where g_{ij} is marker j for individual i with genotypes values {0,1}

There is another assumption that typically if an offspring is getting half of their genes from the mother, and the other half from the father that the expected phenotype they will display is an average between the parental phenotypes (Broman and Sen, 2009).

$$E[f_a(g_i)] = \frac{f_a(p_{parent-1})}{2} + \frac{f_a(parent-2)}{2}$$

A test is performed at each marker site to determine whether the phenotypes of individuals that vary at the marker in genotype g_{ij} present phenotypes that also exist in two distinct populations with μ_0 and μ_1 akin to the parental distributions in figure 1.6. In order words, if there is a change in genotype at loci but no commensurate change in the phenotypes of indivudials with distint genotypes at the loci, there will be a low probability the trait is regulated by elements in the area(Broman and Sen, 2009).

$$LOD = log10 \prod_{i=1}^{N} \frac{P(p_i|g_{ij}, \mu_0, \mu_1, \sigma)}{P(p_i|\mu, \sigma)}$$

In order to reduce the multiple testing burdens of testing 1000s of genetic markers against a single phenotypic trait, we can try to compute the null distribution of the LOD scores rate by randomly permuting the genotypes 1000-10000 times and recomputing the LOD scores. For every marker, we can directly determine the determine the LOD score threshold above which there is only a 0.05% chance of a LOD score by chance (Broman and Sen, 2009). Once a strong QTL is discovered and the gene within the QTL region may be related to the phenotype, a congenic strain of with the complete set of alleles from a single parent strain, except for at this one loci mice is produced as a validation.

1.8 QTL vs. GWAS

QTL mapping and Genome-Wide Association Studies (GWAS) both attempt to determine relationships between genotypes and observed phenotypes. QTL mapping identifies regions of a genome that co-segregate with a treat in an F2 or recombinant inbred population. Using this techniques the key components of the cholesterol metabolism pathway in BXD mice have been dissected in the previous iteration of this study (Williams et al., 2016).

Despite this success, QTL mapping has limitations in the number of alleles that can be assayed at once and mapping resolution. When using QTL mapping in a recombinant inbred population, the total diversity of alleles which can be assayed is limited to only those alleles present in the founder strains. Moroever, if the founder strains have the same genotype in a region of the genome, it cannot be used for QTL mapping. Additionally, the resolution of the mapping is limited to the sizes

of recombination blocks on the chromosomes(Korte and Farlow, 2013). These can be reduced through extensive crossing but long lengths of the chromosome which have low recombination frequency are mapped as large QTLs containing many genes making it difficult to find causal genes for a trait within the QTL (Korte and Farlow, 2013). A solution to the first issue was proposed by the collaborative cross(CC) consortium in which 6 different stains of *Mus musculus*, *Mus domesticus* and 2 from other sub species of mice were crossed in order to produce a mouse stock with significantly higher genetic marker densities. (Collaborative Cross Consortium, 2012). The some of the CC mice strains however have too much genetic distance between eachother which has led to problems breeding the mice.

GWAS overcome the resolution limits of QTL analysis by using higher density markers such as SNPs instead of recombination intervals but run into multiple testing issue. Similar to QTL mapping, GWA studies test the association between a particular trait and a genetic marker across a large population of individuals. GWAS and QTL mapping provide hypothesis discovery and serve as the basis for further direction knock-out or mutation experiments to validate a genes function.

1.9 RI Mouse Population for GxE

Quantitative phenotypes are partially determined by GxE interactions. GxE interaction are the changes in a phenotypes as a result of the interplay between genetics with diet, age, differences in social interaction and exposure environmental stressors like heat, drugs, and pathogens (Williams and Williams, 2017). To accurately elucidate GxE in the context of QTL mapping, mice, and other isogenic model organisms are often used because a researcher can impose well-controlled perturbation across large cohorts of animals with similar genetic backgrounds (Williams and Williams, 2017). This enables the evaluation of complex environemental effects with good

power that would not be possible in wild animal or human populations (Williams and Williams, 2017).

As discussed in section 1.8 one of the most common disadvantages for RI strains is that they contain only a fraction of the known genetic polymorphisms in a species. The BXD family of mice used in this study segregates for a total of approximately 5.2 million sequence variants, which exclude 58% of common polymorphisms found among many standard inbred mouse strains (Williams and Williams, 2017). Despite this low coverage of the total polymorphism space in mice, a study using 150 stains of BXDs, a cohort of mice containing over 12000 potentially harmful missense mutations yielded a wealth of disease-relevant insights by performeing QTL mapping on thousands of phenotypes(Wang et al., 2016). This study not only shows the power of using this refrence population to determine regulatory sites for disease relavent phenotypes but is also a rich dataset to test hypotheses against. Eventhough the BXD mice are crossed for over 20 generateions, some stretches of the RI mice genome will be almost completely identical by descent (Wang et al., 2016). These chromosomal regions are known as blind spots for QTL mapping as they should not contribute to trait variance(Wang et al., 2016). In any given genetic interval, the collaborative cross mice, have much higher marker densities and smaller recombination blocks as compared to the BXD mice which have a sixfold lower polymorphism load. If a phenotype maps to an area in both CC and BXD mice, it is relatively simpler to determine the driving gene in the BXD as there is reduced the number of viable candidate genes in a QTL one has to consider (Williams and Williams, 2017).

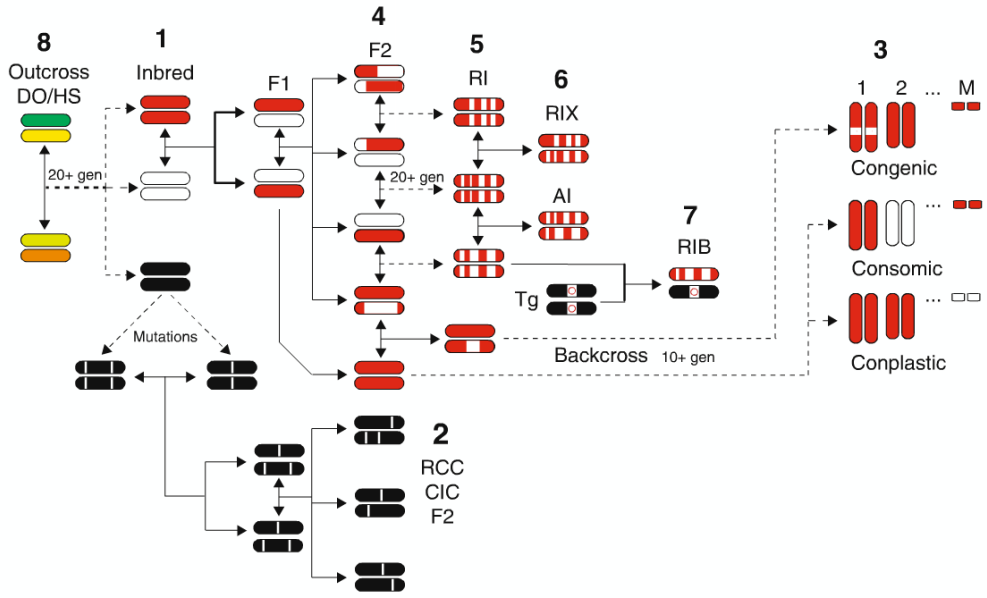


Figure 1.7: Adapted from:(Williams and Williams, 2017) Production of RI stock and validation of QTL results in congenic lines

1.10 Study Design

1.10.1 Model Mouse Population: BXD Mice

The BXD mouse stock is generated from the crossing of C57BL/6J(B) and DBA/2J(D) mice. With repeated inbreeding of the offspring of particular F2 parents for 20 generations, distinct inbred strains can be developed. During this random and selective mating, recombinations events accumulate resulting in a thoroughly dispersed genomes in the mice resultant of the two parental breeds. Each strain is a distinct combination of the parental genomes which allows the construction of a matrix of allelic origins for each stretch of the strain genomes. Since each of the resultant strains accumulates relatively little spontaneous mutations in coding regions, it is sufficient to genotype the parents and reconstructed for the progeny. It is however an advantage that all BXD strains have been genotypes with the data available in

several databases online. In a study, many individuals of each BXD strain are used in order to have stability measured phenotypes which can be used for further QTL analysis (Williams and Williams, 2017).

1.10.2 Mouse Diets: Components of Chow and High Fat Diet

Diet is the major environmental factor controlled for between mouse cohorts in this study. The regular chow (CD) and high-fat(HF) diets exert significant separate and independent effects on the measured metabolites, proteins and transcripts and have to be controlled for in order to accurately map QTLs and evaluate GxE interactions (Warden and Fisler, 2008). Unfortunately, there is a variation in the composition of the feed between batches from the supplier and must be taken to account when compared metabolite data from different mouse studies.

The chow diet used for the CD mice is Teklad. It is composed of agricultural products that represent a normal diet for mice and serves as a proxy for a healthy diet control in the mouse model population. It includes types of corn, wheat, oats and other grains and high fiber content from a range of vegetable sources containing complex carbohydrates(Warden and Fisler, 2008). Additional protein and fats are added to the pellets from sources such as fish and vegetable oil (Warden and Fisler, 2008). In contrast, the defined high-fat diet, serves as a proxy for an unhealthy western diet in the model mouse population. The diet consists of amino acid supplemented casein as it major protein source, cornstarch, maltodextrin or sucrose as the major carbohydrate components and fats from lard or soybean oil. Due to the lack of complex carbohydrates in the formulation, cellulose is added as a fiber source. Both diets are fortified with vitamins and minerals to ensure is mice in either cohort have the baseline requirements for proper functioning metabolism (Warden and Fisler, 2008).

1.10.3 Mouse Sex: Male and Female Mouse considerations

In a previous large BXD mouse experiment undertaken in the Aebersold, Auwerx and Zamboni lab, physiological differnces between the two sexes of mice were investigated (Andreux et al., 2012). Many variant traits are found but in subsequent studies only male mice were used order maintain statistical power. The expansive literature of mice physiology is biased. Male mice in metabolic studies as they do not go through estrous cycles that can change their metabolic profiles(Zucker and Beery, 2010) and female mice however, analogous to female human have longer life span on average as more often used in studies pertinent to identifying anti-aging mechanisms (Yuan et al., 2011). In order to maximize the number of novel aging related features detected in this study, a majority female mice population are used with a small male contingent for ensuring, the factors found are not largely sex driven.

Among the differences between the sexes in mice, one of the most pronounced contrasts to humans is the rapid estrous cycle mice experience. The reproductive cycle in humans, known as the menstrual cycle, lasts approximatelyu one month. Analogous cycling in hormones and utertine phsyiolgy occurs every 4 to 5 days in mice and is known as the estrious cycle (Caligioni, 2009). Although these short cycles make mice ideal candidates for studying changes during reproductive cycles, they also present a complicating factor in assessing sterols and cyclic metabolites in metabolomic screens (Zucker and Beery, 2010). Estrous cycle data (ie. ovulation or cycle phase) is not included in the phenotypic observation of the mice in this study and as a result, may not be reliably excluded.

An additional factor for using only female mice in this study are the large housing costs of keeping male mice in individual cages due to violent dominance behaviors. Male mice are extremely territorial and will even fight identical twins to the death

if housed in the same cage. To keep males separates, however, has dubious animal welfare implications as lack of social interactions shortens the lifespans and increases stress levels of these social animals (Peter Kelmenson, 2015). Barbering is another form of dominance behavior in which mice nibble at or pluck out whiskers and fur from their cage mates or themselves leaving large visible bald spots punctuated over the body of the animal (Peter Kelmenson, 2015). Although the behavior is observed in both males and female mice, it particularly common in female mice closely derived from C57BL/6 (B6) and A2G related strains (Kalueff et al., 2006) Both of these behaviors are a challenge of the technicians at mouse facilities, however barbering is preferential to male mice killing each other.

1.11 Why Multiple Omics

Large-scale inititatives to develop personlized medicine as standard therapy has driven a massive expansion in the number of human genomes that have been sequenced (Telenti et al., 2016; Fakhro et al., 2016). With tens of thousands of human genomes with deep coverage (Telenti et al., 2016) and significant additional SNP variants mapping databases becoming available to public researchers, GWAS have become an important tool for evaluating the association between common genetic variants and risk of disease. Through signification efforts of the scientific community, thousands of disease associated SNPs have also been found(Johnson and O'Donnell, 2009). The results of many GWAS, however, are seldom followed up with investigations that probe biological mechanisms underlying the found associations and can explain a limited amount of heritability. Additionally, most SNP variants are associated with small increased risk probabilities of disease and thus have weak predictive value by themselves without additonal insight about the underlying biological mechanims.

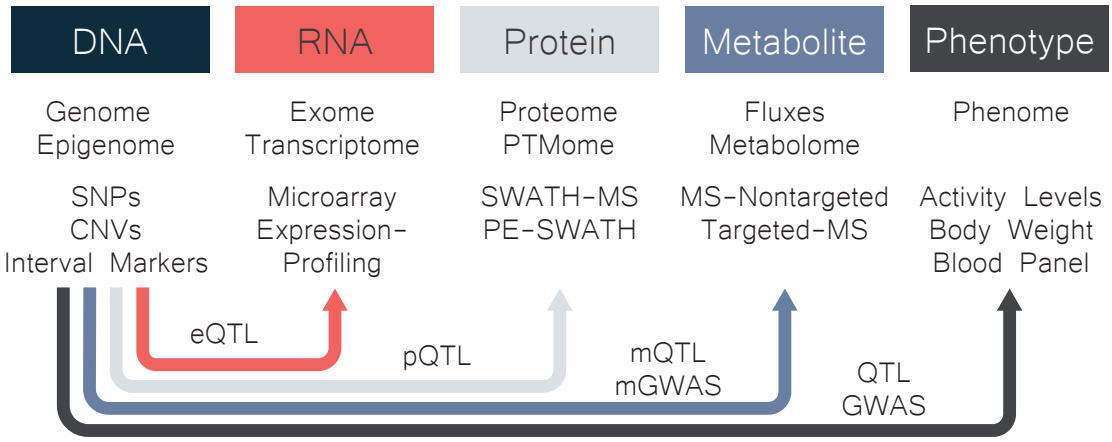


Figure 1.8: Adapted from (Dumas, 2012) Biological information that can be determined from performing high-throughput Omics analysis on many biomolecule read outs

With maturation in other high-throughput omics technologies (Figure 1.8), orthogonal approaches can be used for mechanistic investigation of disease associations. Integration across different genomics, proteomics, transcriptomics and metabolomics can comprehensively determine the relationships between genes, protein expression, metabolite concentrations, and their phenotypic repercussions as alluded to in figure 1.4. As observed in the pervious BXD mouse study done by Williams et al. (2016), the numbers of transcripts, thier cognate protein and associated metabolites are only modestll correlation with each other. All three can also be further modified by enzymatic processes making good experimental design and optimization necessary to increase the resolving power of these investigations (Johnson et al., 2016).

In order for us not to be overwhelmed by the large amount of data being collected, analysis will take place in a sequential layer by layer manner. First, the metabolite data is mined for QTLs and significantly enriched metabolites in the diet and age cohorts. From these metabolites we can determine which pathways maybe be affected by age, diet or a combination of two. Key enzymes that interact with the metabolites are preliminary hypotheses tested for differential expression at the proteomics

level. If metabolite concentrations do not faithful correlate with proteins, mRNA data can be used. For example, if a accumulation of a metabolite is seen in a certain cohort of mice but there are no changes in the upstream enzyme concentration, RNA and genome data can lend an insight into whether there is a mutation in a coding frame withing the interaction enzymes and if there are activation effects of the high metabolites concentrations back on the RNA causing affecting the expression of the transcript. Ideally, structural learning techniques such as lingam(Shimizu, 2014) can be used to validate the causal structures between all of the layers once a few causal network hypotheses have been generated.

Chapter 2

Metabolomics

2.1 Introduction to Non-Targeted Metabolomics

The metabolism is the sum of all the biochemical reactions that occur in an organism. Metabolites are shuttled through multiple enzymatically catalyzed reactions in the fundamental processes of energy production, growth and integrate multiple levels of information about the environment and cellular state in its kinetics and regulation mechanisms. The flux and relative concentration of these metabolites can be studied using multiple targeted and non-targeted techniques(Aksenov et al., 2017).

Unlike most common clinical assays, which are still photometric, the most common analytical techniques used to identify the global consortium of metabolites present in a cells or tissues are NMR and mass spectrometry based (Stanford Medicine, 2017). The goal of non-targeted metabolomics is to detect and quantify as many metabolites from a single extracted sample as possible without prior knowledge of the specific composition of the metabolome. Although, chromatographic separation steps can be used to differentiate between a specific set of compounds inseparable with a single mass spectrometry step, due to analytical limitations, a single analysis

cannot be used to cover the total metabolome which may consist of thousands of molecules with highly variable physiochemical properties. As a result, a decision must be made to determine the most pertinent class of compounds to best address the biological question and a sample is processed according to a protocol optimized for that small portion of the metabolome. Normally protocols for quantifying small polar molecules capture most of the central metabolism and provide an overview of metabolism but cannot resolve ratios between many enantiomeric and diastereomeric compounds like sugar or lipids (FGCZ, 2017). As a result, metabolomics, unlike genomics or proteomics is subdivided into many disciplines such as small polar molecule metabolomics, lipidomics, and glycomics focusing in different chemical classes of compounds using similar analytic techniques (FGCZ, 2017).

Although the types of metabolites recovered in different experiments may be different, the result is always complex data that require signal analysis and cheminformatics tools to process the raw signal from the mass spectrometer into peaks that can be assigned a normalized intensity and a chemical annotation. Bioinformatics and statisticaly analysis tools are then required to determine the correlation between metabolites in the given modules of the metabolites and to examine the connectivity of these metabolic pathways in the context of a phenotype or a process that may be driving a disease(Aksenov et al., 2017).

In contrast to untargeted mass spectrometry, targeted metabolomics operate on prior knowledge and hypotheses and quantify molecules of specific properties and mass ranges. The chromatography step before the mass spectrometry is optimized for extracting and separating the target metabolites and pathways such as the hexose sugars in glycolosis or lipids with equals masses but differences in the locations of their unsaturations (Cani et al., 2009). Targeted analysis can therefore be used as a follow up to untargeted metabolomics in order to validate hypotheses or quantify specific functional isobars(molecules with distinct properties but identical mass to

charge ratios) such as enantiomers or diasteriomers.

The way to perform flux analysis differs on the class of molecule, but generally, a isotopically labelled carbon, nitrogen or oxygen within a metabolite can shed light on which pathways that metabolites is shuttled into(Zamboni et al., 2009). Seeing an isotopic wight shift within pyruvate and different amino acids can quantify the utilization of glucose in anabolic and catabolic reactions. Ideally, one would use tracer compounds to directly quantify the fluxes of all metabolites in a multiplexed fashion, however this is unfeasible because it is technical difficulty and very expensive (Zamboni et al., 2009).

2.1.1 Metabolomics Methods

Metabolomics combines analytical chemistry and mass spectrometry platform technology, with sophisticated data analysis for deconvoluting dense MS1 broadband spectra. It involves the application of chemo and bioinformatics tools to profile the diverse metabolic complement of BXD mouse livers and put the results in biologically relevant context (Coen, 2010). Even with few standard tools and piplelines, metabolomics still offers a platform for the comparative analysis of metabolites that reflect the dynamic processes underlying cellular homeostasis(Aksenov et al., 2017).

MS-based metabolomics offers high selectivity and sensitivity for the identification and quantification of metabolites. Substantial progress in the speed and resolution of instruments and sampling processing techniques used in metabolomics has significantly broadened its analytical capabilities (Aksenov et al., 2017). Additionally, new orthogonal separation techniques like differential ion mobility spectrometry add seconds to the overall analysis while allowing the separation of a wide range of isobars (Domalain et al., 2014).In addition to the technical developments of metabolomics mass spectrometry, a range of chemoinformatic tools and databases that reduce

the complexity of data processing, handing(Xia et al., 2016), annotating signals for complex metabolites adducts and provide context to the metabolites within different biological pathways(Wishart et al., 2012; Xia and Wishart, 2010a). In the metabolomics studies described in this thesis, a high resolution TOF instrument is used, generating more than 17000 peaks in the mass analysis. Of these peaks around a 1500 can be annotated. A road-map of the full metabolite analysis pipeline is given on the next page.

2.2 Metabolite Extraction Protocol & Optimization

The experimental protocol for small polar molecule extraction and sample preparations from mouse liver is very simple in comparison to other metabolite classes and mass spectrometer analytes(Mushtaq et al., 2014; Haynes et al., 2009; Hu et al., 2009). The liver tissue is homogeneous, comprising mostly of hepatocytes which can be effectively ground to a powder in N_{2L}, which enables rapid extractions. The pulverized liver tissue is lysed with a combination of H_2O, MeOH and ACN after which homogenization and different extraction times can be used to extract the metabolites from the slurry.

To ensure the reproducibility of experiments chemical or enzymatic reactions that may occur during the tissue extraction must be minimized because these can drastically alter the original metabolite profile of the organism(Mushtaq et al., 2014). This rapid inactivation of all biochemical and enzymatic activity in organisms is known as quenching and is performed by a the MeOH and ACN in the extraction solution denaturing the proteins in the sample. In low-temperature extractions, long extraction times can be used without large changes from thermal degradation. Short extraction times in high-temperature extractions are used to prevent spontaneous chemical reactions.

Once the metabolites are extracted, the samples are evaporated in a low-pressure centrifuge and can be stored. Small molecules show much higher reaction kinetics than peptides and must be dried to a pellet and stored at -80° until they need to be resuspended and analyzed. On the day of analysis, the samples are resuspended at concentrations of $5\text{mg}\,\text{ml}^{-1}$into 96 well plates and loaded into the auto-sampler for randomized sampling into the mass spectrometer.

2.2.1 Optimization Objectives

Processing 600 samples in a single run is an extremely time consuming process and required at least 20mg of the precious mouse livers samples. Small scale pilot studies were used to determine whether the metabolite extraction protocol used in the previous paper(Williams et al., 2016) were reliable and reproducible 3 years after the experiments were done. Moreover, we had a limited amount of mouse liver, certain amounts of which had to be allocated for proteomics and transcriptomics and reproduction experiments should reviewers ask, thus it was imperative to conserve material.

Additionally, we wanted to tune the protocol to maximize the intensities and reproducibility of the metabolite data. Differential detection of key discriminant metabolites (such as N6-methyl lysine found in significantly higher quantities in chow diet mouse metabolites and has a strong known QTLs) which were observed in previous publications of the BXD mouse data was used to benchmark our current protocol (Williams et al., 2016; Wu et al., 2014). A timeline of all the metabolite studies is given below.

1. Pilot Study 1 - The first pilot study is used to determine the difference between hot and cold metabolite protocols and extraction time dependence on the metabolites intensities.

2. Pilot Study 2 - The second pilot study is used further refine the time schedule of the extractions and generate a standard curve with response factors for each metabolites.

3. Full Run 1 - The first full run with all 600 mice samples is done using the cold extraction protocol optimized in with first two pilot studies

4. Full Run 2 - follow up full run was performed due low number of features detected and sporadic jumps in the total ion counts in the first full run , yielding much higher coverage and lower CVs

2.3 Pilot Study 1

Four extraction conditions were tested, a hot extraction, two cold extraction and one homogenization free extraction, to determine which steps contributed to optimal extraction efficiency and robust metabolic coverage. The boxes outlined in red in figure 2.1 are the steps under scrutiny in this pilot study.

Hot Extraction:

In a hot extraction protocol (Figure 2.1), liver samples kept at -20° on dry ice are weighed into cell culture or falcon tubes. The tubes must be no longer than 10cm long to allow for the homogenization head to reach the bottom of the tube. After weighing, 0.5mL of extraction solution is added to the samples. The extraction solution is sufficiently membrane disrupting and dissolves cellular membranes freeing the metabolites.

A homogenization step using a laboratory-grade blender is included in the protocol to further lyse cells and break up particulates of protein and other non-polar cellular

debris that may crash out of the solution. During this process, viscous heating from the homogenizer brings the sample temperature up quickly. Thus the sample must be rapidly transferred into falcon tube with 3.5mL of extraction solution at 75° and timed for a minute. Although the original protocol calls for an 8ml final volume of extraction solution, 4ml of extraction solution was used because it would take less time to evaporate. To minimize degradation and restarting enzyme-free metabolism, timing is kept meticulously ensuring the sample does not have elongated exposure to high temperatures.

After a minute, the samples are removed from the hot bath and placed into a -20°C bath to quickly cool the tubes. At this temperature, it is assumed most metabolic reactions have stopped and can be kept at his temperature while the rest of the samples are being processed. The samples are then centrifuged to remove high molecular weight materials and the supernatant evaporated.The metabolite powder is stable at -80°C and can be resuspended later on the day of analysis.

Cold Extraction:

The cold extraction (Figure 2.2)is similar to the warm extraction. 4ml of a [40:40:20] solution of MeOH, ACN and H_2O is added directly to the sample to solubilize cells and poison enzymes. The samples are then homogenized and immediately put into a cold bath afterward to bring the temperature down to -20°. The tissue samples are then left to sit in the solution or in the fridge at -20°C to allow further extraction. Two cold extraction times were used, 1 hour and 24 hours at -20°CC to determine the extraction kinetics and optimal extraction times.

Homogenization-Free Cold Extraction:

In most analytical extraction homogenization is usually recommended to achieve an effective extraction(Mushtaq et al., 2014). As the complexity and resilience of the tissue increases, homogenization becomes a crucial step for breaks apart colloidal collections of cells, in the centers of which are not exposed to the extraction solvent. Moreover, metabolites are present in different compartments of cells, thus the disruption of those cells or their protective covering can maximize the extraction of metabolites. Liver tissue, however, contains a low diversity of cell type and is not difficult to break apart unlike muscle tissue. As a result, there may not no additional benefit to the metabolite extraction effectiveness in liver tissue with homogenization.

There are several techniques that can be used separately or in combination to homogenize samples. The most conventional, grinding the liver tissue manually with a mortar and pestle is performed on all of the tissues samples. However, in the protocol described by Williams et al. an additional homogenization step with a laboratory mill ball or laboratory-grade blender is called for after the cell lysis/quenching and extraction solution is added to the tissue. The lab-grade blender in the Aebersold lab can effectively liquefy tissue samples, however, the temperature of the samples is raised and cross-contamination occurs reduced the resolution power of low abundance metabolites. Although ultrasonication can also be considered as an alternative to high-speed mixers or strongly agitated ball-mills, the low numbers of sonicators in our lab precludes the use ultrasonication on 600 samples.

Already, leaning towards using the cold extraction due to its simplicity, we also performed a 24 hour cold extraction without the homogenization step to determine if it was necessary. The homogenizer step introduces impurities and cross-sample containments if the homogenization head is not properly cleaned and adds a minute to each protocol, complicating the timing of the protocol. Thus warranting us to

determine if it can be circumvented.

2.4 Pilot Study 1 Results

Hot and Cold Extraction Performance

The intensities from a sampling of metabolites across the 50-1000 Dalton m/z range is given in figure 2.3. In the third column, the Hot:H1 column compares the intensities seen in the standard hot extraction with the metabolites extracted using the cold extraction protocol with a 1 hour incubated at -20°C. In the lower mass range, the effect is minuscule between the group with significant differences appearing in lipids and cholesterol species that are difficult to extract quantitatively without the use of glass cuvettes that have a polar activated internal surface (Xia and Wishart, 2010b). Almost all of the highlighted metabolites in the volcano-plot (figure 2.4) are thus not of primary interest.

Time Dependence of Cold Extraction

Results shown in figure 2.5 show that both 2 and 24 hour cold extractions extract a higher number of metabolites that the 1 hours protocol. There is a higher variation in the extraction metabolites seen in the 2 hours extracting. One can conjecture this arises from insufficient extraction times, however, the original protocol used in (Williams et al., 2016) used 10-minute extraction, so coming to this conclusion is questionable.

Extraction Time Performance

Effect of Homogenization on Performance

To recall, the liver tissues are already pulverized in a mortar and pestle before all of the extractions. The effects of an additional lab-grade blender homogenization, as required in the Williams et al. protocol were tested in a 24 cold extraction done with and without the additional homogenization step. From the volcano plots (figure 2.6) , it can be seen that homogenization does not significantly increase the intensities of metabolite profiles. Surprisingly it actually reduces the metabolite intensities and increase the CV by 4% (not shown) in the cold extraction, although the source of this variability is not clear.

Pilot Study 1 Conclusions

as a result of the first pilot study, the homogenization step is taken out of the protocol. We also found that the hot extraction and cold extraction yield similar results, thus the hot extraction was supplanted by the cold extraction is it reduces the need for cumbersome heaters to maintain the high extraction extraction solution and extraction bath. Although we know longer extraction times give higher metabolites intensities, the optimal extraction schedule is still unclear as many significant metabolites found in the Science paper (Williams et al., 2016) were missing from the first optimization

2.5 Pilot Study 2

The second pilot study was aimed at tuning a range of parameters for the cold extraction and maximizing metabolites quantified and their intensities.

Pilot Study 2 Objectives

Biological replicates, of mice in the same age cohort, diet and strain were also included to determine which metabolites were most reproducible between biologically identical" mice. A freeze-thaw cycle experiment as performed in order to determine if there are any detrimental effects of freezing metabolites suspended in water. This was done to ensure the sample would be usable in case the samples would need to be frozen overnight and the extraction continued the next day due to unforeseen circumstances. lastly, separate blank samples containing ultrapure millipore water were included for every 10 samples on the 90 well plate for flushing the sample input lines in the mass spectrometer. It is thought this would decrease the baseline noise intensities in the mass spectrum profiles and reduce cross-contamination between samples.

Extraction Time Optimization

In the previous Pilot study 1, the cold extraction was chosen of the hot extraction as it extracted a similar number of metabolites, with higher intensity and had fewer processing steps. However, only 1 hour and 24 hour extraction times were examined across all 24 samples in the pilot study. In order to determine the optimal extraction kinetics for the cold, a time series experiment was conducted. The samples were prepared using the cold extraction methods and allowed to incubate from times between 1 minute and 3 days.

The results from the times series extraction experiment are shown in the figure 2.7. Individual traces from the extraction data can be found in Appendix 6.1. Although metabolite extraction results appear chaotic and randomly varying through multiple magnitudes of ion intensity, there is only a 1.2%-8.7% time-dependent increase between 10 mins and 1 day extraction times. Irrespective of molecule types, the 1 min extraction shows the largest CVs across the samples. This is to be expected as 1 minute is too short a time for a quantitative extraction. Accordingly the coefficients of variation decrease between the 1 hour and 1-day extractions from 22% to 16% however, the CV rise again in the 3 day extraction to 23%. Is this due to the accumulation of known thermal degradation production. Inosine is thermally decomposed to uric acid through Xanthine and Hypoxanthine intermediates (Fang et al., 2015). In the data, Xanthine decreases from an average intensity of 3×10^4 to 8.8×10^3 and concomitant increases in hypoxanthine from 7×10^4 to 2.3×10^5 from the 1 day to 3 days can be seen.

As a result, the 1-day extraction time is kept as the extraction time moving forward as it allows for a full day sample extraction on the followed by a full day of downstream preparation on the second. Moreover, the Hypoxanthine to Xanthine ratio alongside other known degradation pathways of amino-acids (Anton et al., 2015) found in our data can now also we used as a loose proxy to diagnose the degradation state of a sample.

Effect of Freeze-Thaw

Freeze-thaw cycles can be detrimental to sensitive biological samples such as proteomic sample preparations enriching for post-translational modifications (Paltiel et al., 2008). Although metabolites are small molecules it was not obvious whether the formation of water crystals that puncture cell membranes and damage protein has no effect on metabolites. Another motivation for investigating what happens

to the sample when frozen is, that if there is no detected effect, samples can be freeze-dried rather than evaporated in a heated vacuum centrifuge before long-term storage.

Figure 2.8 shows the results from 3 freeze-thaw cycles on metabolites profiles. Four mouse liver extracts suspended in water were placed in a -20°C dry ice H_2O:EtOH solution for freezing and thawing at 4° in the fridge three times. The differences between the two cohorts of metabolites are not significantly different (p-value 0.89), thus verifying the previously held hypothesis that the small molecules extracted in this experiment are not affected by the freeze-thaw cycles. A minority of molecules, including benzene, ETOH and MeOH were found in trace amount in the samples that were frozen, as seen from the outliers highlighted in red. There are thought to have come from impurities from splashing reagents during sample transfer.

Quantification of Cross-Contamination

In the first pilot study, one of the HF mice showed unusually high levels of metabolites such as N6-MethylLysine which are normally only enriched in CD mice, due to cross contamination. In the previous pilot study, there was only one blank (ultra pure water) well for all of the samples on the plates. The total ion intensities for that mouse were lower compared to the other samples and so the relative effects of cross contamination were magnified.

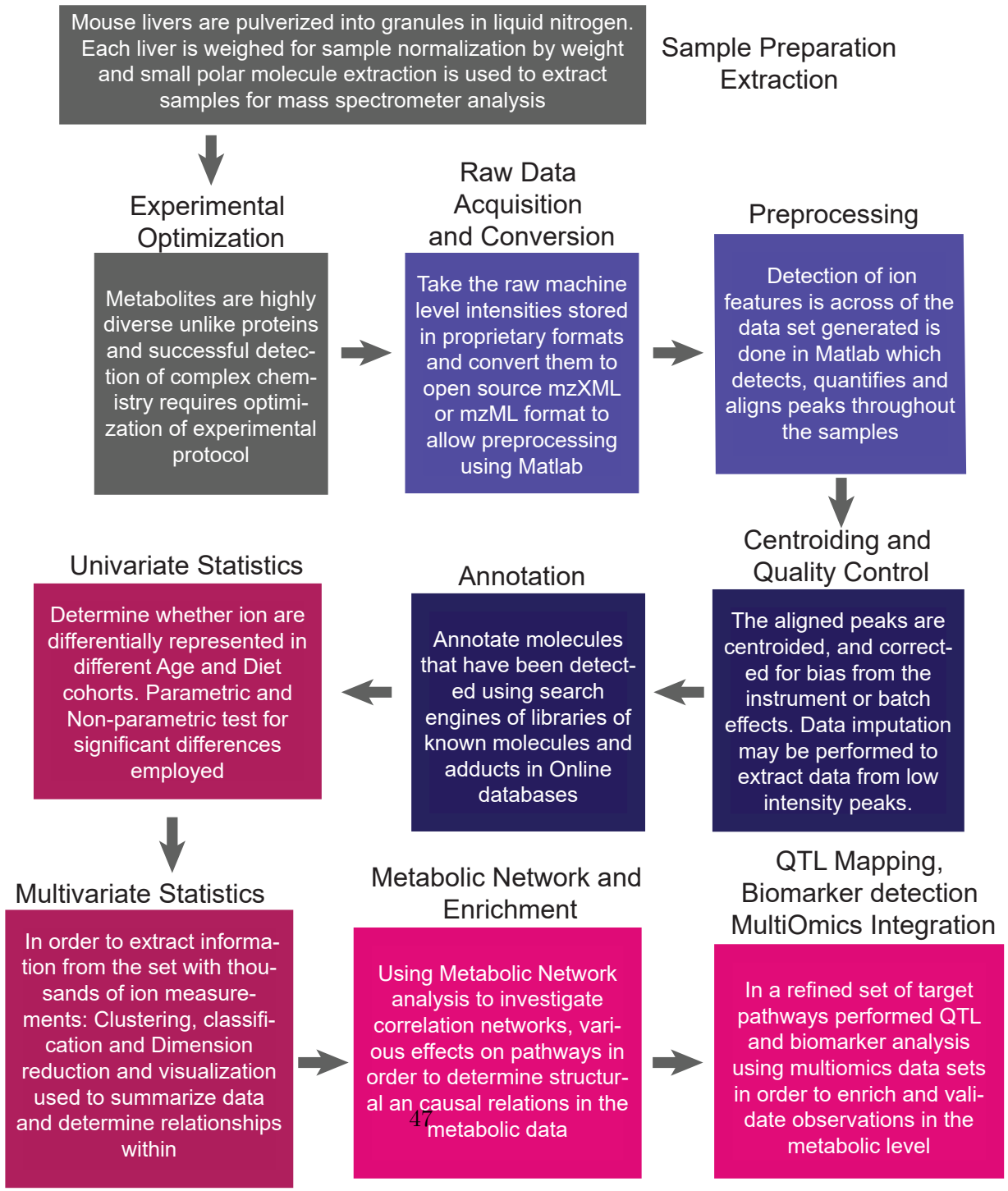
In order to prevent this, two wells of millipore water are added for every row of 10 samples on the 96 well plate. Figure 2.9 shows the intensities of 100 metabolites colored by strain on the top and the same metabolites, plotted for the blank samples. From this plot one can determine that on average the blank wells, even with many interspersed through out the plate have 2-5% of the liver samples due to cross contamination in the sample loading capillary. Thus 4-10%, two times the intensity

of the metabolites seen in the "blank" wells is the inherent limit of discrimination between two samples. As a result minute fold changes cannot be differentiated from contamination precluding analysis of many reaction ratios that are regulated very close to equilibrium at equal concentrations close to the signal intensity baseline (van Eunen et al., 2010; Traut, 1994; Beard et al., 2002; Schellenberger et al., 2010).

Conclusions from Pilot Studies

At the end of the protocol optimization, the 24 hour cold extraction was determined to be the best option for the full run of 631 samples because a large number of metabolites could be quantified at high intensity and the extraction would be easy to plan around. Even if the sample were not extracted at exactly 24 hours, the relative effect of time on the sample intensities is reduced between 8 hours and 24 hours as compared to 10 minutes and 1 hour extraction time regimes where timing is more crucial. We found out the effect of secondary, homogenization was minimal, similar to the effect of freeze-cycles on the metabolites. Unfortunately a large cross-over contamination between samples was also discovered. Once the protocol was optimization the full run with all 631 samples were done. The methods of extracting and aligning the metabolites features from the mass spectra of all 631 samples is detailed in next section.

Summary of Metabolic Data Acquisition & Analysis Timeline



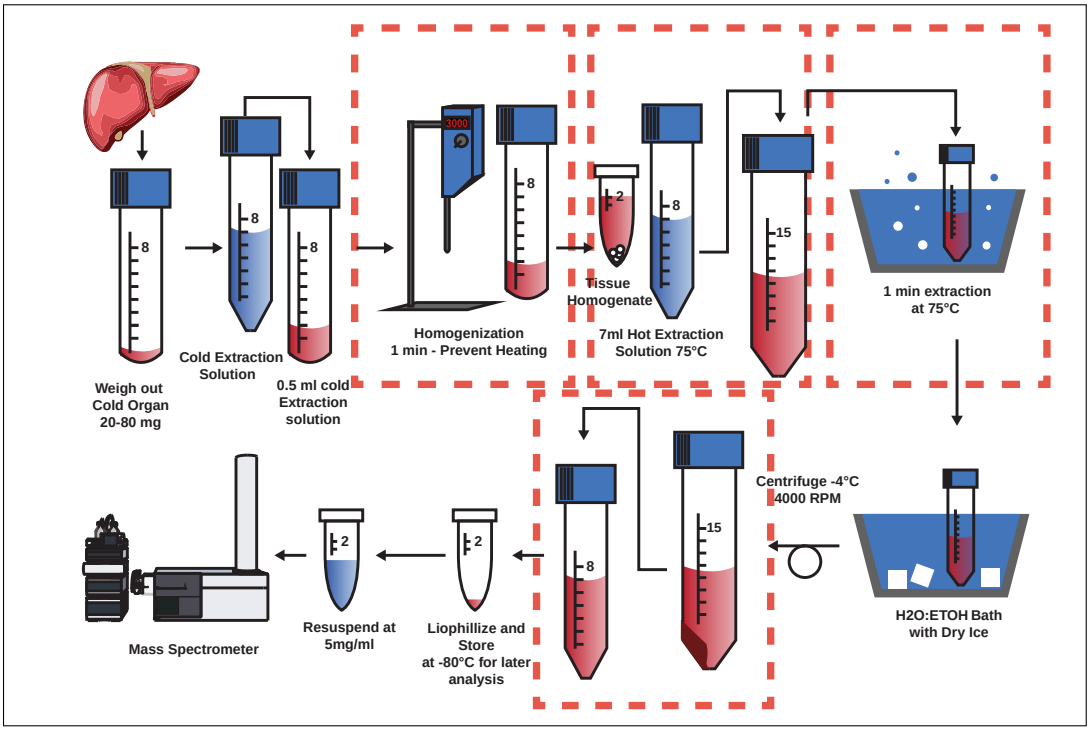


Figure 2.1: Hot Polar Metabolite Extraction Protocol

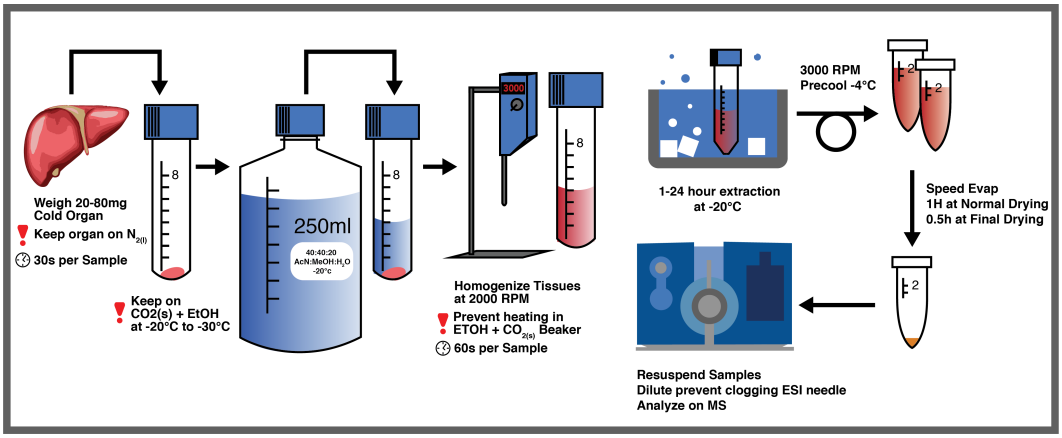


Figure 2.2: Cold Polar Metabolite Extraction Protocol

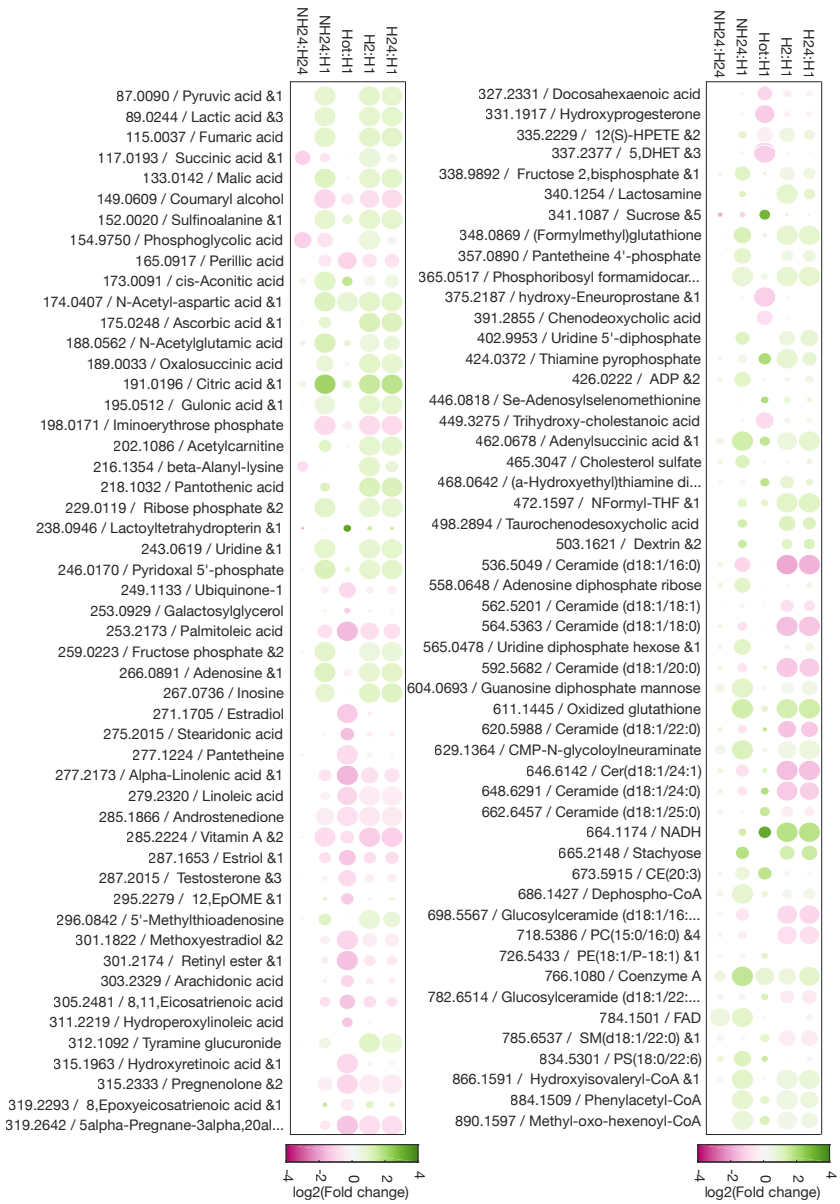


Figure 2.3: Metabolites Log_2 fold changes between different extraction conditions. Hot - indicates the hot extraction used in the Science paper, H1 in the cold extraction with a 1 hour extraction time, H24 also the cold extraction but with a 24 hour extraction time, NH24 is the cold extraction with 24-hour extraction without the homogenization with the lab-grade blender

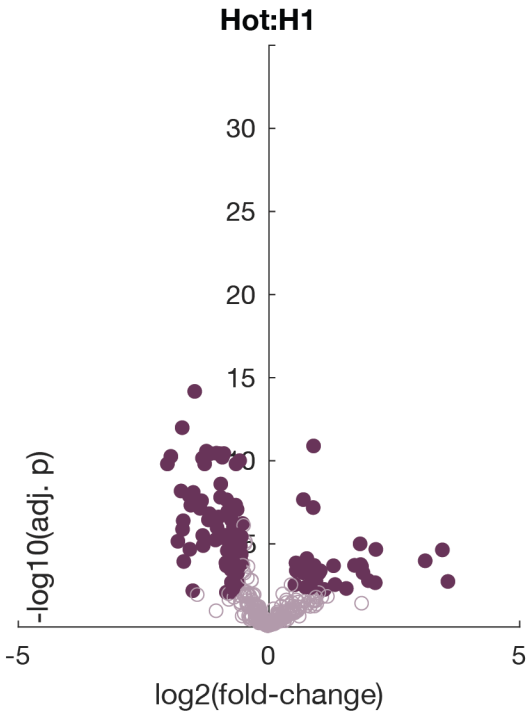


Figure 2.4: Volcano Plot of P-Values and Log2 Fold Changes seen between Hot Extraction protocol and Standard Cold Extraction Protocol

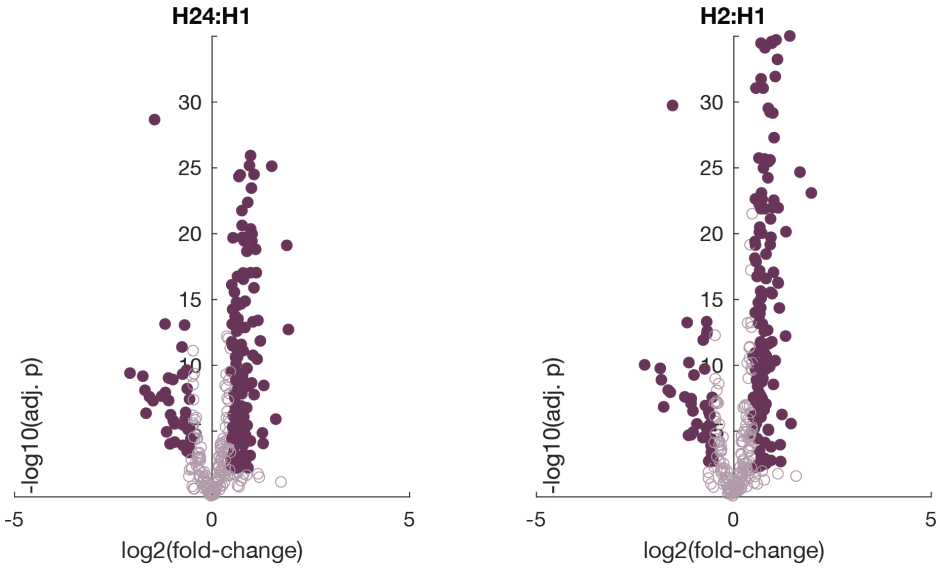


Figure 2.5: Volcano Plot of P-Values and Log2 Fold Changes seen between Standard Cold Extraction Protocols with an additional 1hour and 24 hour extraction period as compared to the standard cold extraction protocol

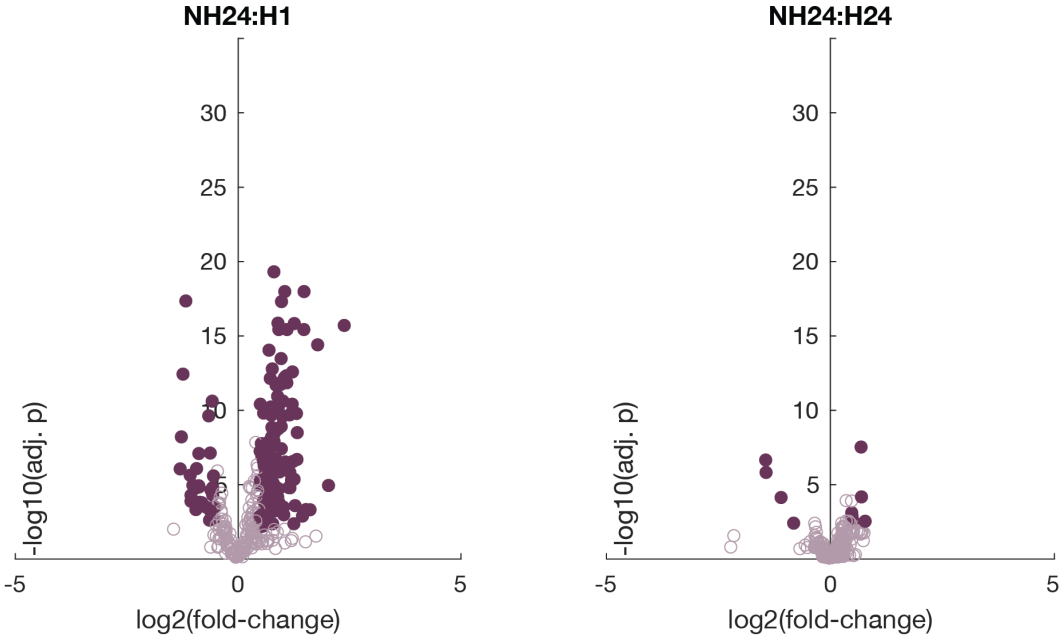


Figure 2.6: Volcano Plot of P-Values and Log2 Fold Changes seen between Standard Cold Extraction Protocols with an additional 1hour and 24 hour extraction perdion as compared to the standard cold extraction protocol

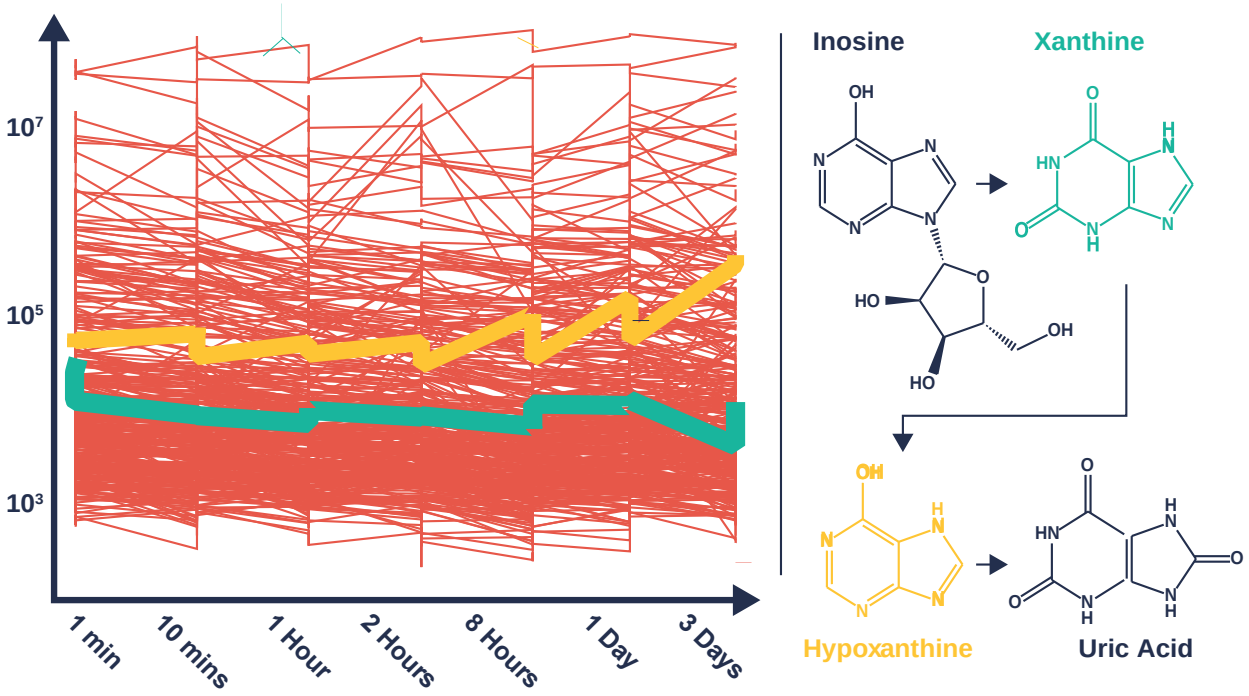


Figure 2.7: **Left** Time series intensities of all metabolites detected in pilot study 2 at different time intervals. In **yellow**, intensities for Hypoxanthine, a thermal degradation product of Xanthine can be seen increasing with time and in **green** the intensities of Xanthine is decreasing with respect to time due to thermal degradation **Right** Thermal Degradation pathway of Inosine to Uric acid as described in (Fang et al., 2015)

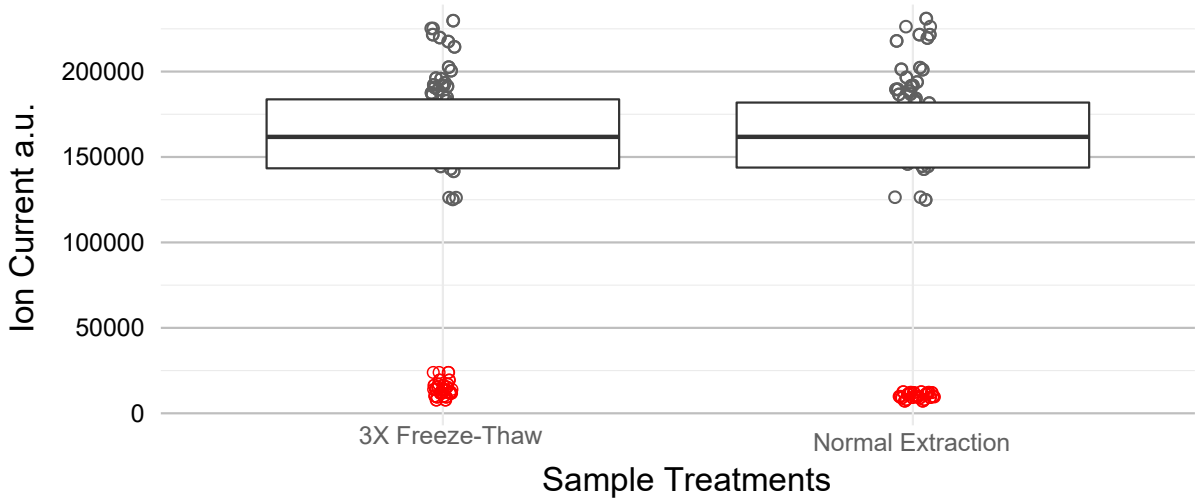


Figure 2.8: The effect of 3 freeze-thaw cycles on metabolites. Rep 1 samples were extracted for 1 day using the cold extraction protocol. Rep 2 samples were also extracted using the 1 day cold extraction protocol but also frozen and allowed to thaw 3 times

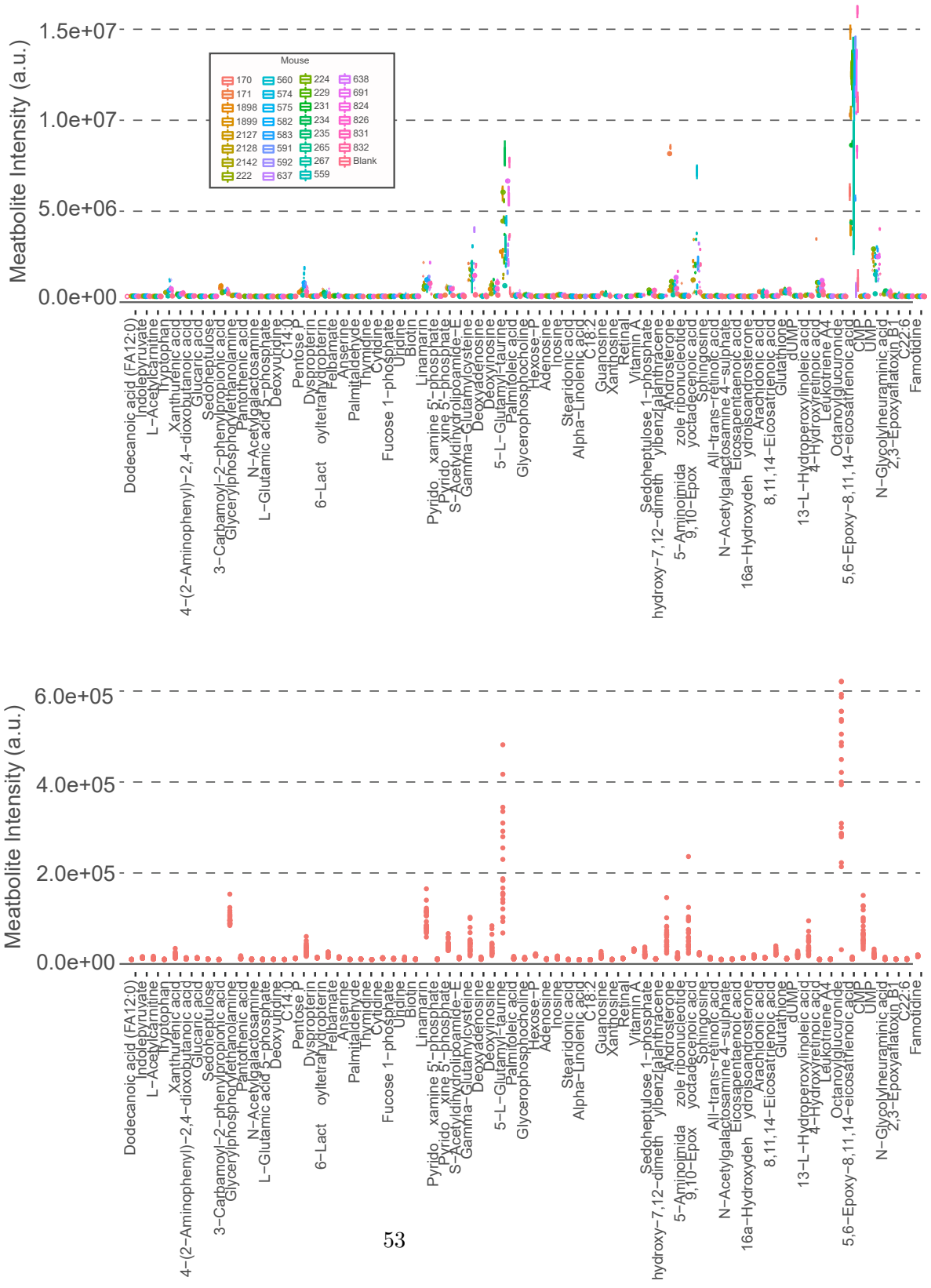


Figure 2.9: **Top:** Metabolite Intensities from all the mice in the second pilot study
Bottom: Metabolite intensities from all of the wells that were only filled with water

2.6 Metabolite Data Acquisition and Handling

The two categories of data processing for metabolomics can be divided into the processing the raw intensity data from the mass spectrometer and high level analysis ,interpretation and contextualizing of the data in biological networks. Raw data generated by the mass spectrometer is processed using a Matlab script called FIAminer which takes the peaks from the raw profiles and outputs a csv matrix file containing all of the samples, metadata and metabolite annotations (Zamboni, Unpublished). The data analysis and interpretation is done in R in addition to the use of KEGG (Kanehisa et al., 2017) and MetaboAnalyst(Xia et al., 2016) for pathway analysis.

2.6.1 Raw Data Acquisition and Processing

The TOF-MS used in this study is the 6550 Quadrupole Time-of-Flight, shown in the figure 2.10. Complex metabolite solutions are electro-sprayed into the mass spectrometer without any liquid chromatography. Next, this ion gas traverses a small lateral distance through a quadrupole drift tube and enter into a large time of flight mass analyzer that projects out of the top of the machine. The raw intensities that are generated by the mass spectrometer in a study are a continuous sampling of the detector current at regular intervals. Once ions flying through the mass analyzer hit the detector surface (Glish and Vachet, 2003), they cause a cascade of electrons that spread in a wave propagating on the detector surface, which is seen as a peak in the ion current. This current when it is plotted continuously is known as the raw profile. The raw profiles from three different samples with multiples replicate can be seen in the bottom panels of figure 2.12.

Similar to optical instrumentation where lines may broaden with an incorrect optical

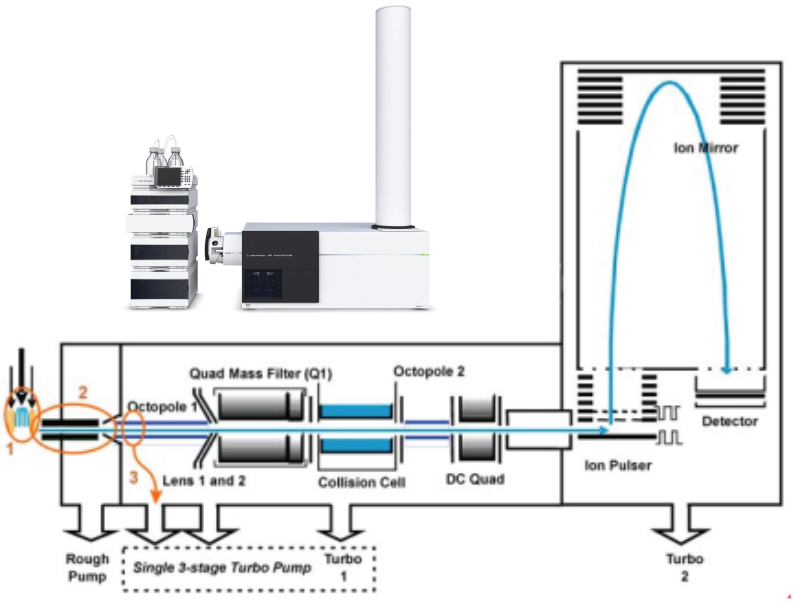


Figure 2.10: Adapted from (Agilent Technologies, 2017) of the 6500 TOF-MS with an included schematic of the flight path. The blue line int he schematic represents the path of the ion after they are softly ionized and electro sprayed into the instrument

arrangement and lens degradations, the widening of the peak shape in the profiles data is due to the dispersion of the ion packet or aberrations from stray electric fields in the mass spectrometer(Glish and Vachet, 2003). Along the flight path of the mass analyzer, many factors may lead to the broadening of a packet of metabolites flying towards the detector which must be considered for accurately determining a baseline intensity and centroiding the profile data. In TOF-MS instruments used in this experiment, the peak shapes show fast intensity spike with a gradual decay of the signal (figure 2.12.C). This is due to the arrival of highly abundant isotopically pure and lighter ions arriving at the detector first, followed by heavier isomers lower abundance ions.

The shape of the peaks in the profile generated by an instrument is a function of the sum of conditions and voltages on all of the ion optics and is known as the MS Tune. In the figure below, well-defined peaks can be seen in the lower M/Z range

highlighted in the boxes on the left side of the figure. The peaks are strong and can be easily aligned and averaged across all the samples and extracted in an automated manner using FIAminer. However, the signals in the high m/z regime is much lower intensity as seen in the centroided spectra shown in figure 2.1.A, and requires a high resolving power instrument to detect distinct peaks in the convoluted grass. Through a comprehensive calibration process in that involves not just m/z, as is the case for all conventional MS calibration, but also the peak shape, these smaller peak features can be extracted (Zamboni, Unpublished).

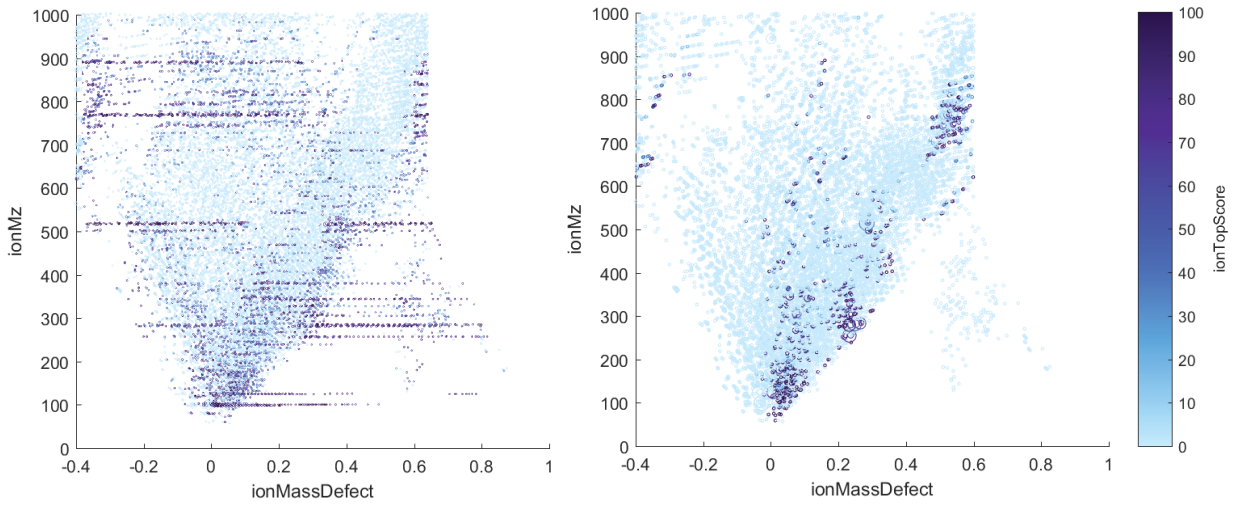


Figure 2.11: Left: All annotated peak annotated automatically. Right: Ringing peaks and impossible mass combination filtered

A final processing step removes artifacts that appear during the analog to digital signal conversion detector in which pre-amplifier electronic noise is recorded and appears as horizontal streaks in the figure above. These is known as ringing peaks(Goodner et al., 1998). Since they show up at m/z ratios that cannot be attributed to any sum of know elements or isotopic mixes, they are easily removed.

During the course of this peak assignment, each peak assignments is given a score and only perfect score peaks are used in the final analysis. In an average sample analysis 19 000 featured were detected, 1500 of which could be annotated with some

certainty when searched against the HMDB (Wishart et al., 2012), Metlin(Smith et al., 2005) and KEGG databases(Kanehisa et al., 2017). The majority of detected features however cannot be annotated and are cryptically known as dark matter in the mass-spectrometry community (Aksenov et al., 2017).

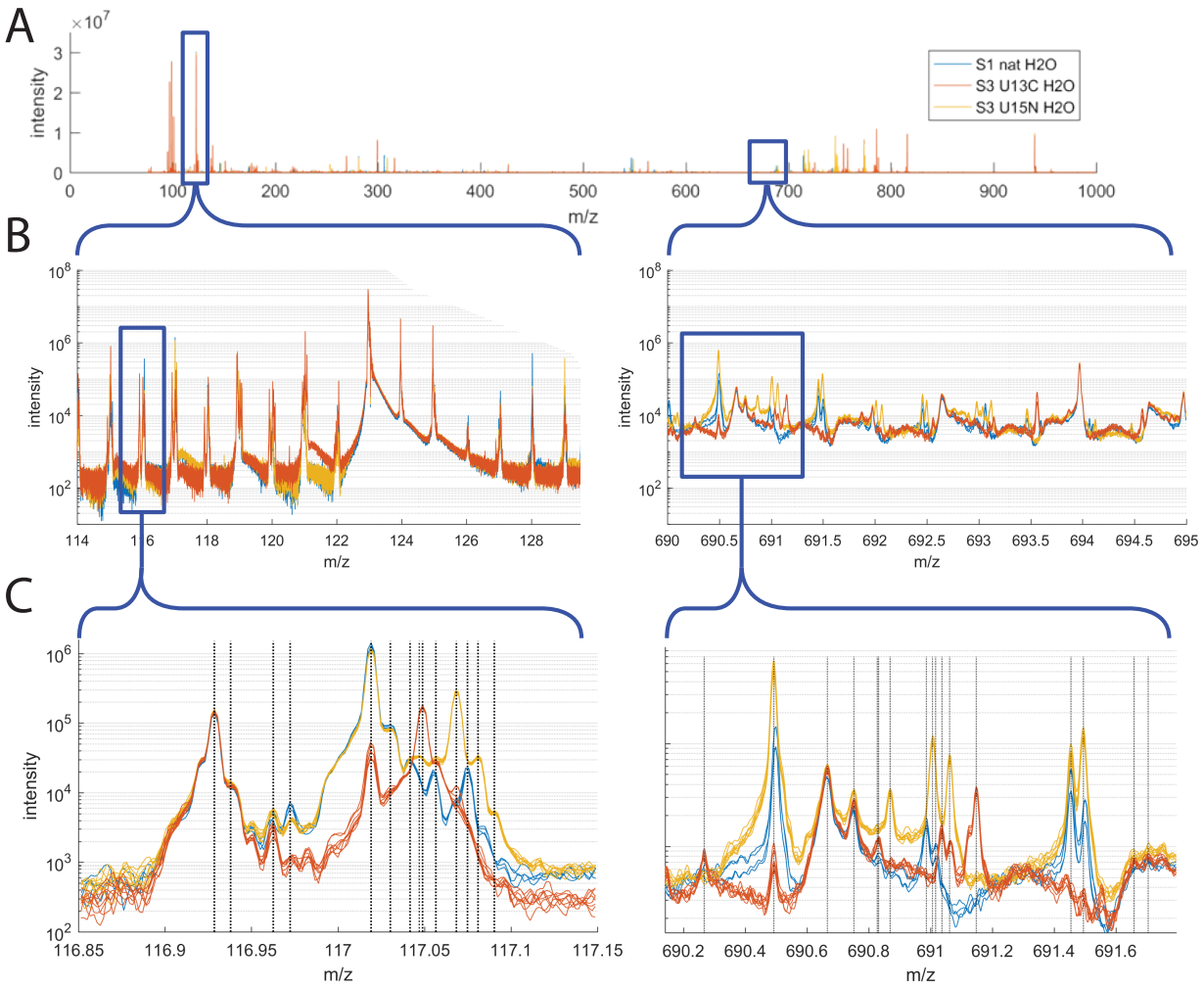


Figure 2.12: A. Centroid level data for the metabolite peaks from [0,1000] m/z
B. A zoomed-in view of the lower mass range around 100 m/z. The peaks in the lower mass range are well defined and can be fitted with the mixture a mixture of Gaussians to determine their maximum values. high m/z range peaks which are less well defined and required an additional operation to stabilize the baseline before peaks can be centroided. C. A further zoom into the spectra again shows the unstable characteristic of metabolites peaks in the high m/z ranges which pose a challenge to process in an automated manner.

2.7 Data Analysis

2.7.1 Normalization of Metabolite Data

In metabolomics, the physiochemical property range of the molecules is sufficiently large such that a small subset of metabolites, analogous to spike-in peptides at known concentrations in proteomics, cannot be used as an internal control for intensity normalization (Välikangas et al., 2016). This is because the intensity of a molecule is not a simple linear function of the metabolites concentration, but rather a convolution of ionization efficiency, fly-ability and voltages biases on the electronics in the mass spectrometer. Thus the only factor we can normalize the metabolites to are the originally weighed masses of the liver tissues, the metabolites were extracted from (Välikangas et al., 2016). This normalization, although crude yields highly normal data allowing us to perform statistical tests without breaking the normality assumption.

2.7.2 Quality Control

Once the data has been normalized, summary statistics of the full run data are generated. Firstly, a plot of the metabolite intensities overlaid over a map of all of the reaction in the metabolism of mice is used to show the metabolome coverage. The map of the coverage can be found in Appendix **??**. In both of the full metabolome runs of 621 mice livers, all the major KEGG pathway categories except drug and xenobiotic metabolism are covered. This is a good sign as the controlled diets did not contain common drugs or xenobiotics.

The number of metabolites in the first full run performed on all 621 mouse livers samples disappointingly yielded only a third of the metabolites annotations as com-

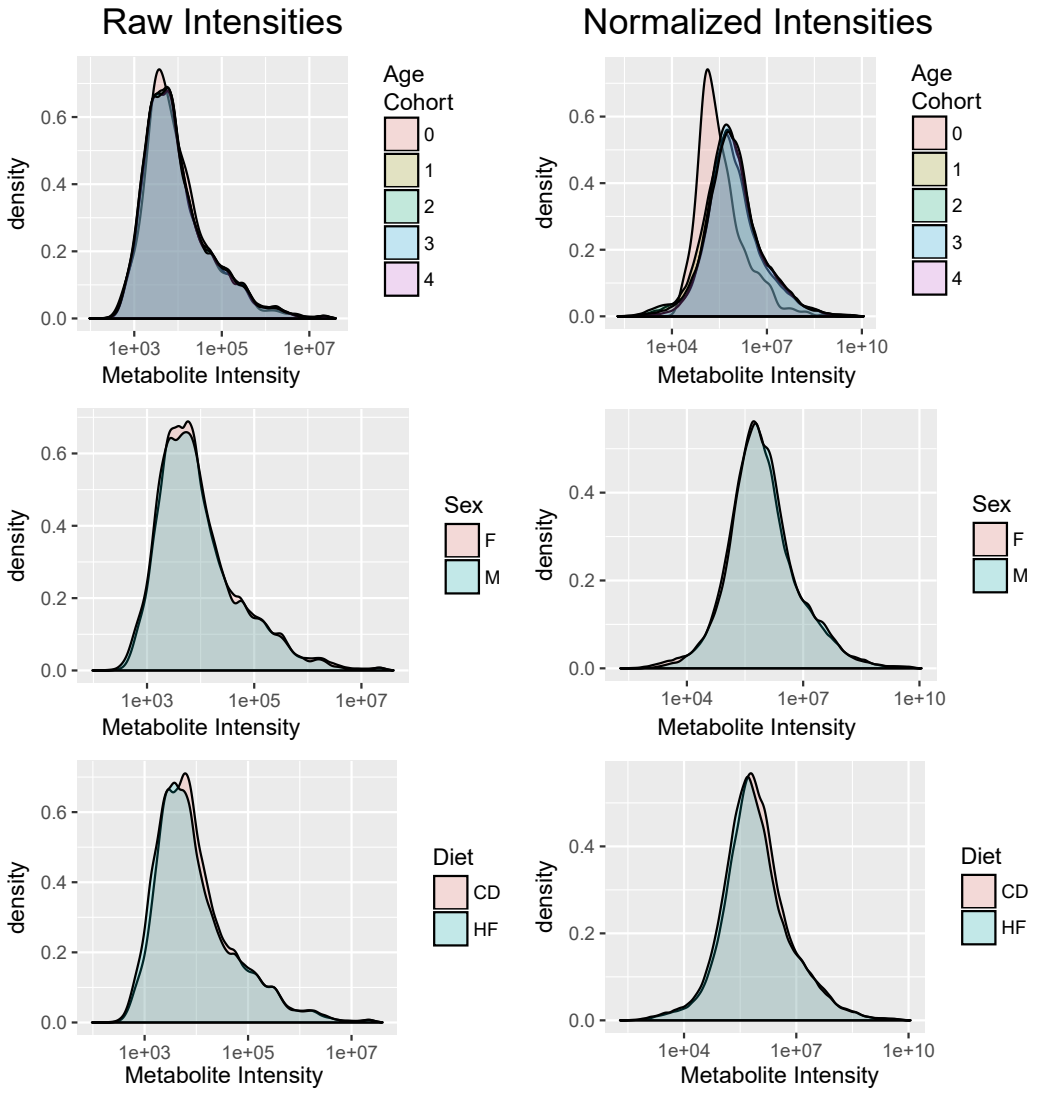


Figure 2.13: Metabolites intensities normalized to the amount of tissue is extracted from show normalized even when segregated into separate groups

pared to the Science paper. This was due to the logistical challenges of dealing with this many samples and issue with refrigerators containing the samples losing power, allowing them to heat up before analysis. Fortunately, the second run of all the metabolites was performed with improved planning and backup refrigerators. A much higher number of annotations could be made, with 475 metabolites

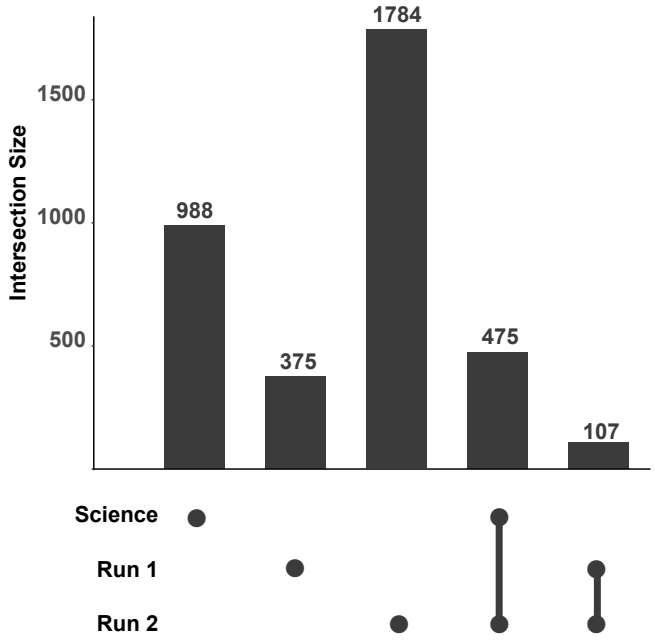


Figure 2.14: Number of metabolite annotations in the previous Science paper by Williams et al., and two full metabolome experiments conducted for this thesis. The second run, done with much higher care and attention to timing and temperatures during the preparation yields a much higher number of hits.

overlapping with the science paper, giving us a large set of data to cross-validate conclusions.

To get a summary view of ion intensities, the average total ion counts of the metabolites in the mouse and the blank samples are plotted together in figure 2.15. Here only the data from the first full run is shown. As seen previously, the mass spectrometer has a very large dynamic range and detects metabolites in over 4 orders of magnitudes. The on average the signal from the blanks is 2 order of magnitude smaller and cross-contamination signal in the full runs is only 2%-3% of the signal.

A bootstrapping sub-sampling method is used to generate the coefficient of variance(CV) between, injection replicates (samples that are sequentially injected one after the other), process replicates(samples that come from the same mouse liver

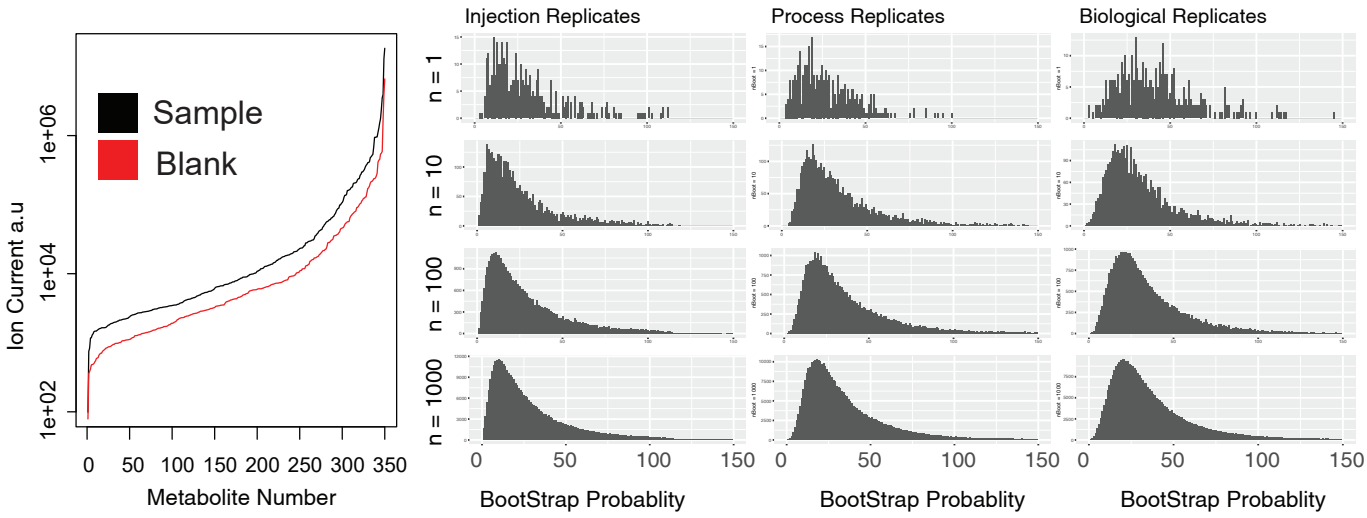


Figure 2.15: **Left:** The Ion intensities of different metabolites plotted from lowest to highest in the mouse samples (in black) and in the black well (in red) that had only water. The signal from the blanks is the result of cross-contamination.**Right:** Bootstrap CV calculations of injection replicates, process replicates and biological replicates

and are extracted using the same protocol) and biological replicates (samples that come from two different mice that are the same strain and within the same diet and age cohort). The bootstrap distribution of the injection replicates has the lower CV with a mean of 7% and a sharply defined mean. The process replicates and biological replicates have a slightly higher CVs at 8% and 12% which is within the precision of metabolomics in the previous paper, but slightly on the higher side. These CV values are replicated in both runs on the mouse livers samples but lower in the second full run.

Most high CVs are the result of metabolites that are not quantitatively extracted with the small polar molecule extraction protocol. Figure 2.16 shows the two classes of metabolites and their correlations between biological replicates. Small polar molecules such as oxoadipate and aconitate, are quantitatively extracted and show a robust correlation between biological replicates. Phospholipids, ceramides, cholesterols, and prostaglandins often are not well extracted using this protocol and have high CVs between biological replicates. In addition to being poor measurements,

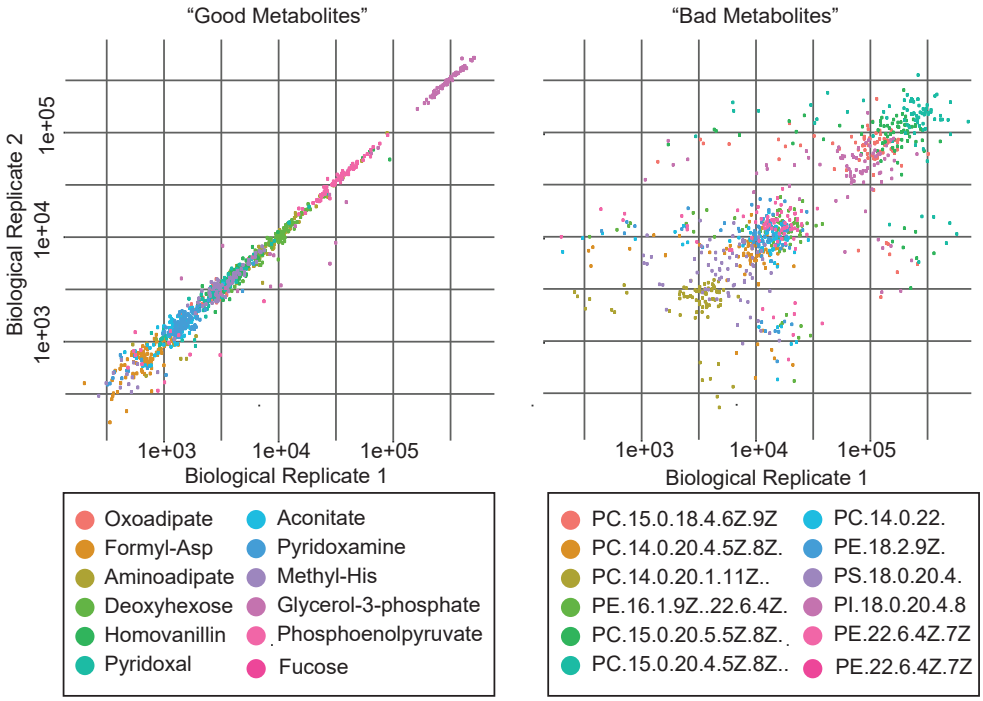


Figure 2.16: **left** Small polar molecules are extracted effectively form the liver samples and correlate strongly within biological replicates. **Right** Non-polar molecules are not well extracted using the extraction protocol this method and are unreliably extracted]

the chemical reaction that catalyzes the anabolism and catabolism of these lipid molecules are often cyclical, a single enzyme may oxidize fatty acids of multiple lengths. This complicates QTL analysis and does not allow us to pick a single affected enzyme, reactant, product trio in our network analysis. It is for these two reasons the fatty acids are excluded from QTL analysis.

Clustering is the last method used to quality control the shape of the data. Manhattan hierarchical clustering and Mankowski clustering is used. Ideally, injection replicates, bio-replicates and diet cohorts should cluster together, rather than samples that were extracted in the same batch being clustered together, indicating large batch effects. Hierarchical clustering, k-means and PAM based methods on the raw data may not be sensible when the variables are measured in many orders of magnitudes. Metabolites with the largest ranges get the most weight and might dominate

the analysis. As such, the variables can be are scaled and the standard intensities are used for the clustering analysis.

In figure 2.17 all the samples from Mouse 1481 are clustered together (shown in pink). In orange, all of the replicates for mouse 1270 are clusters together, however, there is a single run that had an error in the injection that is clustered with another mouse strain. The many biological replicates of C57BL/6J mice are all clustered with their injection replicates, however, the Minkowski distance is much for effective at ascribing these samples to a single cluster. Except for the 1 mouse in this subset, all the other mice are clustered with their injection replicates. Both the Manhattan and Minkowski distances yield the smallest averages distances between, injection replicates, strains, and diets respectively. The clusters of samples in different age cohorts are robustly determined, indicating the metabolites that differ between age cohorts of mice a do not show stark dichotomies as the metabolites that differ between the diet cohorts. As a results age, mice in the same age cohort on ages may not be clustered together using either metric. The cluster analysis yields a satisfactory result, and there is confidence in the data. Thus, more biologically orientated questions can be asked.

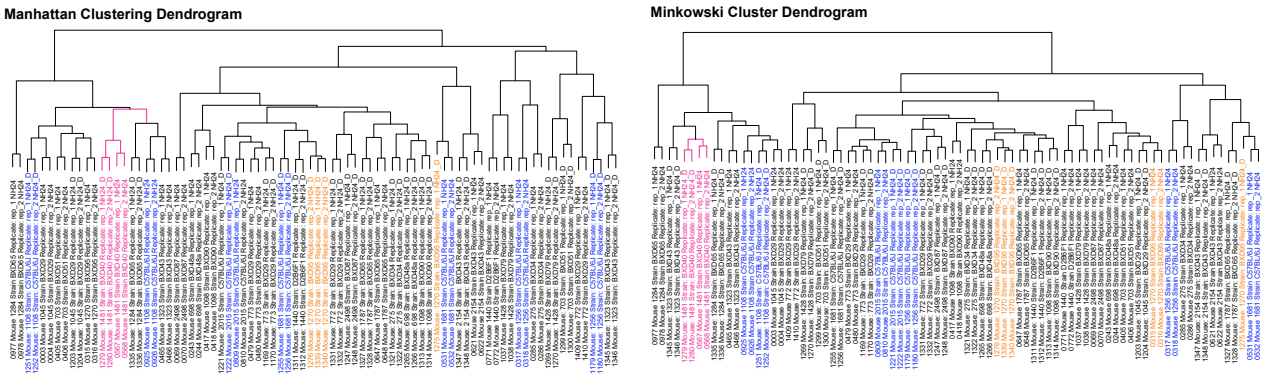


Figure 2.17: **Top:** Hierarchical agglomerative clustering using Manhattan distances. Below: Hierarchical cluster performed using the Minkowski distance which is more robust to noise. Biological and injection replicates are colored to qualitative assessment of the clustering result

2.8 Analysis of Metabolic Data

2.8.1 Metabolite Fold Changes

The purpose of performing this extensive metabolomic search to identify differences in the metabolomes of mice in different diet and age cohort. Metabolites that were robustly different between the HF and CD mice in the previous metabolic colony experiments are observed again in aging the colony but show lower fold change ratios and intensities. In the volcano plot (figure 2.18) on the left, allyl isothiocyanate, allysine, 2-oxoadipate, and fucose are among a large group of metabolites that have large fold changes in mice that are on a chow and high-fat diet.

The volcano plot for metabolites with significant fold-changes between old and young mice 2-oxoaditpate and fucose also accumulate in higher quantities in older mouse livers in addition to bilirubin and taurocholic acid derivatives. However, taurocholic acid is a cholesterol derivative and cannot be trusted to be reproducibility measurable. Molecules like bilirubin which are markers for jaundice can also be an interesting target because of the fact that as mice age, they seem to have a reduced ability to remove bilirubin from their livers. The physiological function and reason why these metabolites make good metabolic disease markers are given below.

AllyIsothiocyanate: AllyIsothiocyanate accumulates in the chow diet mice in both the old metabolic colony from (Williams et al., 2016) and current aging colony mice. AllyIsothiocyanate(AITC) is found in high quantities in cruciferous vegetables and is highly bioavailable once digested (Zlatkis and Liebich, 1971) It is a secondary metabolite produced in high quantities in certain plants like horseradish and mustard but can be found in a large range of vegetables that may be present in the chow diet. AITC also has a very strong QTL, however, the QTL region is sparse in terms of SNPs and genes, containing many Riken genes which do not allow us to uncover

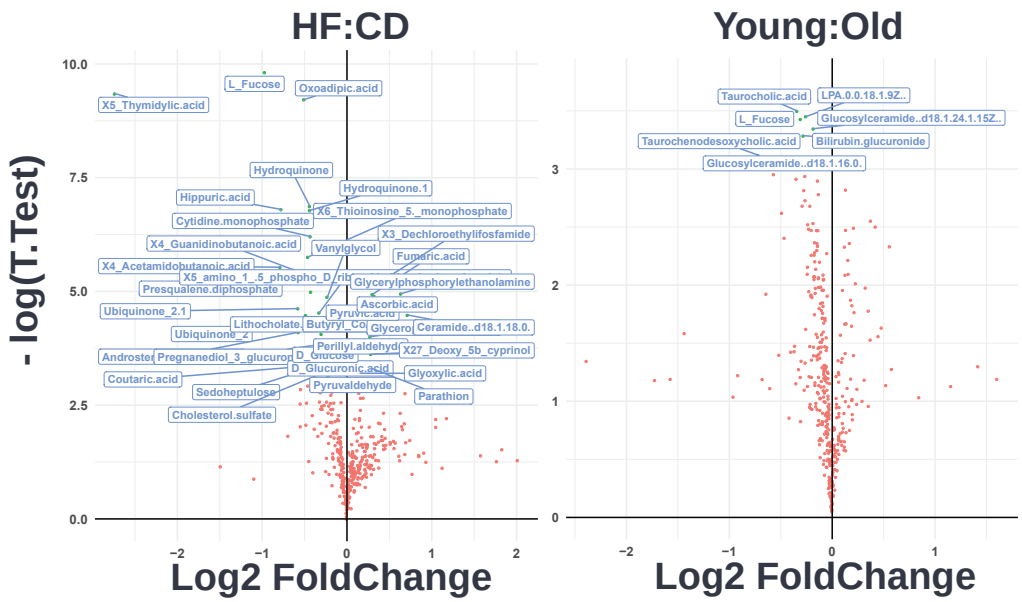


Figure 2.18: Volcano Plot of the fold changes between mice in different diets cohorts on the left and mice in different age cohorts on the right. The significance threshold is chosen at a t-statistic of 3 and an absolute log2 fold-change of 1

the underlying affective loci conclusively. Although the stress-related regulation of AITC in plants is mechanistically understood, it is not likely its production is similarly regulated in mammals(Kissen et al., 2016).

Aminoadipate & 2-Oxoadipate: Aminoadipate and 2-oxoadipate are both degradation products of lysine, tryptophan, and hydroxylysine produced in mammalian livers(Higashino et al., 1965). They are significantly enriched in HF diet mice and higher in older mice (p-value = 0.65). In humans, mutations in the Dehydrogenase E1 And Transketolase Domain Containing 1 (DHTKD1) leads to the accumulation of lysine and tryptophan degradation products and increases in detectable reactive oxygen species (ROS) (Goncalves et al., 2016). The DHTKD1 gene encodes a protein that catalyzes the conversion of 2-oxoadipate to glutaryl-CoA through a dehydrogenation reaction. The catalytic domain of DHTKD1 is structurally related to 2-oxoglutarate dehydrogenase and generates H_2O_2 and O_2^-, during oxidation of 2-oxoadipate which can cause age-related degradation and high-fat diet associated

inflammation (Goncalves et al., 2016).

Allysine: Allysine is a product of lysine oxidation which is increased in the aging skin and diabetes patients that do not have renal failure (Sell et al., 2007). In the maturation of fibular collagen in the skin LOX(Lysyl Oxidase), oxidatively deaminates the lysine residue generating allysine. The interruption of this process can lead to the accumulation fo allysine and deterioration in the collagen matrix of the skin resulting in wrinkles (Bailey et al., 1998). Additionally, the -amino group of lysine residues in proteins may undergo stochastic deamination by MCO (metal-catalyzed oxidation) reactions that produce allysine. This MCO catalyzed oxidation is also implicated in an age-related diseases such as Alzheimer's disease and metabolic diseases such as atherosclerosis and diabetes (Stadtman, 2004).

Fucose: Fucose containing glycan is a post-translational modification on a range of serum proteins and is itself present at low concentration in the serum. Serum proteins are produced in the liver and aberrant fucosylation on these proteins are the result of mutations in the many fucosyltransferases(Becker and Lowe, 2003). Proteins with incorrect fucose-containing glycoconjugates can create complications in a range of cellular processes and are a putative biomarker for many diseases including cancer, rheumatoid arthritis and diabetes (Wiese et al., 1997).

After validating known HF and CD metabolites from (Williams et al., 2016) such as allysine and fucose, volcano plots of metabolites foldchanges between the two diet cohorts are used to determine a wider number of metabolites to perform QTL analysis with. The HF:CD volcano plot shows significant changes in central energy metabolism such as pyruvate, cMP, fumarate, and ubiquinone are enriched in the high-fat diet which may not be surprising. Pyruvate is also an intermediate in the ketone bodies pathway that is activated when there is low glucose but high fat in the blood serum.

D-glucose is also a significant hit and although it can be assumed glucose is the major isomer constituent of the peak, a targeted chromatographic separation assay would have to be optimized and implemented in order to resolve all the hexoses and determine this for certain. Additional carbohydrate metabolites such sedoheptulose are also an interesting trait. It also has many enantiomeric centers meaning this peak is mostly like a composite of all the c7 sugars in the pentose pathway. This pathway is interesting as it has implications for the rate of nucleotide synthesis which may be in pertinent in aging, however, the need for targeted mass-spectrometry to validate the hit is not appealing. What is largely missing from these metabolites which are enriched and depleted in HF are the amino acids. These metabolites are intermediates between the glycolytic and TCA cycle metabolites which one would assume would be hugely affected pathways with large diet based interventions.

2.9 QTL Analysis

The distribution of the raw QTL LOD scores for all of the metabolites with significant age related and diet fold changes are shown in figure 2.20. In order to ensure the high scores were not simply generated by chance, the null distribution for the QTL LOD scores are computed. The genotyping array with which the QTL analysis is performed is scrambled after which the LOD score for the metabolites are recomputed. If the metabolite still shows an association with regions of the chromosome that has been completely scrambled, this means the association may have occurred by chance. A majority of the real QTL have large overlapping segments with the LOD scores generated with the scrambled genotypes array indicating a need for a high LOD score threshold of 5-6 to be confident in a QTL. Thus, one should be careful in interpreting a QTL are causal if it exists in segments of the chromosome that has the proclivity to show high QTL LOG Scores even after scrambling.

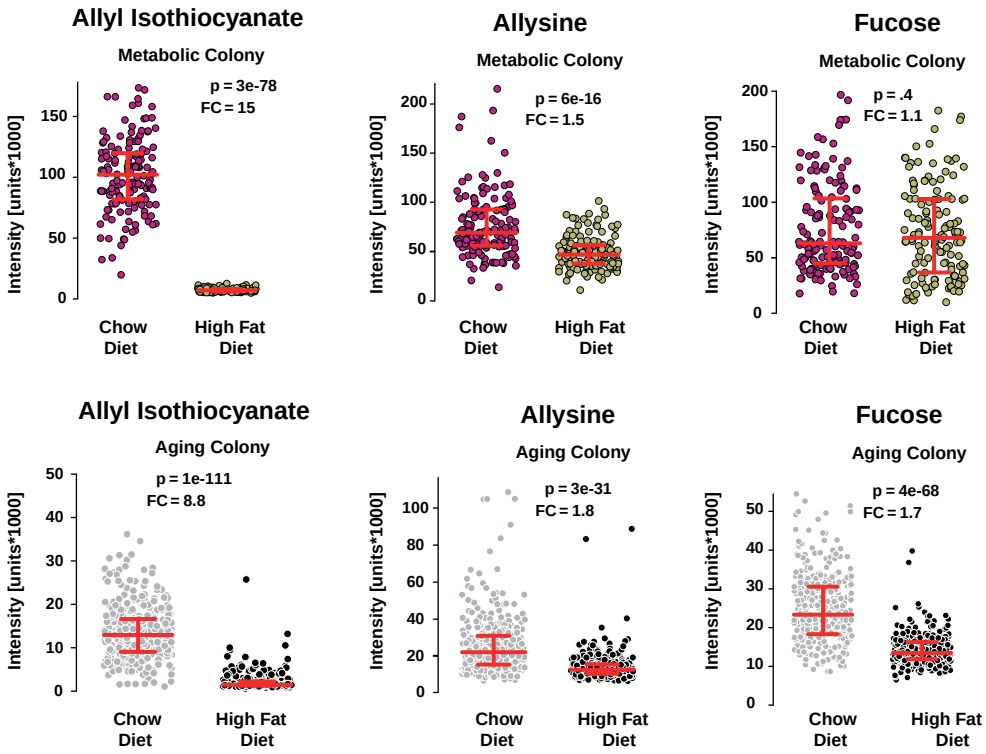


Figure 2.19: Three metabolites, known to be highly enriched in mice on the HF diet are used as a positive control in this new mouse cohort. Allyl Isothiocyanate, allysine and fucose are all detected in the new mouse colony, however all three metabolites show weaker signal and smaller fold changes than those detected in the previous experiment by Williams et al.

One can see from this analysis that there are many significant metabolites QTLs that differ between diet cohorts. The top QTL results can be seen in Appendix 6.2. On chromosome 3 of the QTL results, there is a range of significant QTLs that are manifest in the CD diet mice. The converse is true for chromosome 2 with mice on the high-fat diet. This speaks to the complex nature in which environment (diet) can allow the dissection of genetic drivers of certain phenotypic traits (Abiola et al., 2003).

From the assortment of metabolites found to have significant fold changes between the diet cohorts D-2-Hydroxyglutaric acid, pyruvate and glutathoine also have significant QTL results. The QTL for D-2 Isocitrate dehydrogenases (IDH) was already

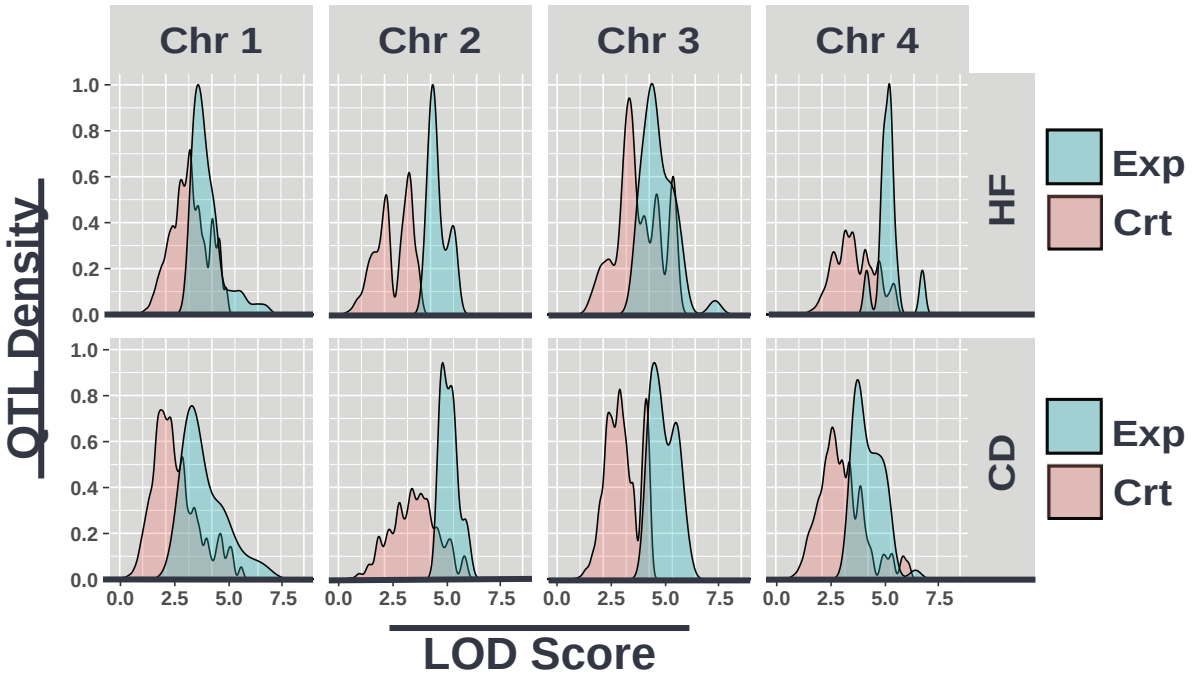


Figure 2.20: This figure shows the raw computed LOD scores for the metabolites tested against the true genome array and the QTL LOD scores of testing those metabolites against a scrambled array of genotypes. The distribution of NUL LOD scores is given in red and the actual computed LOD scores are given in blue

found in a previous study of a mouse cohort and servers as a validation that we have generated high quality data. It is known that a mutant form of isocitrate dehydrogenase can also preform the reduction of D2-hyudroxglurate and it is seen as having a suggestive QTL. In the many genes that can be found in the QTL regions with the high LOD scores, D2HGDH is an already known catalyst for the production of the metabolite and allow the QTL to be solved quite readily.

The QTLs for glutathione and pyruvate are not as obvious to solve as compared to D2GDH. If one looks at the strain wise values for pyruvate given on the right side the haplotype maps, there is a large jump between the two strains with the highest pyruvate levels and the rest. This is an indicator that this QTL may be outlier driven rather than driven by key genetic drivers for two separate population pools of low and high pyruvate mice. On visual inspection, the difference between the top two and the third mice is almost the same amount as the difference between the third highest mouse and the last. Even though each strain in this contains many biological replicates, there can still be chance fluctuations in the identified intensities of two or three mice in the same direction. Next, in inspecting the shape of the QTL between 40 and 47 mb on chromosome 11, it is evident that it is just barely significant. The dotted line is the QTL LOD score computed for mice that were both old **and** was on the high-fat diet in an attempt to see if there may be an interaction with HF fat diets in pyruvate only at old age. Lastly, when one looks int he QTL regions, it is strewn with ion transporters and Riken gene which would make the QTL difficult to solve. Pyruvate has such large number functionalities in cells and is involved in numerous metabolic pathways, which only confounds the problem (Voet and Voet, 2011).

On the other hand, the peak for glutathione is robustly above the significance threshold. The values for glutathione within the strains range from 13.20 to 11.07 with a gradual transition between the strains that have the B6 locus in the 80 Mb region

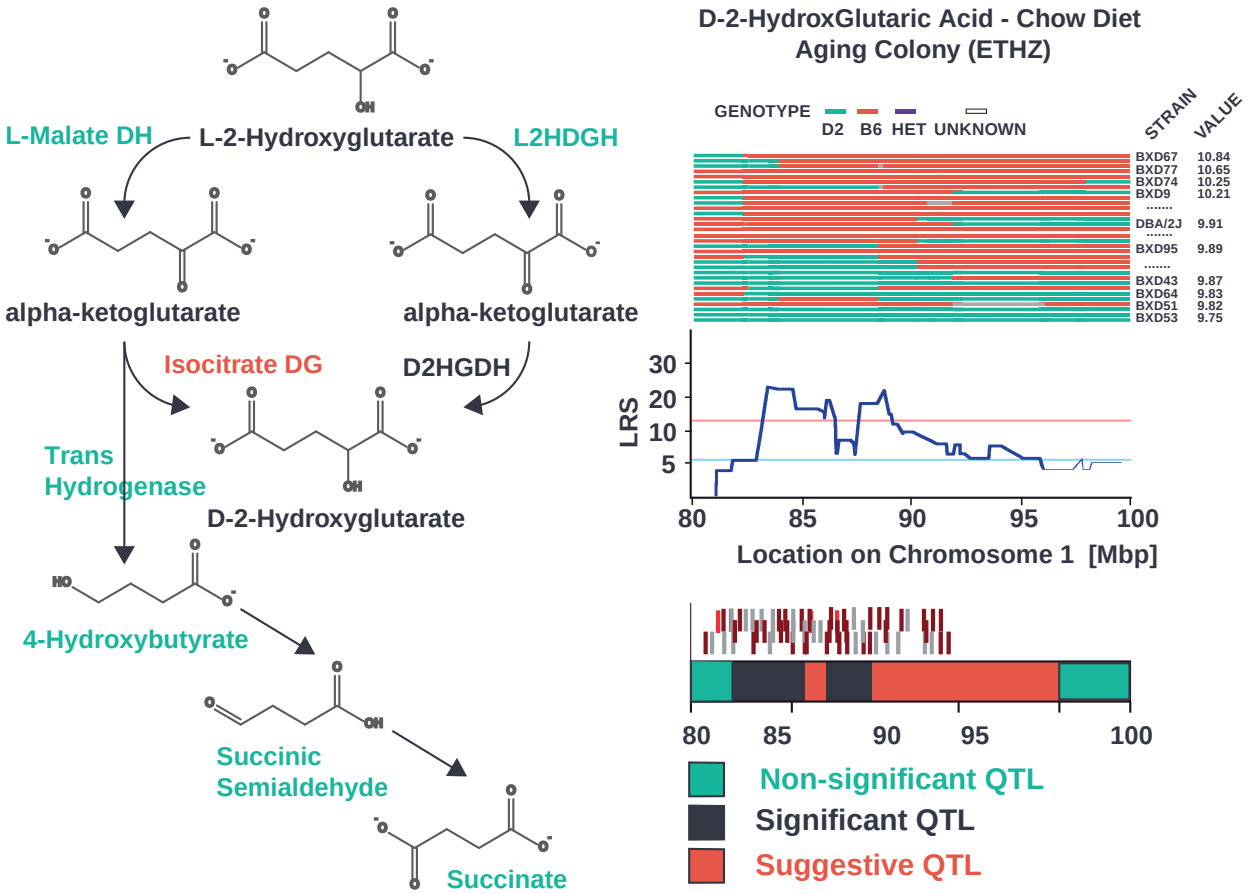


Figure 2.21: **Left:**Pathway showing the conversion of D-2-Hydroxyglutarate to α-ketoglutarate through Dehydrogenases. Mutate forms for isocitrate DG and D2HGDH are known to perform this catalysis. Proteins that are contained in a significant QTL for D-2-Hydroxyglutarate are highlighted in black. Proteins in suggestive QTL are highlighted in red and non-significant loci are highlighted in green. **Right**The haplotype map at the loci on Chr 1 where the QTL is found is shown. On the right of the haplotype map, the values for D-2-hydroxylate are sort from highest to lowest. One can see that most of the strains with higher D2Hydroxyglutarate have the B6 allele and those that have the lower concentration of the metabolite have the D2 all. At the bottom of the panel, once see the areas of the chromosome where there are suggestive and strong QTLs and the lines above the schematic of the chromosome are the genes found in this reagion

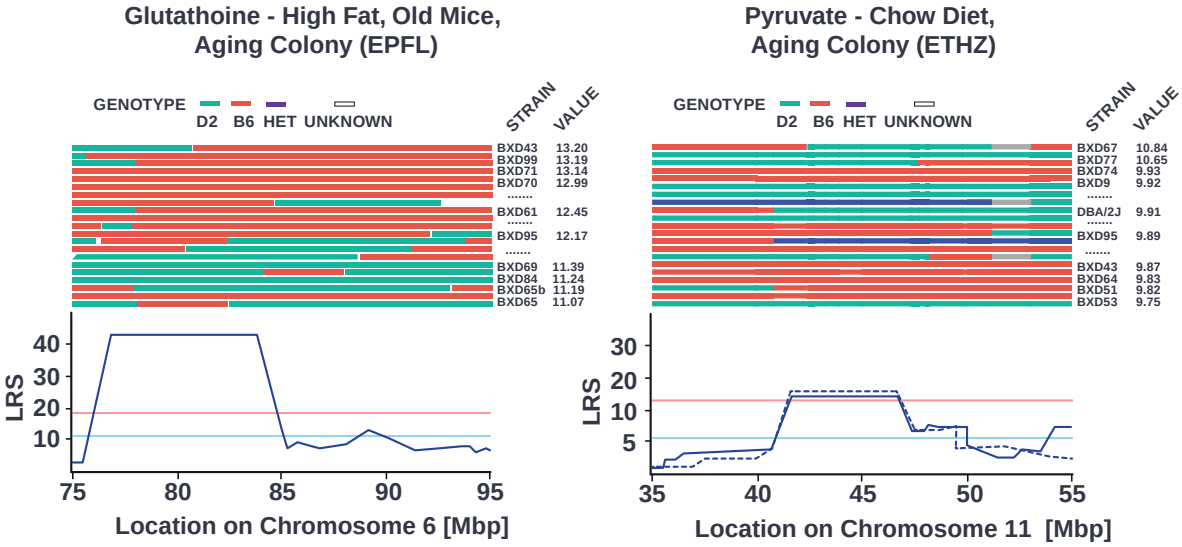


Figure 2.22: The QTL haplotypes maps and QTL LOD scores for Glutahione and Pyruvate in the region the chromosome whre the higherst LOD score for the meatbolite is found

in chromosome 6 and those that have a D2 allele. Additionally, the chromosome is much sparser in this region that it is in the region for pyruvate. The function glutathione is also more focused hopefully making it easier to solve the QTL detected.

2.10 Metabolite Set Analysis

After the glutathione QTL was found, metabolize set enrichment analysis is used to determine whether there are changes in the metabolites that are upstream and downstream from glutathione to get more biologically meaningful results. Set enrichment analysis of any kind requires the use prior knowledge to determine the probability that pathways are significantly affected given a set its components. The classification is highly dependent on the quality the prescribed ontologies. Luckily there is extensive data on the disease-related network and physiological networks for *Mus Musculus* within the metabolite analyst framework (Xia et al., 2016). In the figure given below the absolute correlation coefficients for metabolites from all mice

in the study, metabolites from mice in a single diet and metabolites from a single diet and single pathways were computed. In the total metabolites space, there is a very slight correlation. Removing half the mice in the set that do not consume the metabolites in the CD, HF-specific metabolites shown in green show a slightly larger correlation coefficients. Lastly, when only metabolites, within a simple KEGG pathway, are queried for their correlation, it becomes quite evident that the respective concentrations metabolites within the functional pathway modules are actively transformed, generated from each other as they pass through the pathway. This is why on average the correlation coefficient for the in pathway metabolites is much higher. With this rationale, if we look at the full network glutathione.

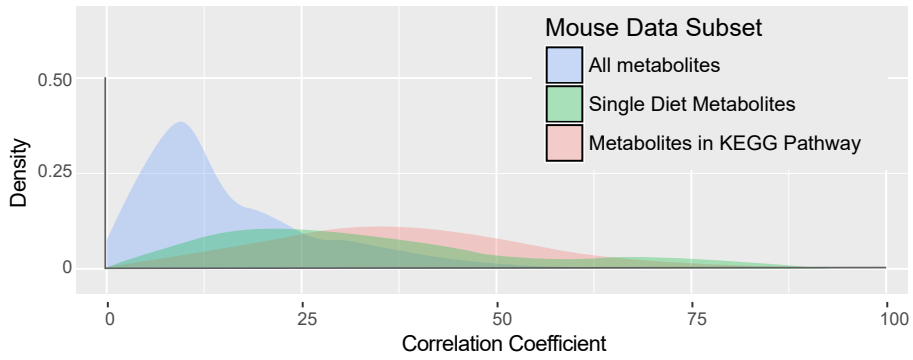


Figure 2.23: Correlation meatbolites In Metabolic Networks

Once the data for high enough coverage of a metabolic pathway is generated, it can be used for higher level interpretations in terms of pathways. This can be performed by putting the metabolite concentrations in the context of their catalyzing enzymes and gene pathways, visualizing all the fold changes on the pathway. The test is applied to determine whether a pathway is affected by either the age or diet segregation, is similar in nature to the question computing a QTL asks. In this case, we determine whether not the two populations of metabolites (HF/CD diet or Old/Young) projected on the pathway belong to the same population or not. If the mean fold change is conserved through many of the metabolites across two cohorts, they can be thought of as significantly effected(Wishart et al., 2013). Through this

method, we see that the oxidized version of glutathione (the red central node in figure 2.24) is metabolized by rate-limiting enzymes and that it accumulates in older mice. Thus this pathway is a hypothesis we can probe further one using Protein and RNA data.

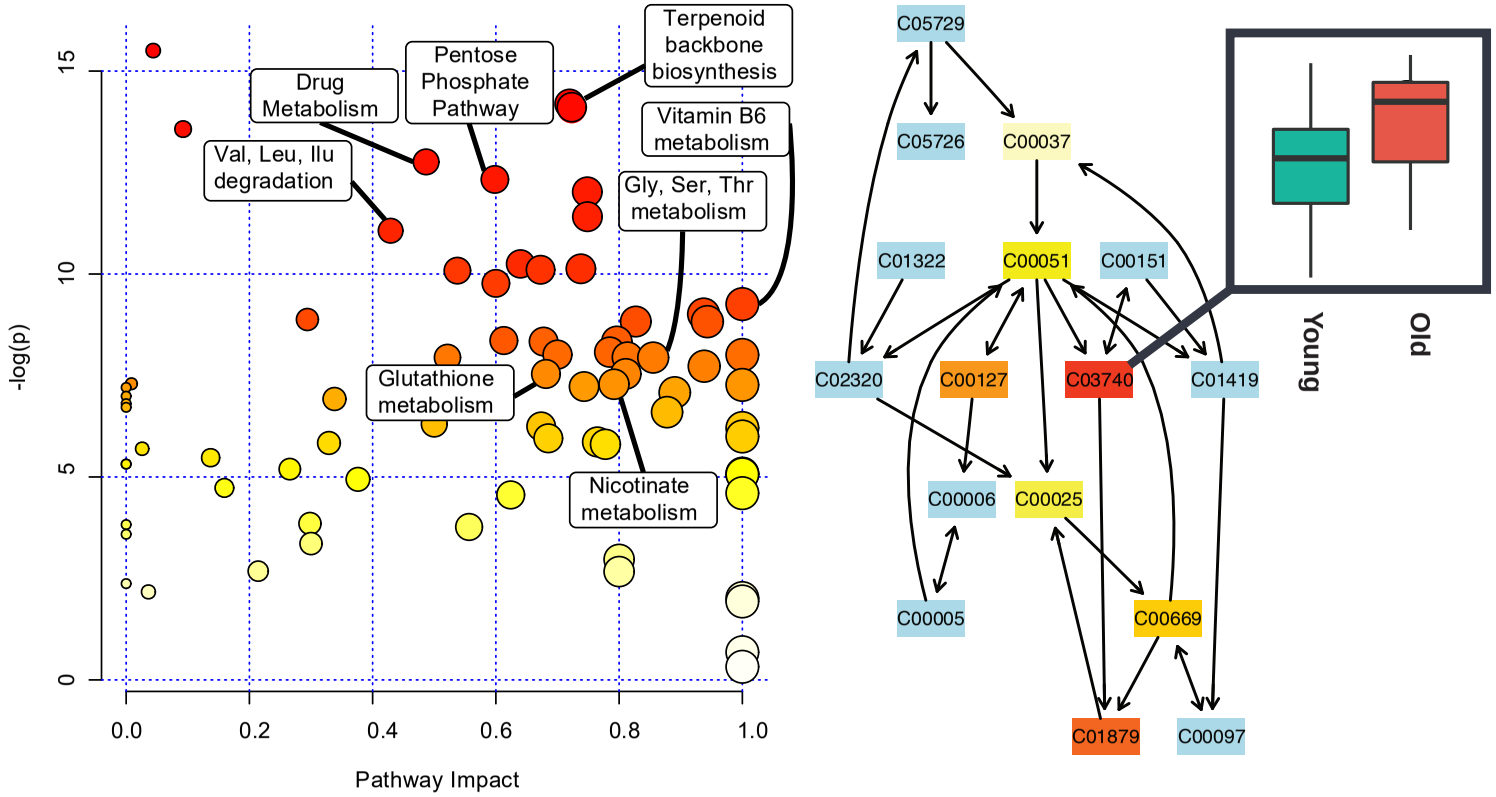


Figure 2.24: **Left:**The results of the MSEA analysis show good coverage in many metabolic pathways, given in on the x-axis. On the y-axis is the probability the pathway is differentially affected between young and old mice. Pathways such as terpenoid backbone synthesis, glycine, serine and threonine metabolism and vitamin b6 metabolism are both well covered and have significant fold change differences between the young and old mice. Two pathway of particular interest to the Auwerx lab, glutathione metabolism, and nicotinate metabolism are also significantly effects but have lower coverages, making flux estimations in the pathway difficult.
Right: Glutathione metabolism is shown. The synthesis of GSH from glutamate, cysteine, and glycine is catalyzed sequentially by -glutamylcysteine synthetase (GCS) and GSH synthetase. This pathway occurs in virtually all cell types, with the liver being the major producer and exporter of GSH. Glutathiones scavenges ROS and radical species and oxidized to glutathione disulfide(GSSG). It can be reduced to GSH by the NADPH-dependent glutathione reductase which may be less effective in older mice leading to the build-up of oxidized glutathione shown in the box plot. (Wu et al., 2004)

Chapter 3

Proteomics

3.1 Introduction to MS Proteomics

Proteins are involved in almost all the functional components of a cell, participating in enzymatic, structural and signaling modules that determine the phenotype (Hartwell et al., 1999). Historically, quantitative protein studies have been performed on a small subset protein for which high-quality validated anti-bodies exist. Unfortunately, anti-bodies experiments show a great deal cross-reactivity, are time-consuming to perform and are poorly multiplexed(Solier and Langen, 2014). Even though ELISA and western blots can show a much higher sensitivity than MS techniques and can be performed on a large scale using robotics, these assays suffer some poor resolution of relative protein concentrations and specificity as compared to mass-spectrometry methods (Solier and Langen, 2014). As a result, mass-spectrometry techniques have become the tools choice for proteomics studies. With more sophisticated data analysis techniques and faster, higher resolution instruments available on the market, simultaneous and comprehensive coverage protein concentrations from all functional classes in the proteome can be done, allowing

the holistic assessment the overall biochemical state a sample (Schubert et al., 2017).

The earliest use mass spectrometers with proteins were in top-down sequencing application. In MS-sequencing, a protein is incompetently digested, then single amino-acid residues are removed from the c-terminus and analyzed by a mass spectrometer to elucidate the sequence of the fragment. Once, the sequences of many small fragments are uncovered, overlapping peptide sequences are used to sequentially elucidate the amino acid sequence of the whole peptide(Steen and Mann, 2004). Bottom-up or peptide-centric methods are used to analyze protease digestions and prototypic peptides to perform quantification on a variety proteins in a mixture. Contemporary applications MS-Proteomics include DIA(data independent methods) which uses spectral libraries to accurately quantify large number proteins in a sample SRM(selective Reaction Monitoring) in which a pre-selected group proteins can be monitored at high sensitivity(Picotti and Aebersold, 2012) and DDA(Data Dependent Acquisition) methods which can be used for deep proteome mapping (Nagaraj et al., 2011) and proteome-wide quantifications of model organisms (De Godoy et al., 2008). In all three methods, fast, high-resolution mass spectrometers that have the ability to analyze specific bands the mass range are required to obtain detailed information on what ions arise from fragmenting selected parent ions.

Proteome Data acquisition Paradigms

With Data Dependant acquisition a full spectrum the peptides that elute off a column into the mass spectrometer are analyzed. The instrument alternates between scanning the total ions recorded at the MS1 level (full scan mode) and precursor mode in which ions that have high intensities in the full scan are analyzed after being isolated and fragmented into smaller fragmentines(Bateman et al., 2014). It is an unbiased method for observing as many peptides at the MS2 level as possible

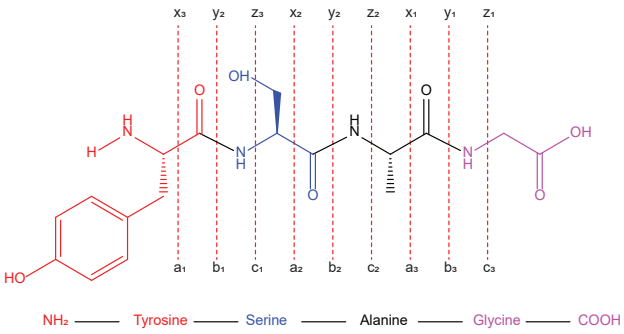


Figure 3.1: The proteomics nomenclature first purposed ion 1985 by (Roepstorff and Fohlman, 1984) used to describe sequence fragments is maintained as a convenient method for referencing fragment by the site they are cleaved from the peptide. Starting from the N-terminus, the ion is the fragment generated by a cleavage before the carbonyl, the b fragment is a cleavage after the carbonyl-forming the acylium ion and the c ion is a fragment derived from a cleavage after the amide bond forming an ammonium ion fragment. The converse is true for z, y and x ions which denote the same fragments starting instead from the carboxylic acids terminus. The subscripts denote the residue number from the N or C terminus the cleavages occurs (Roepstorff and Fohlman, 1984).

within a single duty cycle. Although DDA methods allow for detection significant portions mammalian proteomes it does not provide robust quantification all detected peptides(Richards et al., 2015).

Data independent acquisition methods such as SWATH(Gillet et al., 2012) acquire a signal from a large band MS1 peptides and to do not use the MS1 peptide intensities to determine which peptides precursors are selected for further fragmentation(Venable et al., 2004). This is in contrast to Data Dependant mass spectrometric techniques in which fragments with the highest intensities in the MS1 space are selected by the machine to be fragmented and further analyzed in the MS2 space. Signal intensities in the MS1 space can be highly stochastic meaning peptides may not always be selected by the mass spectrometer in DDA mode between samples if they do not consistently show the same high-intensity peaks. Another key difference between DIA and DDA methods is that while DDA machines run on an alternative fulls scan and precursors duty cycles, a SWATH machine runs on fixed pre-programmed duty cycles, where swath acquisition windows are programmed to

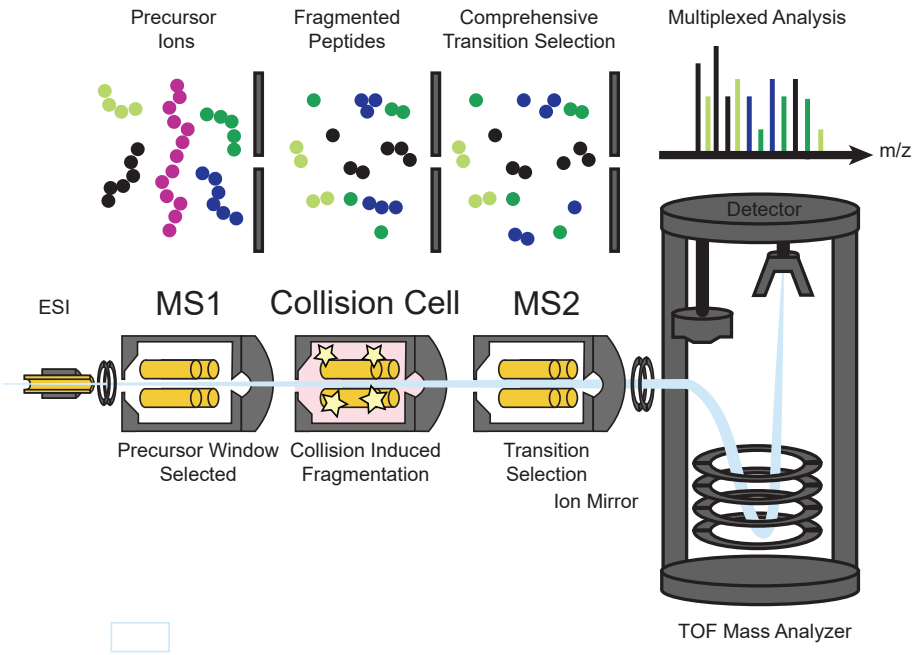


Figure 3.2: Schematic AbSCiex 5600+ Triple TOF Mass Spectrometer used for DIA/SWATH Experiments. One can see the electro-spray injection(ESI) needle on the left followed by three quadrapoles and a Time of Flight(TOF) Mass Analyzer. Digested peptides are separated using a high pressure LC column and then injected into the mass spec through the ESI needle. In DIA mode operation, the first quadropole (**MS1**) acts as a bandpass filter allowing a selected range peptides through to the collision cell. Ion transmission through a quadrapole can be tuned by applying combinations DC and RF voltages that are resonant with a certain interval m/z ions. In this case the **purple peptide** is too large and is filtered out. The transmissive ions are known as the **precursor ions** are fragmented through collision with neutral gases such as N_2 or Xe in the collision cell. The collective fragments from all the precursors are known as **fragmentines** or **transitions** are focused into the time-?flight chamber for mass analysis. The resulting spectra (shown above the TOF analyzer) is a composite all the fragmentines from the precursor ions

scan over a large mass range in increments 5m/z to 25m/z(Röst et al., 2017). As the wide precursor isolation window moves over a mass range between [400-1000] in 5-25m/z intervals, all the peptides in the interval are fragmented and analyzed together. Complex MS2 spectra are generated when all of the fragmentines of the precursor isolated space are recorded together and necessitate the use of computational tools to deconvolute.

The most common instrumentation used in SWATH is the quadrupole time-flight

mass spectrometers, A schematic an AbSciex triple TOF is shown in figure 3.2. Irrespective the instrument being used, accurate SWATH implementation have a quadrupole which can act as a selective band filter for precursor selection and a high-resolution mass analyzer than produce high-quality MS2 spectra.

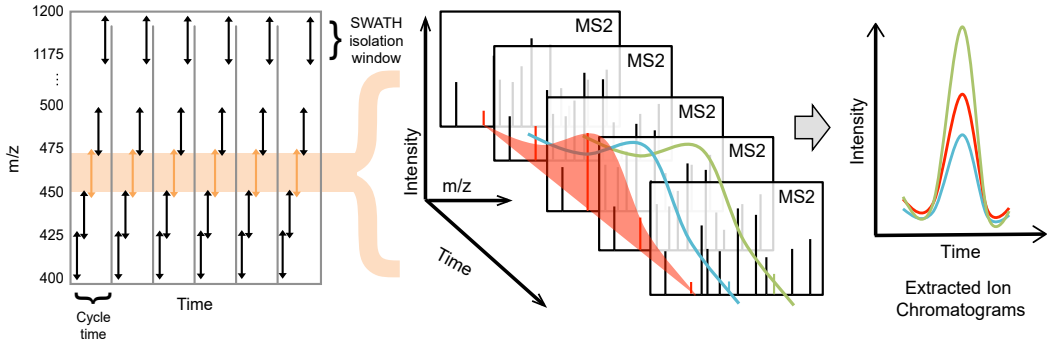


Figure 3.3: Figure adapted from (Röst et al., 2017) Illustration the SWATH-MS Duty cycle. (1) As peptides elute of an orthogonal chromatography column into the mass spectrometer, a window mass ranges OR SWATHS are isolated and fragmented. (2) The ions that results from the fragmented peptides is recorded as a convoluted spectrum including fragments from the three **Red, Green and Blue** peptides. For each the swatch, there is a 100 ms acquisition cycle in the MS2. From 400-1200 m/z this makes a full duty cycle 3.2 seconds. (3) Once the acquisition is complete, specific ion fragments can be extracted from the multiplexed peptide spectra to produce ion chromatograms for peak groups.

In the figure 3.3 the red and blue peptides can be use to illustrate the principles of SWATH sampling of the precursor space. Each consecutive window moving in the z-axis out the page shows an MS2 spectra at subsequent time points of the chromatographic separation. The red and blue digested protein (we can assume they a have m/z 500 and 502 respectively) elute off a chromatography column and are injected into the first mass analyzing quadrapole. An the MS1 a precursor isolation space 5m/z is used and selects both the red and blue peptides for fragmentation in the collision cell. Both peptides are co-isolated, co-fragmented and co-analyzed leading to the MS2 seen in the windows of figure 3.3. As a result, fragmentines from all the precursors are present in the final multiplexed data. By determining all the fragmentines that are known for a single peptides, the multiplexed MS2 recordings of

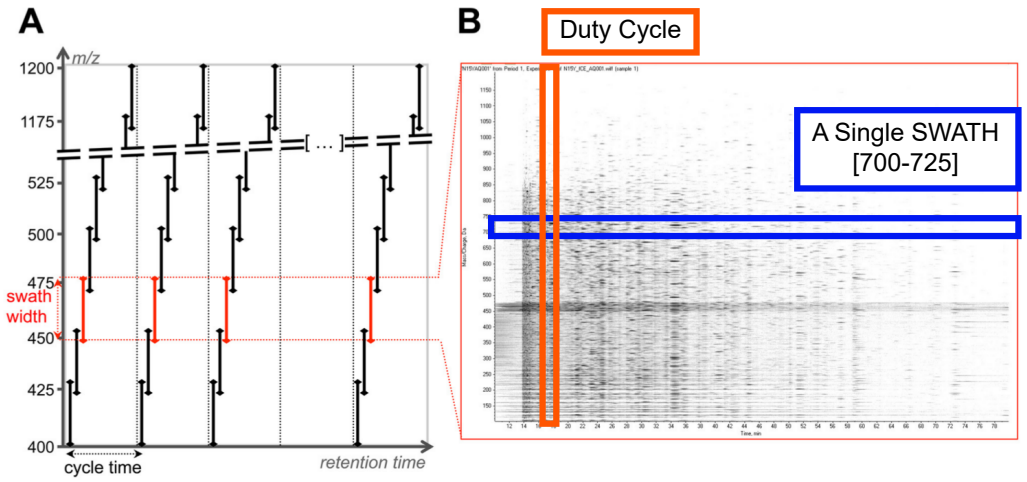


Figure 3.4: Figure adapted from (Röst et al., 2014) A.SWATH duty cycle schedule. B. Final 2D chromatogram LC-MS SWATH Data C.Reconstructed and aligned XIC D. A. The x-axis represents the chromatographic dimension and the y-axis the m/z domain with the vertical arrows shows the m/z acquisition windows. The signal intensities the peptides at each mass/charge unit at a given point in the elution shown in the grey and black scale. Each one the black dots seen is a different precursor ion.

the transition mixtures can be deconvoluted. In this way, the continuous monitoring of the peptides is done both in terms time and MS2 signal thus comes at the cost added complexity in separating and quantifying the signal from different precursor peptides. However, When the instrument is forced to fragment all the precursors in every duty cycle within the limit detection, the result is a consistent quantification all precursors in the sample (Gillet et al., 2012).

DIA SWATH Operation

In figure 3.4, an example the data acquired from the initial implementation SWATH Acquisition can be seen. A SWATH (highlighted in blue in panel B) is the collection fragments that are acquired in a single mass isolation window, in this case, 700-725 throughout the chromatographic separation. The cycle time is highlighted in red. The quantitative accuracy the proteomic analysis is inverse dependent between the SWATH-window and dwell time on each window (Röst et al., 2017). As SWATH-

MS has a chromatographic separation, the cycle time must be sufficiently small to resolve multiple close peaks in a single SWATH.On a 400-1200 m/z range with 25m/z swaths, a machine with 100 ms dwell time on every acquisition window require 3.2 seconds per cycle. Increasing the cycle time increases the accuracy the fragmentines being quantify but reduces the resolution the chromatographic peak each peptide's elution(Lange et al., 2008). Additionally, the precursor isolation window can be tuned to use smaller 5m/z wide swath windows in the lower m/z peptides [400, 600] where there is a large number eluting peptides and larger windows in the [1000,1200] regime where the number of quantified peptides is much smaller. This can increase the number peptides detected by 10-15% in complex samples where multiple peptides are co-isolated in a single swatch-window.

3.1.1 Extracting Peptide Chromatograms

In contrast to DDA acquisition, in which all the MS2 spectra come single peptides fragments, the MS2 spectra produced with SWATH contains peaks from many precursors and are not directly searchable against a peptide database. Processing the spectra is done using a C++ distribution called OpenSWATH which allows the automated processing the SWATH data and spectral library(Röst et al., 2014). In figure 3.5 the top panel shows MS2 spectra of a swath from 700-725 which includes fragments from the peptide **WIQDADALFGER**. The lower panel shows DDA spectra, our prior knowledge, two fragments **WIQDADALFGER** that were isolated by the mass spec and quantified in DDA mode. In figure3.5D. the spectral features from two fragments y4 and y10 identified both DIA(top) and DDA experiments(bottom) are highlighted.

In a DDA experiment, the y4 and y10 fragments would have high intensity and would be selected by the machine to be isolated for fragmentation thus allowing us to determine the major characteristic peaks from these precursor ions. The pattern

of peaks detected at the MS2 is a unique fingerprint signature for each precursor fragment(Gillet et al., 2012). This fingerprint can be used to assign fragmentine peaks to a specific fragment and peptide in a convoluted DIA MS2 spectra. Once the major fragmentines from the precursor peptides can be identified in the complex spectra using previously determined DDA signatures we can plot fragmentine intensities against the chromatographic time dimension and signal group overlapping peaks begins to appear, as seen in fig 3.5.C. If many fragments are derived from a single precursor ion, the peaks result in overlapping extracted ion chromatograms(XICs). Statistical analysis can then be used to score peaks taking into account the peak shape in elution profile. Lastly, the maximum intensity of the highest peak for each the fragments is used for peptide quantification.

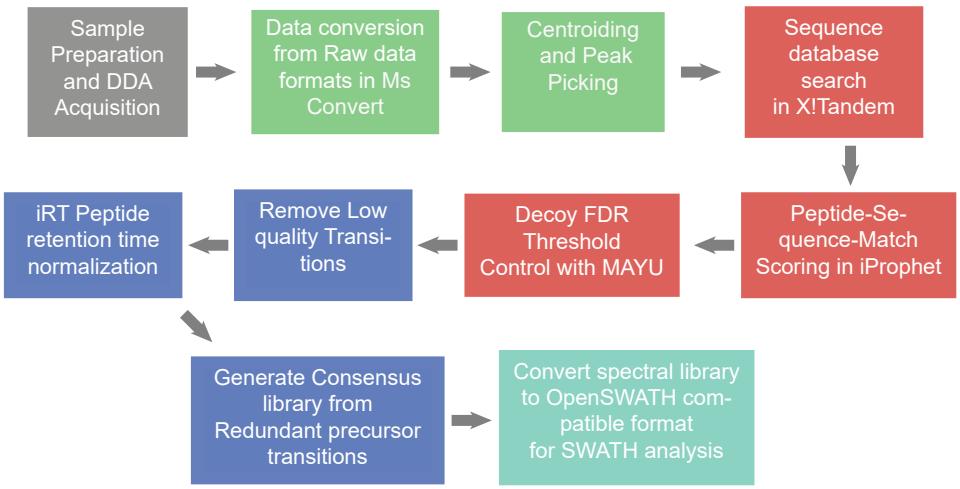
3.1.2 SWATH MS for BXD Mouse Liver Proteomics

To analyze complex proteins mixtures extracted from the BXD Mouse livers, SWATH is ideal as it allows the quantification of a large set proteins across multiple samples and providing a consistent number protein quantified, with high accuracy, reproducibility and sensitivity. Using DDA methods, many more proteins can be detected but the measurement is not reliable across many samples. SRM, a classical targeted proteomics has slightly higher sensitivity compared to SWATH but the number protein that can be quantified is relatively discrete. With certain optimizations SWATH shows coverage comparable to MS1 label-free quantification as in shotgun identification and sensitivity comparable to the performance of SRM (Liu et al., 2013). In a side-by-side comparison, plasma protein detection by SRM and SWATH (Liu et al., 2013) show similar CVs and slightly lower quantification proteins (33 vs 37) in SWATH.

Another distinct advantage of using SWATH proteomics is the ability to re-process data post hoc. If an XIC) hows two probable elution peaks for a precursor, it may be

difficult to determine which chromatographic peak should be used for quantification. In SWATH, all transitions data is saved, meaning different transitions can be added to the composite XIC to determine which is the correctly eluting peak (Gillet et al., 2012). Moreover, if a major peak in the XIC is distorted due to some interfere, one could simply select a different peak to use for quantification. Unless a sufficient variety of peaks were sampled in SRM this would not be possible post-hoc(Gillet et al., 2012). Lastly, SWATH allows for iterative mining proteomic data enabling the biologist to simply reprocess the data and select peaks for different proteins of interest rather than having to re-perform the experiment targeting a different set of proteins in the same sample (Gillet et al., 2012).

3.1.3 DDA Spectral Library Generation



Raw Data Processing and Conversion:

For spectral library generation, representative samples from the full cohort of samples need to analyzed in DDA mode in order to tabulate pertinent information about the peptides and their transitions. High-quality protein libraries from a 58 mouse

liver samples run in shotgun mode have already been constructed in previous BXD mouse liver studies which allow us to use the previous data (Williams et al., 2016). The data from these runs have to be centroided and converted from a proprietary to an open source format such as mzML or mzXML in order to perform the subsequent processing steps. A range converters can be used however, the centroiding algorithms may change the appearance XICs depending on whether the maximum peak height or peak integral is used. ProteoWizard-MSCovert which is specifically written for processing LCMS data using the max peak height and generates the optimal centroid data for database searching (Kessner et al., 2008).

Peptide Sequence Annotation:

The next steps, database searching, scoring and spectral library generation can all be performed with using components an open-source suite protein data processing packages known as TPP (Trans Proteome Pipeline). Using, iProphet (Shteynberg et al., 2011) all the MS2 spectra are searched against multiple sequence database search engines such as X!Tandem and Comet(Eng et al., 2013) which can annotate spectra by comparing experiential peaks to expected peaks from in-silico digested libraries within given mass defect tolerances. Rather than exhaustively searching all peptide-sequence combinations iProphet assigns a match with the highest probability and quality score to each peptide-spectrum match(PSM) through a penalized greedy search algorithm (Shteynberg et al., 2011).

If an MS2 spectrum is erroneously annotated, the error will propagate to the DIA analysis and peaks found in the DIA quantification will be thought to have originated from the wrong peptides. MAYU is used to need to perform a false discovery rate estimation at the PSM, peptide and protein levels (Reiter et al., 2009). It takes the iProphet output and generates robust empirical FDR estimate using the target-decoy technique(Elias and Gygi, 2007). A 1% Protein FDR, 0.2% Peptide FD, 0.08% PSM FDR and iProphet score threshold 0.9774 is used to keep only high-

quality spectra in the library.

Elution Time Normalization:

In order to use a DDA spectral library to interpret DIA data, the peptides in the DDA library must elute at a similar time interval to the DIA quantifications. This is an issue because the absolute retention different peptides may vary on the order minutes between runs, columns and mass spectrometers. In order to normalize the elution behavior between runs, 11 intensity and retention time standards (iRT) peptides are spiked into all sample at a high concentration (Bruderer et al., 2016). These peptides elute at regular intervals on the LC gradient. The elution time of the second peptide is set at an arbitrary unit of 0 and the 11th at 100. A linear fit between the interim peptides allows us to project all sample peptide retention times onto the 0-100 scale instead of the chromatography time scale. The same iRT peptides are also spiked into the DIA samples allowing us to predict when the samples peptide will elute reducing the search space for where a peak corresponding to a peptide in the DDA library and DIA data (Bruderer et al., 2016).

Spectral Library Consolidation:

In a given DDA experiment, there is a significant amount redundancy in MS2 data recorded for each MS1 precursor ion. MS2 fragmentation can be triggered multiple times or triggered stochastically as mentioned before, resulting in redundant MS2 spectra in our data. Using SpectraST a consensus spectra consolidating the features of all MS2 spectra for a given MS1 peaks can be generated. This is done by averaging the intensities of replicate peaks and removing individual MS2 spectra that do not conform to the mean spectral patterns(Lam et al., 2008). In the end, 5/6 transitions in multiple charge states are taken for every precursor ion (Schubert et al., 2015). Transitions that are found in the MS1 data or have an m/z below 400 are noisy and thus excluded from the library. Once a library of MS2 spectra with sufficiently high PSM score, low FDR rate, and minimal redundancy is produced, it can be converted

into the TraML for integration into the OpenSWATH package for analyzing DIA data.

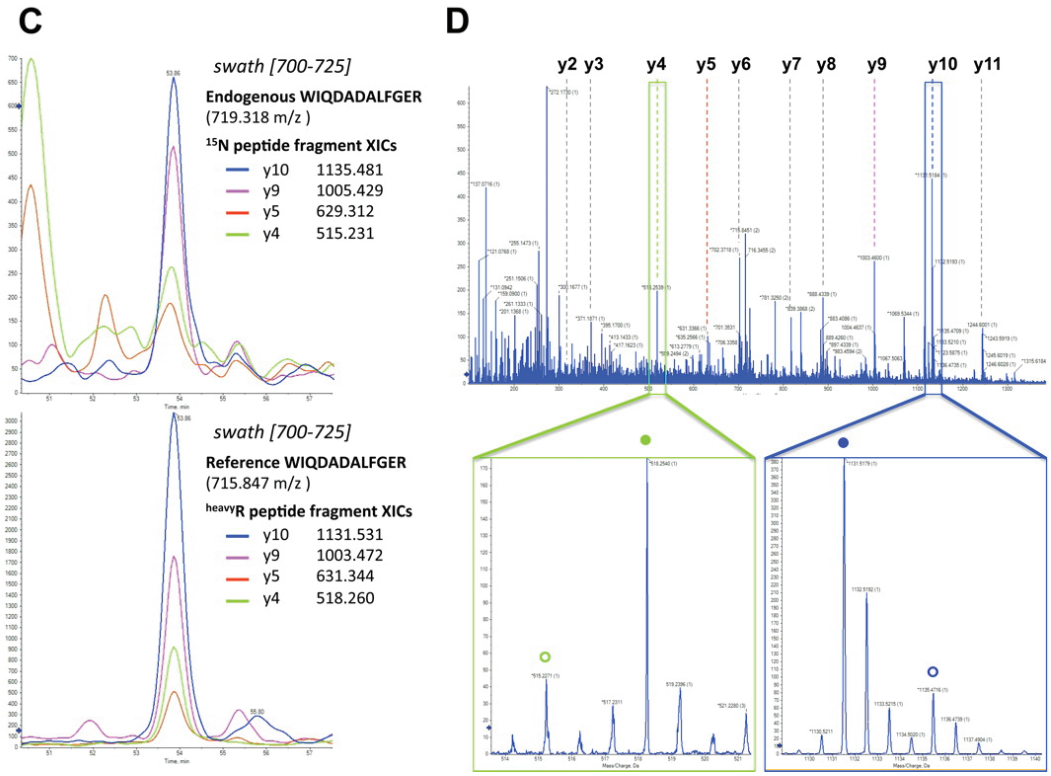


Figure 3.5: Figure adapted from (Gillet et al., 2012) **Top:** Right: DIA M2 spectra of the 700-725 m/z swatch for the peptide WIQDADALFGER Left:The extracted ion chromatogram for the peptides determined using the y10, y8, y4 and y5 overlapping peaks around the 54min elution time **Bottom:** Right: M2 spectra from the y4 and y10 peaks showing all of the fragmentines from the two fragments break down into when the peptides are isolated and subjected to collision induced dissociation Left: The XIC from for the peptide WIQDADALFGER after scoring and aligning the various MS2 peaks to determine the most likely peak

3.1.4 Experimental Proteomics Protocol

In order to samples with SWATH-MS and DDA, proteins must be extracted from the liver tissue and digested into shorter peptides. Trypsin is used to cleave proteins at long basic amino acids after which the samples are cleaned to remove salts, lipids and other metabolites that can interfere with the LC-MS chromatography column. As the DC bias can change on the mass spectrometers detector, many

internal standards as prepared in order to normalize the peak times and peptides intensities. For 20300 µg protein extracted from the BXD livers, 20 pmol/µl fetuin B and 20 pmol/µl alpha1-acid glycoprotein (AAG) in addition to a set universal protein standards (UPS1) are spiked in on the day analysis.

On average, 100µg of protein is extracted from each sample. For the initial extraction from the liver tissue, the sample cuvettes are filled with acetone which disrupts the cells membranes and precipitates the proteins. Next, urea and bicarbonate buffer are added to denature the peptides and facilitate digestion by trypsin. Iodoacetamide and dithiothreitol are di-sulfide reducing agents added to the mixture to prevent cross-linking between the extracted protein (Voet and Voet, 2011). The protein samples are incubated in the denaturing mix for an hour before trypsin is added to digest them overnight. The digestions should not be longer than 24 hours, as trypsin can hydrolyze bonds within its own backbone creating large peaks in the mass spectra. Next, peptides are cleaned using a C_18 column. The samples are washed with MeOH followed by mixtures ACN:H_2o with increasing proportions water to remove polar impurities. Using a 1:1 mixture ACN:H_2o, the peptides can be eluted. The LC solvent gradient goes from a 2% ACN solution to a 50% solution. The column therefore also removes any peptides that could not be resolved with this solvent gradient. Lastly, the digested peptides are centrifuged and dried for storage at 80°C.

On the day of the mass spectrometry runs, the dried samples are resuspended in a mixture ofACN:H_2O 2:98 + 0.1% FA to a target concentration around 2501000 ngµl. The samples are adequately agitated and sonicated to free proteins that may be stuck to the walls of the cuvette. A final centrifugation removes residual impurities that may cause issues with the mass spectrometer. The UPS1 peptides are spiked into the samples and act as a control for the sample injection. Indexed retention time (iRT) peptides which elute in a well-defined interval on LC are added to control for

variation in the chromatography between samples (Escher et al., 2012). Once the samples and internal injection, time and digestion control are loaded into MS tubes. The samples are ready for injection into the mass spectrometer in either DDA mode for library generation and DIA/SWATH mode for quantification.

3.2 Data Analysis

Once the mouse liver proteomics samples have been digestion, analyzed in the mass spectrometer and processed using OpenSWATH an excel sheet with all the transition intensities, peptide and protein can be exported. Once the data is in this tabular form, the remainder of the analysis is done in R.

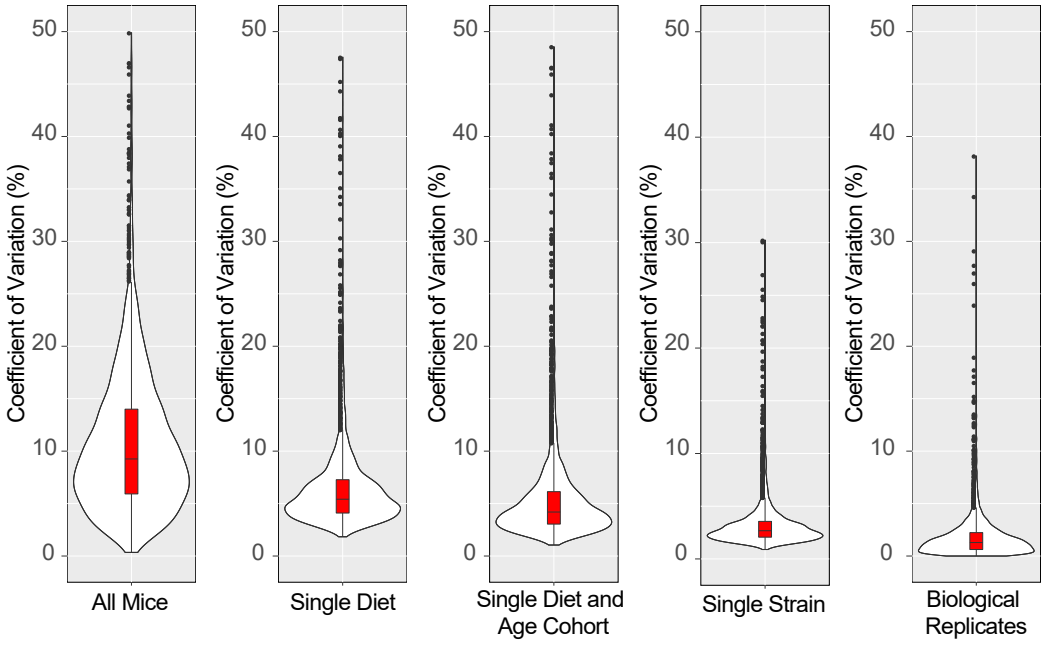


Figure 3.6: Coefficient Variation SWATH-MS Run between Mice measured in two batches

3.3 Quality Control

The protein level intensities are computed by taking the mean intensity the three most intense peptide fragment peaks. As a preliminary quality check, the coefficient variation is computed for the protein level (3.6). Across all the mice, there is an 8% coefficient variation. In contrasts to metabolomics where 8% coefficient variation can be seen in the injection replicates, peptides are more homogeneous in their physiochemical properties and are able to be prepared and analyzed in a more reproducible manner. Additionally, there are many more layers post-processing SWATH data in comparison to shotgun-MS data. The CV for proteins in mice that eat single diet is 5%. Once the age is controlled for, this dropped even lower to 4%. Between biological replicates, the variation is minutely allowing for the detection small changes between cohorts.

3.3.1 Batch Effects

The principal component analysis (Appendix figure 3.7) of the data shows that large parts of the variation in data come from batch effects, rather than the intrinsic biological heterogeneities between mice. The first two principal components explain 57% of the variation in the data, the first principle component axis segregating the samples according to their batch numbers. The analysis of the peptides intensity densities between batches shows larger differences than peptides intensity densities across mice on different diets. Similarly, hierarchical crusting analysis(not shown) clusters the mice in batches 1, 2.0 and 2.1 into three distinct clusters. Although batch effects are a problem in proteomics which can be tackled using a range of statistical techniques(Leek et al., 2010), no such methods of correcting the peptide intensities or correcting p-values generated from statistical tests are used in the subsequent analysis as this data is preliminary.

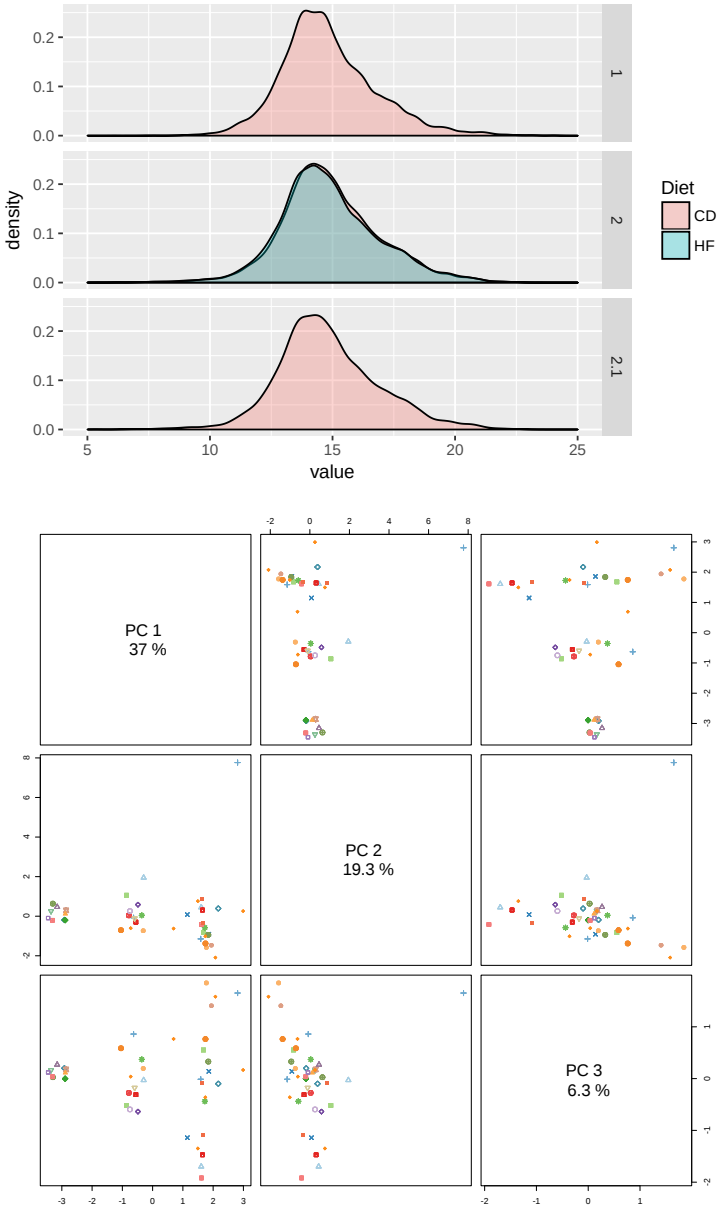


Figure 3.7: **Top:** Peptide Intensity densities in the three preparation batches colored by diet. Although diet does not significantly effect the peptide quantifications and the preparation does seen to skew the densities slightly.
Bottom: PCA analysis of the proteomic data indicates the first two principle components explain 56% of the variation seen in the data. Unfortunately, the first principle component separates the data based on the batch

3.4 Proteomics Fold-change Analysis

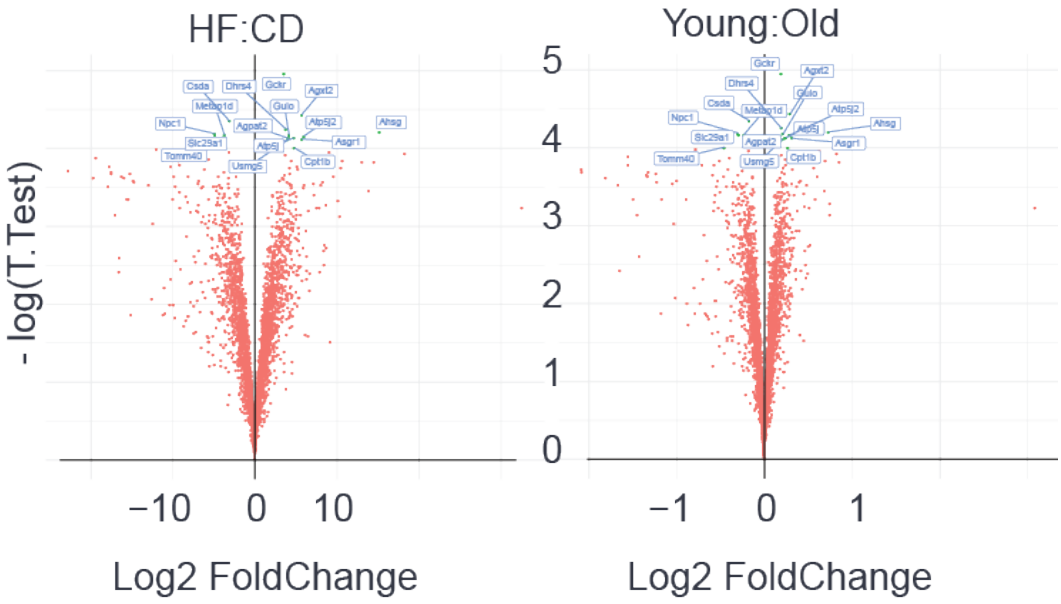


Figure 3.8: Volacano Plots Diet and Age Related Protein Foldchanges

Analogous to the metabolites, there are proteins that have a large fold change in between the diet cohorts as compared to the age cohorts. In fact very liberal fold change threshold have to be set in order to highlighted a few significant proteins in the volcano plot below (figure 3.8). At the same significance thresholds for diet as age, there are hundreds significant proteins. This speaks to the difficulty performing age-related studies. The effects sizes between the cohorts are not large, and in order to detect the few up or down regulated proteins in old mice, one must be willing to have a very high false discovery rate. In a similar study performed in over 4000 mouse tissues by Walther and Mann, minimal age related changes in the proteome were found.

3.5 Protein Biomarkers

In order to find predictive biomarkers for age, the significance threshold is sequentially lowered to allow 5, 10, 15, 25, 50 and 100 proteins into the biomarker prediction pipeline. 4/5ths the data will then be used to train SVM (support vector machine) and RF(Random Forest) models to classify old and young mice. The n 1/5 validation set will be used to construct confusion matrices and ROC curves to determine the classifier accuracy each method. This is done multiple times, stochastically leaving a 5th the data out for validation. In the end, the factors protein that is chosen most frequently in these machine learning techniques will be used in further QTL and network mining techniques.

3.6 ROC Curves

ROC(receiver operating characteristic curve) curves originally developed in world war 2 for radar signal detection is a graphical method evaluating the performance binary classifier systems (Hajian-Tilaki, 2013). They are routinely used to evaluate classifiers and in the case of this thesis, we are evaluating the prediction accuracy of a SVM and RF classifiers assignments of mice to old or young age cohorts from a small set measured proteins concentrations. A ROC curve is defined as a plot of sensitivity (the number true positive decisions/the number actually positive cases) as the y coordinate versus its specificity(the number true negative decisions/the number actually negative cases) or false positive rate (FPR) as the x coordinate. From the ROC curves, the area under the ROC curve(AUROC) can be computed and used to determine which classifier and ensemble factors are best able to differentiate mice on different age cohorts (Hajian-Tilaki, 2013).

3.7 Random Forest

Decision Tree methods are a widely used for classifying large multivariate data. This method is non-parametric and dissects data into segments that start with a root node, and multiple branching nodes terminating in my leaf nodes. At each branch node, the data is segmented into two populations along a variable, this is down hierarchically, moving down the tree until a final classification is declared (Song and Lu, 2015). Although tree methods have good interpretability and provide easy intervention options, classification trees use a greedy search algorithm and thus can be unstable to noisy data(Song and Lu, 2015). Good prediction results with the diet classifications can be made using tree methods because there are a few very strongly represented molecules in each diet. The age-related classification is not as good as fats(which show large variances and are not well measured) are often selected for partitioning the groups. Moreover, single tree methods are not stable when small sub-samples of the populations of mice (a few strains of interest) are used.

Random forests are an extension of decision tree methods in which a large collection of de-correlated decision trees are built using random sample selection and then averaged (Hastie et al., 2009). For a given set of data, a subset of the data is randomly chosen and a decision tree is generated. Additionally, a random subsets of variables in the dataset is used for determining the split points in the decision trees. Many noise trees are generated from a random selection of data variables and generate an ensemble of classifications. Each tree in the ensemble gives a vote and a classification is made using a majority vote(Hastie et al., 2009).

The randomForest forest package in R is used to determine in a mouse sample's to determine the most important proteins cocnentrations for predicting age and diet cohorts (Liaw et al., 2015). Although Random Forests are difficult to interpret simply,Breiman suggest looking at the missclassification rates in each tree of the out

of bag(OOB) samples with and without each variable to determining th (Breiman, 2001). This means after each tree is generated the OOB mouse samples are classified with and without certain proteins is computed in order to determine the importance of the protein in the prediction (Hastie et al., 2009).

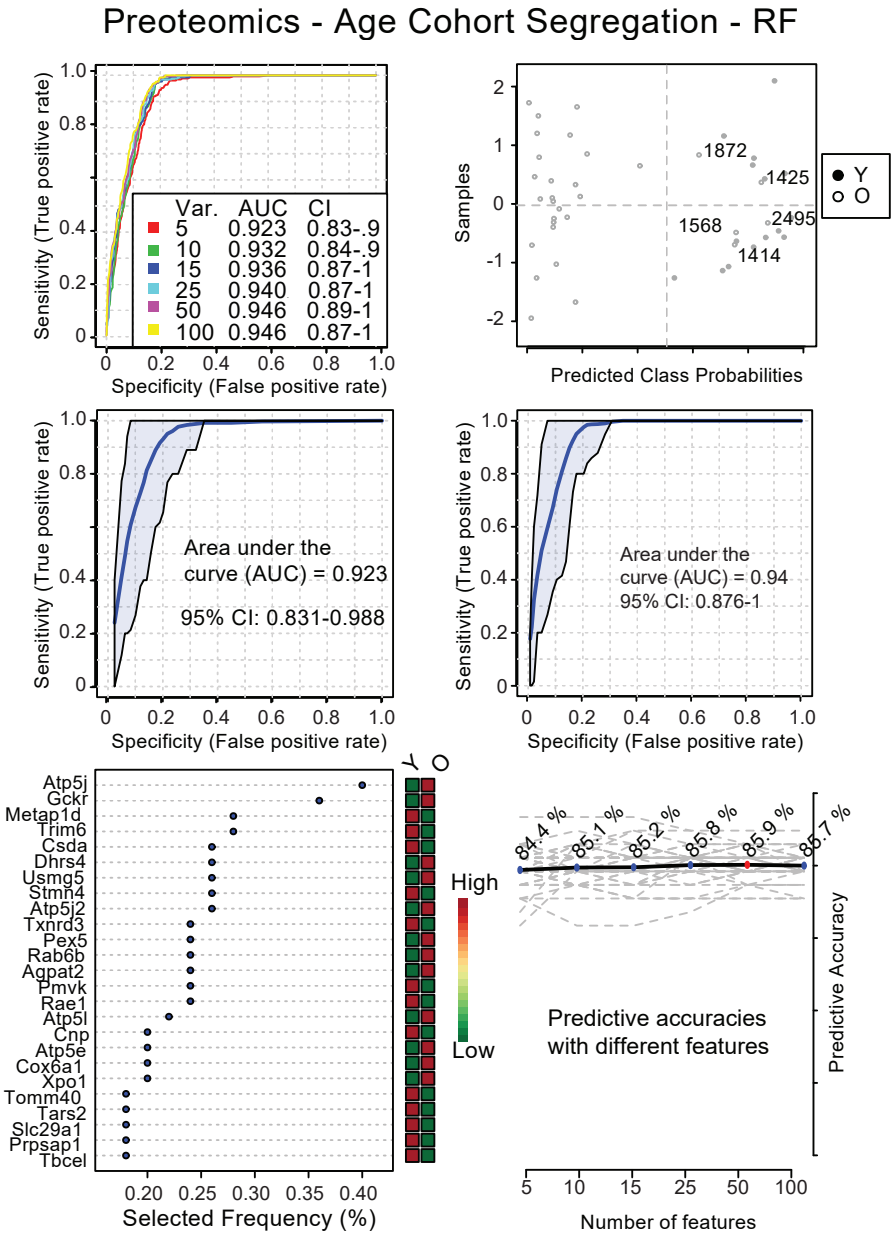


Figure 3.9: In the top left panel, the ROC curves for various model sizes are given, along with its AUC and the 95% confidence intervals. In the top right, the Dark and white dots are given in terms their classification probability. In the Middle row, the ROC curves for 5 and 25 proteins are given. In the Bottom left, the most frequency chose factors are given. These factors are seen as prototypic for aging and will be followed up with in further analysis using transcript data. In the bottom right the accuracy for each model size is given

3.8 Support Vector Machine

The SVM classifier is similar to older linear discriminant methods. In brief the algorithms attempts to construct the optimal hyperplane that can separate the diet and age mouse cohorts(Shawe-Taylor and Sun, 2011). The R package rminer is used in this case, with the varaible importance genreatedf rom the Imporatnce function (Cortez and Maintainer, 2016).

3.9 Conclusions From Proteomics

In this section, a small 22 mouse subset the 631 mice were analyzed using SWATH-MS. The protein quantification data that was put forth had many significantly different proteins for the different diet cohorts but very few significantly changed proteins in different age cohorts. As a result, SVM and RF methods were used, producing very good classifications of whether the mice were young old on the basis 5 to 100 proteins. Although the interpretability of these algorithms is difficult, one can try to look at overlaps between common proteins selected across both SVM and random forest and perform QTL analysis in the future. The number of samples analyzed is extremely small in relation to all the other BXD proteome data that will be quantified in coming experiments. The preliminary proteins hit found here are used with the metabolite data and transcript data to try to construct age-related protein network.

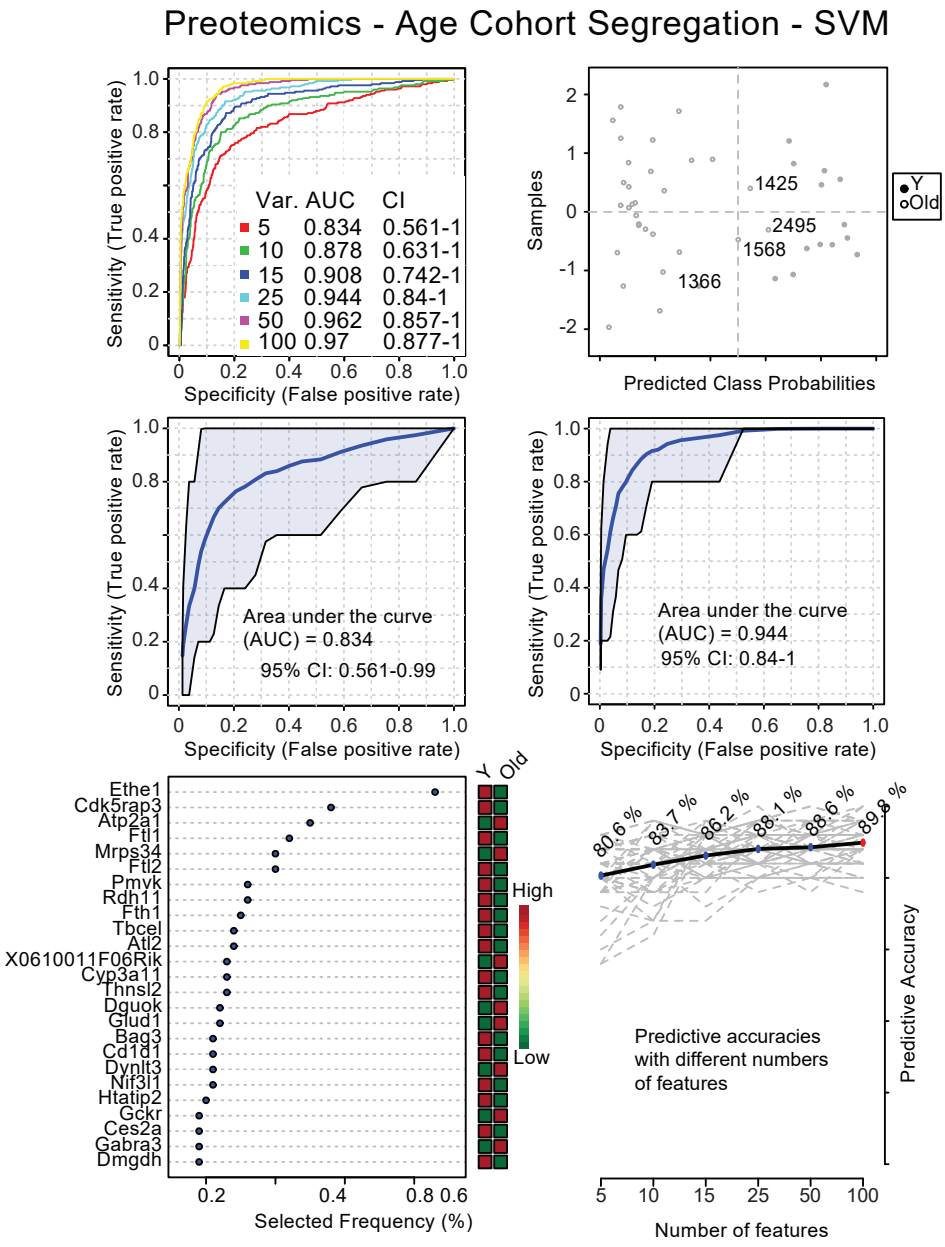


Figure 3.10: In the top left panel, the ROC curves for various model sizes are given, along with its AUC and the 95% confidence intervals. In the top right, the Dark and white dots are given in terms their classification probability. In the Middle row, the ROC curves for 5 and 25 protein are given. In the Bottom left, the most frequency chose factors are given. These factors are seen as prototypic for aging and will be followed up with in further analysis using transcript data. In the bottom right the accuracy for each model size is given

Chapter 4

Transcriptomics

4.1 Introduction to Micro-arrays

Preliminary transcript data from 99 mice was also produced using Affymetrix microarrays. These microarrays are silicon substrates that have a library of known cDNA fragments immobilized to their surface in discrete pixels (Miller and Tang, 2009). Experiments profiling the expression and relative abundance of a large set transcripts can be performed in a piece-wise manner. All of the transcripts in samples are functionalized with a fluorescent tag and introduced on to the surface of the microarray. When transcripts hybridize with complementary cDNA on the chip, the fluorescence intensity read off the chip using a charge-coupled device(CCD) camera is proportional to the concentration of the tagged transcript. Fluorescence standards on the chip seen in the red outline in figure 4.1 used to convert florescence throughout the chip into concentrations. Sometimes, the CCD camera can image certain parts the chip slightly lighter or dark, systematically biasing the measurements in segments of the photograph. The green box shows intensity normalization pixels that are placed throughout the chip in order to correct against irregularities

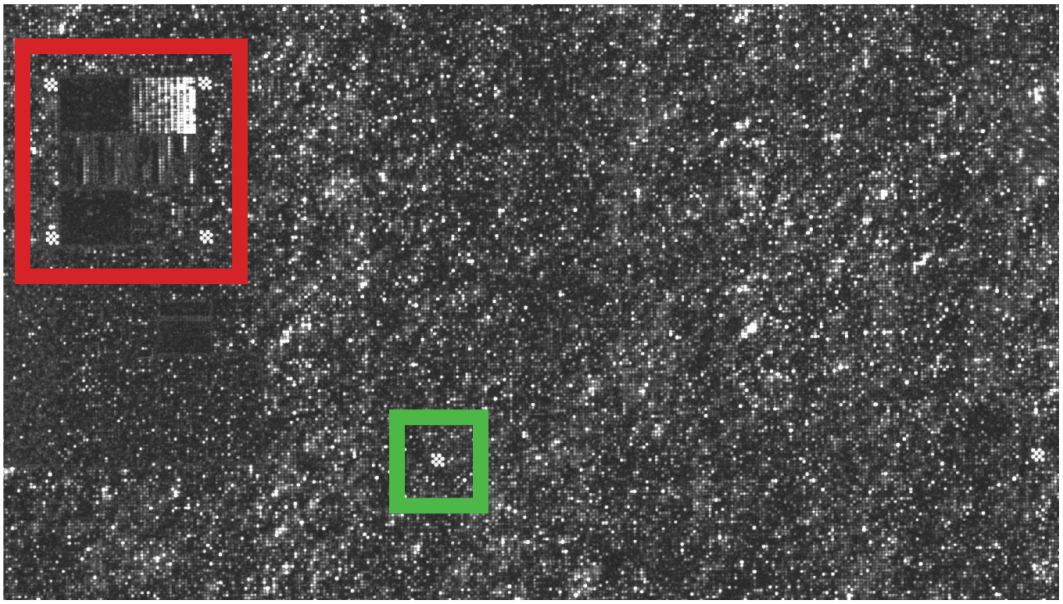


Figure 4.1: A zoomed in picture of a micro-array that can be processed to determine transcript levels. This is a one color array meaning single sample intensities are given. The red box is surrounding the intensity standards that are used to calibrate the intensity to transcript concentration ratio. The green box is around an intensity marker which is used to normalize the intensities across the chip irrespective defects that may capture different parts the chip with lower or higher intensities

in the CCD.

4.1.1 Transcriptomics Data Processing

Microarray experiments were not performed in-house. instead, the BXD mouse liver tissues were frozen in $N_(2l)$ pulverized to a powder in a motor and pestle and shipped to the United States for processing. At the facility, RNA is extracted from the mouse livers using the TRIzol Plus RNA Purification System (Rio et al., 2010)). It is converted to cDNA and fictionalized with a fluorescent tag before microarray analysis is done. The raw unprocessed DAT, CDF, and CEL files were returned for analysis using RMA express and R.

4.1.2 Data Extraction

Microarray expression data can be extracting with two files, the CDF and the CEL files. The CDF (Chip Description File) contains information about the probes on the gene array chip (Ben Bolstad, 2003). One CDF is required for all 600 mice in the full experiment because the layout of the probeset is the same for every mouse transcriptome being profiled. For each gene being profiled on the chip, there is a set of oligonucleotide probes 25 nucleotides long that are used to determine its transcript expression. This set contains both a perfect match probes that have the identical sequence of the gene and mismatch probes that have a substitution in the central nucleotide of the probe. As a technical control, mismatch controls should always show lower intensity than the perfect match probes and they sets are placed in different parts of the chip and quantify the amount of cross-hybridization that occurs from non-target genes (Coen, 2010). The CDF file also contains the location of control probes that can be used to correct of heterogeneities in the CMOS image detector across the gene array.

The pixel intensity values that are detected at each spot on the gene array are daved in an efficient .dat format (Miller and Tang, 2009). For an intial insepction, the .dat file can be open in an image view such as photoschop to insure positive control show high intensities through the chip and that there is no smears or local abberations in the light intesnity, possible introduced form mishandling of the chip. This format however, is difficult to work with and The .dat files can then be converted to .cel files a CEL file stores the results the intensity calculations on the pixel values that are extracted from a DAT file

This includes an intensity value, standard deviation of the intensity, the number pixels used to calculate the intensity value, a flag to indicate an outlier as calculated by the algorithm and a user defined flag indicating the feature should be excluded

from future analysis(Miller and Tang, 2009). The file stores this aforementioned data for each feature on the probe array.

RMAExpress reads in the pixel intensities and in the CEF files and the CDF files can return a table of normalized or raw intesnities for each gene. As an alternative to using R packages like affy to perform this, the simple graphical user interface is used load in the data and convieietnaly export it easy to use formats. In this study, RMA expressed is used to generate a csv table of all transcripts and their intensities. Normalization and model based background subtraction is done in R (Irizarry et al., 2003a,b; Bolstad et al., 2003).

4.1.3 Normalization

When running experiments that involve multiple high density oligonucleotide arrays, it is important to remove sources variation between arrays non-biological origin. Normalization is a process for reducing this variation. It is common to see non-linear relations between arrays and the standard normalization provided by Affymetrix does not perform well in these situationsBolstad et al. (2003).

Many traditional statical methodologies such as t-tests which will be performed on the microarray data afterwards are based on the assumption normally distribution or at least symmetrically distributed data, with constant variance. if the assumptions are violated . All the follow model based normalization techniques are adapted from (Durbin et al., 2002).

4.1.4 Model Based Error Subtraction

An error model is required to remove the non-biological noise from the system. in our model we assume

if : $x_k i$ is the true abundance the probe k in sample i

if : $y_k i$ is the measured intensity on the micro-array

then : $y_k i = a_k i + b_k i * x_k i$ If we assume true abundance is proportional to signal intensity

In the equation above, we assume the signal detected is a function the abundance the transcript in addition to noise that depends on the abundance and also noise that is independent or systematic noise. The parameter $b_k i$ summarizes abundance-dependent noise: which includes number cells, hybridization efficiency, label efficiency. The parameter $a_k i$ Summarized the abundance-independent noise. This noise can arise from unspecific hybridization, background florescences that may have been detected or stray signals.

If we assume only multiplicative noise in the linear model above and assume all the noise in the measure is derived from abundance-dependent noise.

$$Y_k i = a_k i + b_k i * x_k i$$

$$a_k i \approx 0$$

$$b_k i = b_i \cdot \beta$$

$$b_k i = b_i \beta_k (1 + \epsilon_k i)$$

The concentration dependent parameter $b_k i$ is then composed the sample specific noise which is described by b_i and the probe specific noise given by β_k. The remaining noise is modeled with a stochastic portion the model. In end the final model taking into account the multiplicative noise only can be given as

$$Y_k i = b_i \cdot \beta_k \cdot x_k i(1 + \epsilon_k i)$$

where $\epsilon_k i\ Norm(0, c^2)$ and c is the coefficient variation as defined by $c = \dfrac{std}{mean}$

With this formulation, if we want to determine the relative abundance a transcript with respect to another we can use the expression

$$M_k = \frac{Y_{k2}/Y_{k1}}{b_1/b_2}$$

In the more natural case we can assume the existence both multiplicative and additive noise and in this case our linear expression for determine the true abundance the transcript must be slightly altered. The additive noise term is also composed a systematic error term a_i and sample specific but abundance-interdependent term $b_i \eta_{ki}$.

$$Y_k i = a_k i + b_k i \cdot x_k i \qquad a_{ki} = a_i + b_i \eta_k i \qquad b_k i = b_i \beta_k (1 + \epsilon_k i) \tag{4.1}$$

this yields the final model given below:

$$\frac{Y_k i - a_i}{b_i} = b_i \beta_k^{\epsilon_k i} + \eta_k i$$

$$\text{where } \eta_k i\ N(0, c^2) \text{ and } \epsilon_k i\ N(0, s^2)$$

This equation allows us to model all the sources noise in the micro-array data from defined endogenous and exogenous sources in order for us to subtract it of, as indicated int he $Y_{ki} - a_i$ term.

4.1.5 Variance Stabilizing Normalization

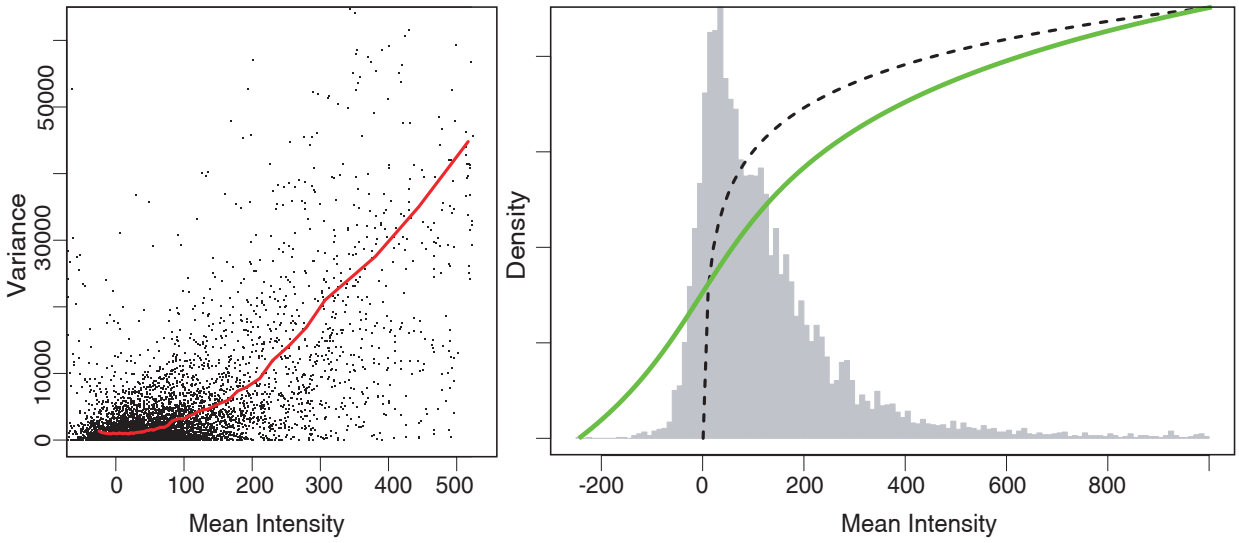


Figure 4.2: Adapted from (Durbin et al., 2002).(Left) The variance and mean relationship found in experimentally produced microarrray data. The red line shows a plot the v(u) described in at the end section 2.3.6(right) The variance stabilization transformation performed on the data. The green line represents a log transformation, and the dotted line the arcsinh transformation which is hte preferred transformation method

Although the background noise has been subtracted a variance stabilization still needs to be performed on the data. This is because the coefficient variation is not constant throughout the dataset. (Durbin et al., 2002). There is a quadratic relation between $v = var(Y_{ki}$ and $u = E(Y_{ki})$

$$v(u) = c^2(u - a_i)^2 + b_i^2 s^2$$

this relationship can be shown using empirical data. The figure below shows variance and expected values plotted against each-other illustrating the quadratic relationship. From visual inspection it seems a log transformation may benefit the micro array data, it is not obvious which transformation is optimal for stabilizing the variance in our data. (Durbin et al., 2002). Since the log transform is not defined for values under zero, values that become negative after the background subtraction

are not defined, forced us to throw out large swaths data. Moreover, a log transformation provides good variance stabilization at high levels, but inflate the variance close to the detection threshold. (Durbin et al., 2002). Therefore, an arcsinh transformation is used instead as it behaves like the log transformation asymptotically but is linear in the lowest intensity regions.

The variance stabilization transformation used with out micro-array data is then

$$h_i(y_{ki}) = arcsinh(\frac{c}{s}.\frac{y_{ki} - a_i}{b_i})$$

The final result this transformation is that intensities are normally distributed with a constant variation c^2 and a mean $b_i\beta_k$. Now if we would like to quantify differential expression we can use the expression

$$\Delta h_{k,ij} = h_i(y_k i) - h_j(y_{kj})$$

4.1.6 Parameter Estimation

In the end the final model used to perform the variance stabilization with out expression data is

$$arcsinh(\frac{y_{ki} - a_i}{b_i}) = b_i\beta_k + \epsilon_{ki}, \epsilon_{ki} \sim N(0, c^2)$$

In order to fit the parameters, one can use a maximum likelihood estimation. The model parameters can be fitted by using the majority genes unchanged assumption in which the sample specific noise parameters can be more or less assumed to be the same across all transcripts.

$$b_i\beta_k = b\beta_k$$

$$arcsinh(\frac{y_{ki} - a_i}{b_i}) = b_i\beta_k + \epsilon_{ki}, \epsilon_{ki} \sim N(0, c^2)$$

4.2 Quality Control

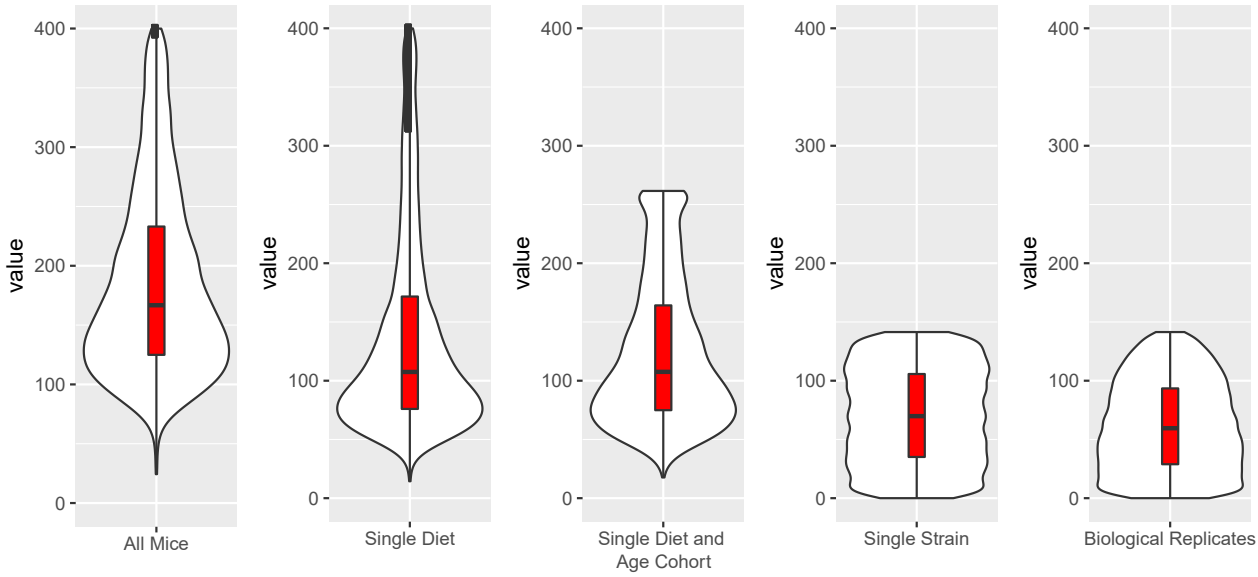


Figure 4.3: CV Analysis of Transcript Data

The coefficients of variation in the RNA data are as much larger than those seen in the protein and metabolites. This due to the many order of magnitude over which different transcripts can be produced. The biological replicates themselves show nearly a 100% coefficient of variation indicating large fold changes are needed to reliably detect differences in the transcription between mouse cohorts.

4.2.1 Tissue Contamination

One of the biggest challenges with analyzing transcriptomics data from organs or samples containing more than one cell type is the effect of variations that come from changing proportions of cell types within each sample. Although liver is a

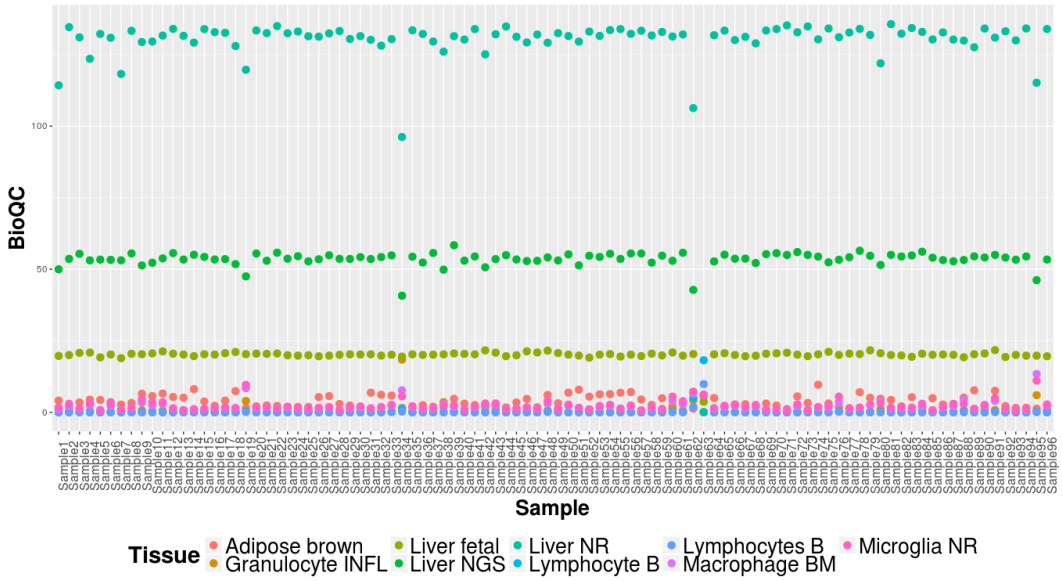


Figure 4.4: BioQC data showing the organ classes the BXD mouse samples are mostly to come from. Most sample show high liver signatures, however there is evidence for a high degree of immune cell infiltration in many samples

large relatively homogeneous organ, overwhelmingly made up of hepatocytes, there are also a range of endothelial cells which make up parts of the vasculature that supply blood to and from the liver, it and a host of invading immune cells that are present in almost all tissues irrespective of the inflammation state of the animal. All of these cells that influence the gene expression patterns independent of the effects being studied and may lead to a perceived enrichment of certain gene pathways, when we are simply assay different cells. This is a complication encountered in this study because specific lobes of the liver are not always used for the same omics analysis, instead, the livers are ground in liquid nitrogen, with the resulting powder being partitioned into many aliquots with slightly different portions of each lobe.

BioQC is a package developed at Roche that quantifies samples based on their resemblances to organs and provides a quantitative method with which to determine the context of samples and their comparability (Zhang et al., 2017). This package uses manually curated gene signatures that are prototypic for 150 tissues types or

organs, and in some cases, a an organ in a single development state (embryonic and mature kidney have varying transcript levels in many genes) which are used determine organ scores for each sample. This method has the distinct advantage over unsupervised learning techniques like PCA which can be used to visualize and manually inspect the distance between two samples but does not show the source of the contamination (Zhang et al., 2017). As can be seen in figure 4.4 in the BXD mouse samples, all of the liver samples have high mature liver scores above 80 followed by intermediate embryonic liver scores and lower scores for adipose tissue and monocytes which make up minority constituents of the samples. The two mature liver scores were given, Liver NGS, which is derived from the RNAseq Atlas(Krupp et al., 2012)) and Liver NR, derived from an in-house Roche liver panel, show similar results with concomitant drops in scores for samples 18, 33, 62 and 80. The Liver NGS score is more specific to humans and gives lower scores as compared to liver NR.

Overall, one can conclude a low amount of contamination for the expectation of four samples in the first 96. In the heat-map visualizing the same data(figure 4.5), one can see the large areas of dark blue indicating low BioQC scores for cell types such are brown fat and microglia. One sample, 62 does not have any organ enrichment score, indicating either technical error. The other samples can be roughly grouped into two categories of liver cells. The minority group which for lower liver scores using the Roche and NGS panel on average and have high scores for immune cells like macrophages, granulocytes, and lymphocytes. Immune cells surveillance is well documented in the liver, as portal circulation bring disgested material directly to the liver where it can act as a frontline defense for materials entering circulation to the rest of the body(Jenne and Kubes, 2013). These mice most likely have an infection or have eaten something that contained immunogenic material. The rest of the samples show very high liver scores, meaning there has high expression of genes that are expressed in normal human liver.

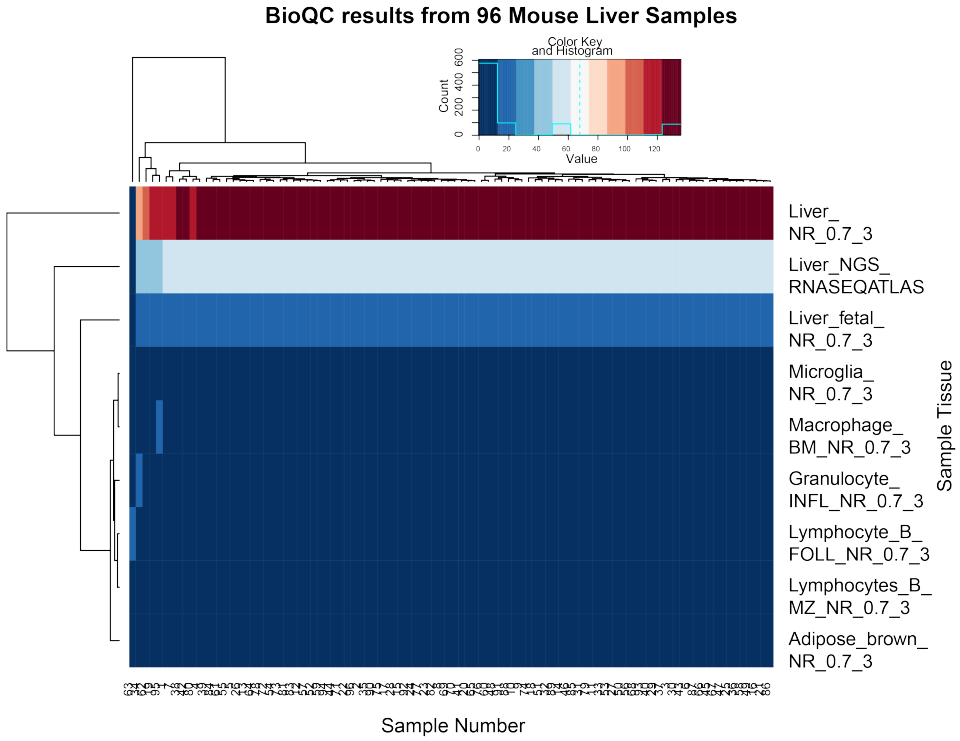


Figure 4.5: Heatmap of BioQC scores for the 96 samples. Most of the samples show high BioQC scores indicating, genes known to be expressed selectively in liver are detected in most of samples. BioQC scores for immune cells are also detected, indicating the presence of infiltrating cell in the samples as well

4.3 Transcript Level Fold Changes

In a dramatic reversal from the proteomic and metabolomic data, the transcript data is highly enriched with differentially expressed transcripts between older and younger mice. Much like the proteomic data, too many hits are found using the a of 0.05 threshold for significance and a 1 fold log fold2 change. Such a large set of traits can be difficult to follow up with literature reviews and QTL analysis. As such all of the transcripts enriched in the old mice, are added to networks previously used to investigate metabolites and protein data. The goal for this integrated analysis is to determine causal circuits for the certain metabolite concentrations in the liver.

As an initial look at the relationship between the three layers information we have,

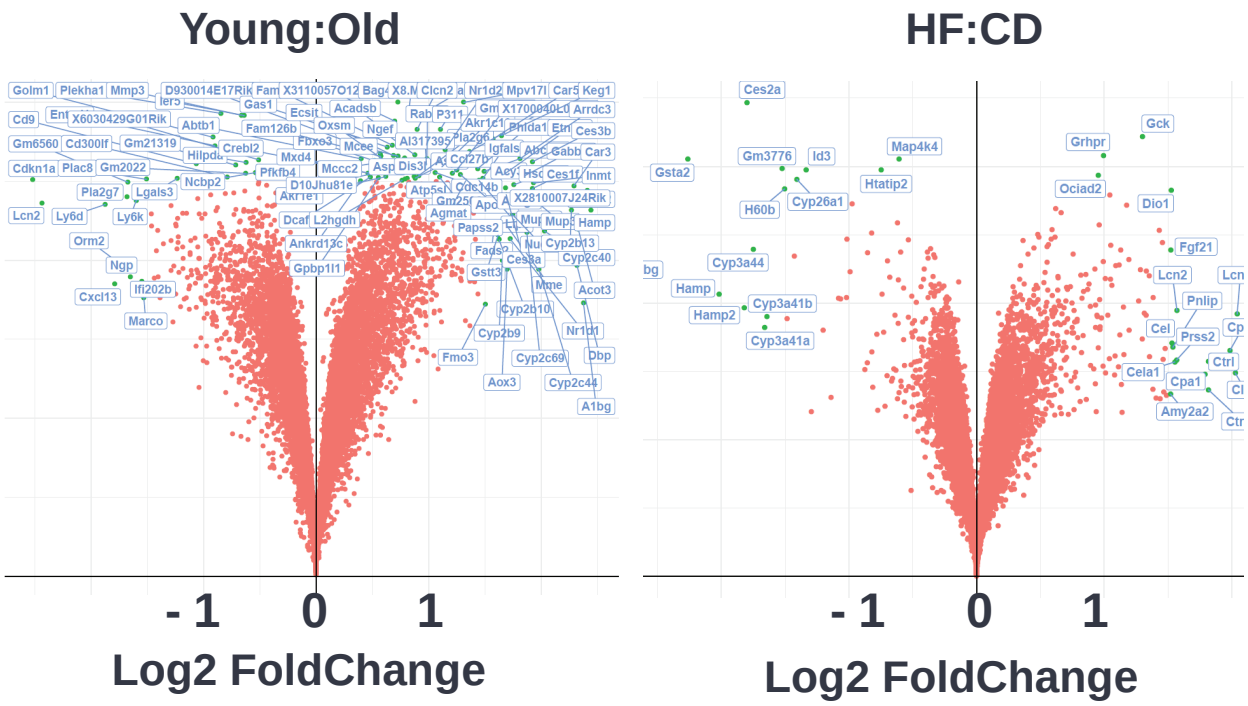


Figure 4.6: Volcano Plots Diet and Age Cohort Fold Changes

one can look at the correlation between proteins and transcripts and realize that they are not very tightly coupled. For the few very highly correlated sets protein and transcripts, many are metabolic genes such as hexokinase. From this one can conclude that even though not all pairs protein and transcript will be instructive and add an explanatory information for certain metabolites levels seen, they may be in a small subset of metabolites.

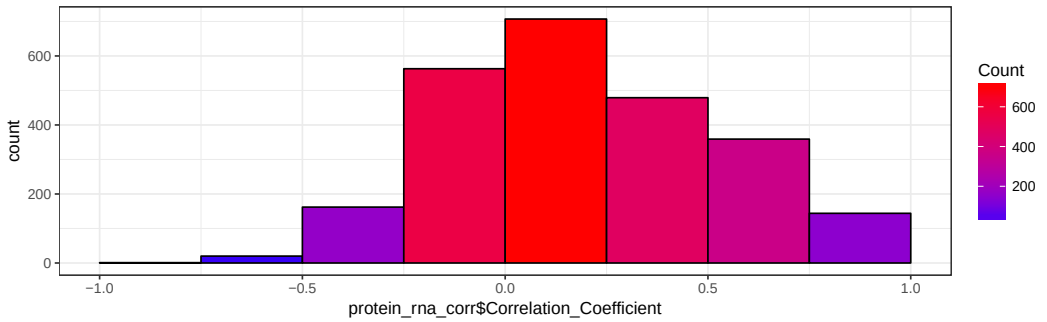


Figure 4.7: Correlation Coefficients between protein and RNA

4.4 Gene Set Enrichment Analysis

4.5 Multi-Omics Data Integration

4.5.1 Integrated Analysis of Linoleic acid Oxidation

As an initial proof concept to show that proteins, metabolite and transcript data could be used together to generate causal networks, a small part of the Linoleic acid metabolism was modeled (figure 4.9) In the first reaction phosphotidyl choline is cleaved by Pla2g12b into linoleic acid. The metabolites, protein and transcript are all similarly expressed. When the fatty acid has become linoleic acid it can be converted by various lipoxygenases and cytochrome p450 enzymes into either 12,13-epoxide, Vernolic acid or the 9,10-epoxide, Coronaric acid. The regulation of this circuit is known in *Vernicia fordii)* where the genes that increase production of the oil experience product repression (Li et al., 2010). Data from the effects of these oils in terms of its substrate or product interactions with proteins in mammal is not know .The Cyp genes performing this latter transformations are down-regulated 40% in high fat diet mice compared to the chow diet mice and the transcript is unregulated which is in contrast to findings elsewhere on the effect of high fat diets on cytochrome proteins, drug transporters (Ghose et al., 2011). The lower amount

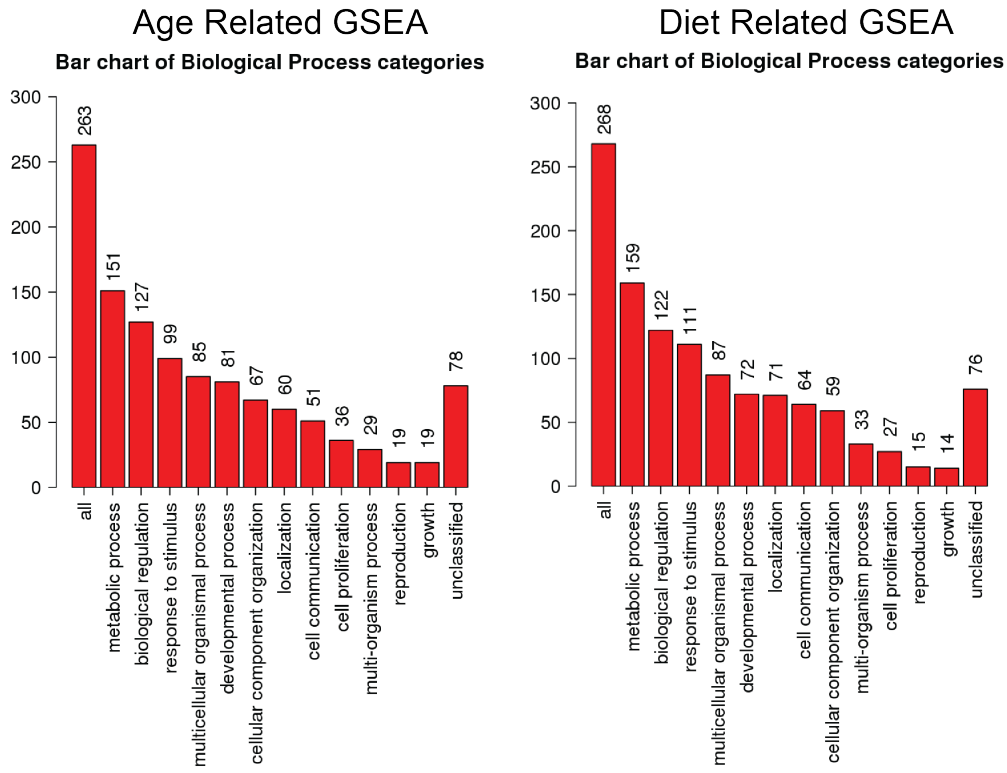

Age Related GSEA
Bar chart of Biological Process categories
300
250
200
150
100
50
0
263
151
127
99
85
81
67
60
51
36
29
19
19
78
all
metabolic process
biological regulation
response to stimulus
multicellular organismal process
developmental process
cellular component organization
localization
cell communication
cell proliferation
multi-organism process
reproduction
growth
unclassified
Diet Related GSEA
Bar chart of Biological Process categories
300
250
200
150
100
50
0
268
159
122
111
87
72
71
64
59
33
27
15
14
76
all
metabolic process
biological regulation
response to stimulus
multicellular organismal process
developmental process
localization
cell communication
cellular component organization
multi-organism process
cell proliferation
reproduction
growth
unclassified

Figure 4.8

protein may be due to the accumulation the 9-10-epoxide. Although in this toy example, one cannot definitively say that the 9-10-epoxides is down-regulating its catabolic enzymes, as in a trp repressor type circuit, it is only a hypothesis, one can test once all the layers molecular data are available.

The metabolite fold changes overall in the pathway however are not large enough to rule out chance fluctuations in the relative abundances of any of the biomolecules. Even though there is a 5 fold enrichment of Pla2g12b in the HF diet mice, there seems to be no large changes in the ratio of PC(18:0) and Linoleic acid. Another reason for the lack of difference between up and downstream metabolites in this

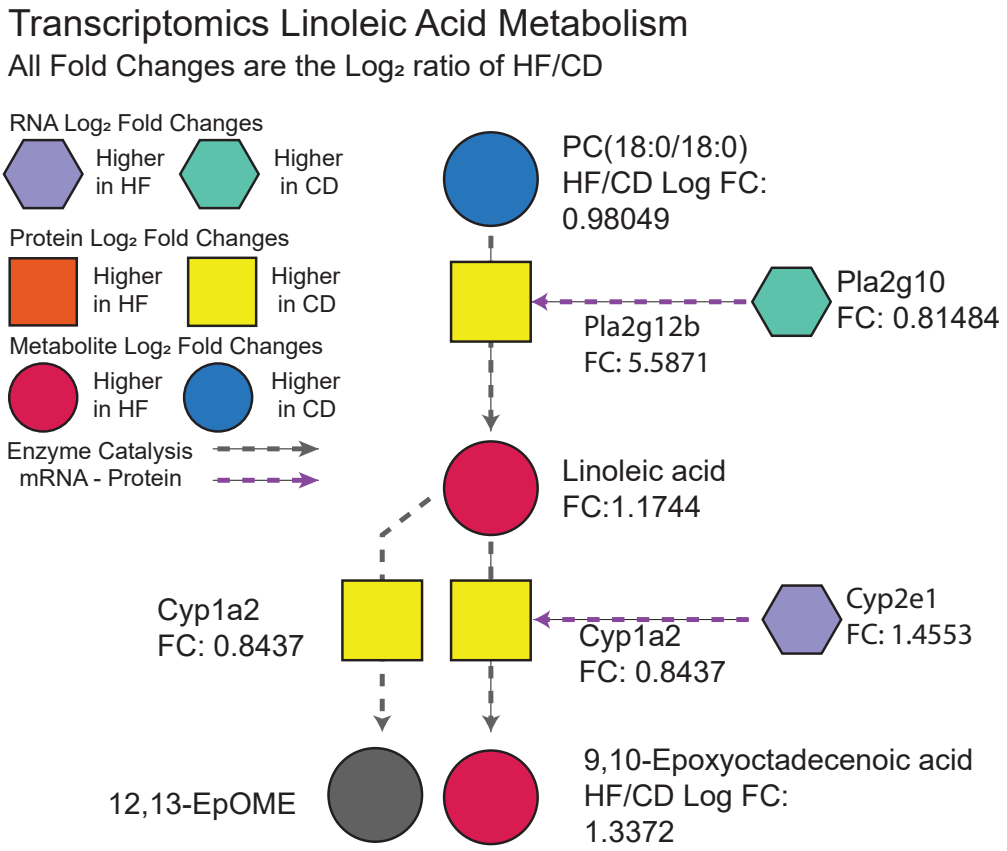


Figure 4.9: Integrated analysis of Linoleic Acid Oxidation

pathways is that all three of metabolites have multiple isobars. In this case it is impossible to determine the exact location of the fatty unsaturations which are important in determining the function of metabolites. This and the exclusion of shunt pathways in this rarefied models make it difficult determine the exact reason the large enzymatic fold change does not translate into a large flux though the pathway (Saa and Nielsen, 2016).

4.5.2 Integrated Analysis of Tyrosine Metabolism

In a slightly more complex example, we can consider the tryptophan metabolism which had was a significantly effect pathway in the metabolite data, pathway enzymes were unregulated in high fat diet in proteomic analysis and transcripts related to those proteins were hits in stringent fold change analysis for transcripts. The central metabolite, L-tryptophan is highlighted in the graph with a deep purple boarders cannot be produced endogenously and needs to be obtained from diet. From the central L-tryptophan amino acid, tryptamine, Oxitiptan or L-Formylknynurenie can be formed shuttling the metabolite into three distinct directions, one of which is the production of serotonin and melatonin. Serotonin is synthesized via tryptophan hydroxylase, which converts 5-hydroxytryptophan (5-HTP) into serotonin. (Schaechter and Wurtman, 1990; Fernstrom, 1983) . Melatonin is generated from serotonin, via N-acetyltransferase and 5-hydroxyindole-O-methyltransferase activity (WURTMAN and ANTON-TAY, 1969). Despite this, it is difficult to alter the level of tryptophan in the tissue (and as a result the flux of this pathway) from changing an animals diet as it is brought in by transporters that also transport other competing amino acids (Robinson et al., 2009). This may be the reason, the integrated analysis was not able to find genetic factors that control the liver tryptophan concentrations. If the effects of the transporters on the concentrations of tryptophan and its cognate metabolites is much larger than any of the effects of mutations in

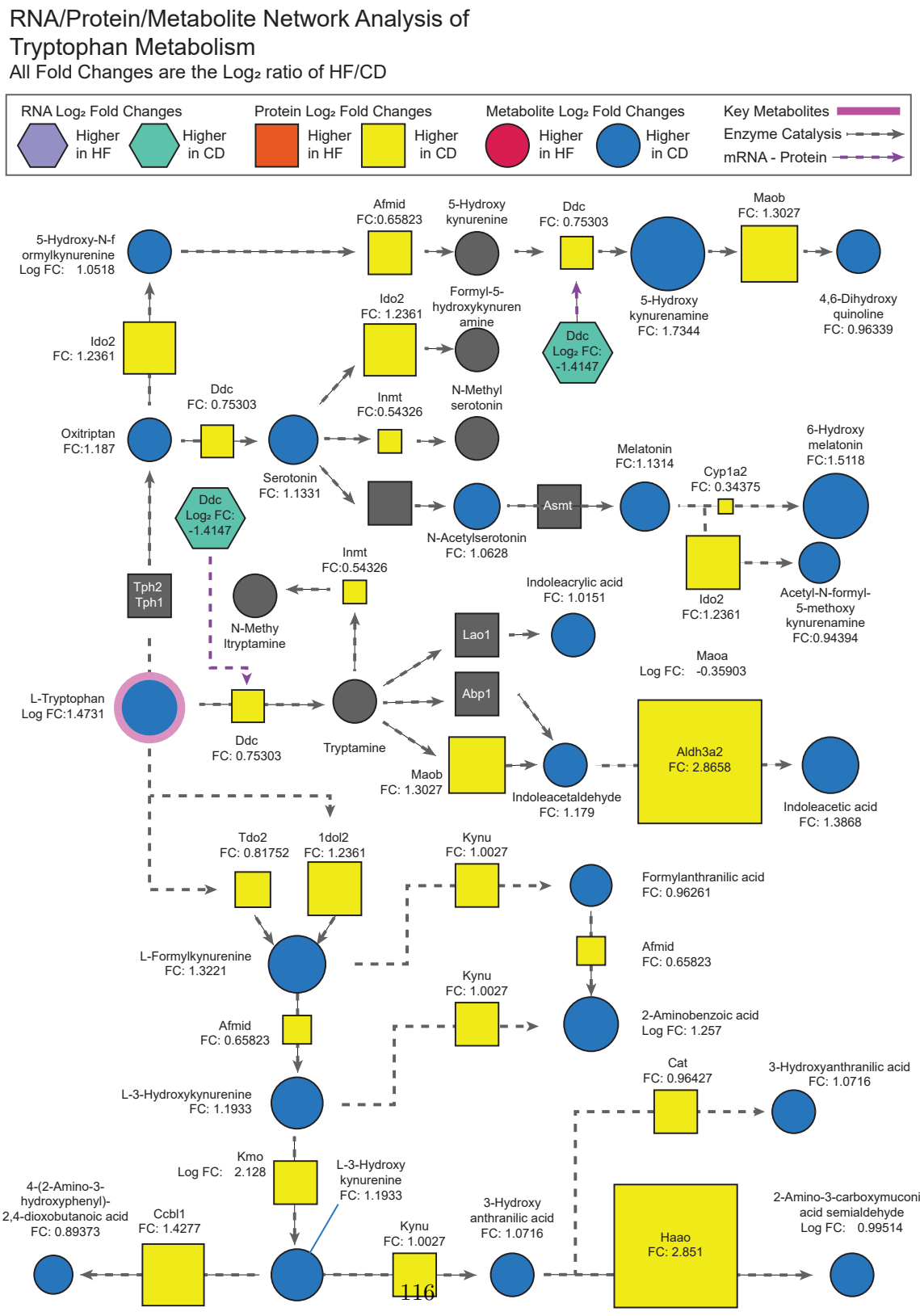


Figure 4.10: Integrated analysis Tryptophan metabolism

the enzymatic pathways, it may not be possible to determine.

Despite the effort th tat has gone into reconstructing the pathway, including significantly enriched transcripts, proteins and metabolites in the tryptophan metabolism, definitive conclusion about the direction and nature of interaction between the layers of biomolecules could not be reached. The issue in this circuit is that the networks fold changes are highly counter intuitive. One partly interpretable section this network is the reaction, in which melatonin is produced though the reduction of 6-hydroxy-melatonin by a Cyp enzymes. In this case the lower levels the enzyme and high levels of the product can be interpreted in two ways. Either the production the enzymes is being inhibited by the excess product (which it is not, no specific decease in the transcript between the diets can be detected) or that the cyp gene is expressed lower on average in the high fat mice and the 6-hydroxy melatonin is building up due to the low expression of the catalyzing protein.

For most of the other graph, the inability to put discrete barriers on the reactions thus does not allow us to accurately predict what the causal structure. Moreover, imposing such barrier would be helpful as a theoretical exercise but would not represent the true intermingled nature the metabolism. Moreover, one may look at the large boxes drawn for Ddc, a high over expressed transcript in HF mice, Aldh3a2 or Haao, another two proteins that are also highly epxressed in the HF mice, don't actually seem to have an effect on thier downstream metabolite molecules.

4.6 Transcripts Conclusion

A large number differentially expressed transcripts in aging mice were found, in contrast to the proteomic data in which there was almost no differences between the age cohorts. Despite having only 22 proteomic samples and 90 transcriptomics samples, detailed descriptions of metabolic pathways with protein and transcript

nodes overlaid were generated. The large amount information generating a network diagram can condense is very convenient for visualizing inputs and outputs of pathways. The construction the aforementioned networks however does not allow for easy structural hypothesis testing, as the boundary conditions for the networks are not well defined.

Chapter 5

Future work

In this thesis, a large amount metabolic, proteomic and transcriptomic data was collected on 88 strains of BXD mice. QTL analysis of metabolites elucidated three strong QTLs, one for pyruvate, hydroxbutaric acid and glutathione of which glutathione looks promising as a potential target to follow up with using glutathione oxidase inhibitors or recombinant mice with mutant alleles for this gene.

The preliminary SWATH results indicate some promising follow up proteins. Once the full SWATH MS and transcriptomics data is produced, most thorough network analysis can be done to determine pathways adversely effected by aging.

5.1 References

]

Bibliography

Abiola, O., Angel, J. M., Avner, P., Bachmanov, A. A., Belknap, J. K., Bennett, B., Blankenhorn, E. P., Blizard, D. A., Bolivar, V., Brockmann, G. A., Buck, K. J., Bureau, J.-F., Casley, W. L., Chesler, E. J., Cheverud, J. M., Churchill, G. A., Cook, M., Crabbe, J. C., Crusio, W. E., Darvasi, A., de Haan, G., Dermant, P., Doerge, R. W., Elliot, R. W., Farber, C. R., Flaherty, L., Flint, J., Gershenfeld, H., Gibson, J. P., Gu, J., Gu, W., Himmelbauer, H., Hitzemann, R., Hsu, H.-C., Hunter, K., Iraqi, F. F., Jansen, R. C., Johnson, T. E., Jones, B. C., Kempermann, G., Lammert, F., Lu, L., Manly, K. F., Matthews, D. B., Medrano, J. F., Mehrabian, M., Mittlemann, G., Mock, B. A., Mogil, J. S., Montagutelli, X., Morahan, G., Mountz, J. D., Nagase, H., Nowakowski, R. S., O'Hara, B. F., Osadchuk, A. V., Paigen, B., Palmer, A. A., Peirce, J. L., Pomp, D., Rosemann, M., Rosen, G. D., Schalkwyk, L. C., Seltzer, Z., Settle, S., Shimomura, K., Shou, S., Sikela, J. M., Siracusa, L. D., Spearow, J. L., Teuscher, C., Threadgill, D. W., Toth, L. A., Toye, A. A., Vadasz, C., Van Zant, G., Wakeland, E., Williams, R. W., Zhang, H.-G., Zou, F., and Complex Trait Consortium (2003). The nature and identification of quantitative trait loci: a community's view. *Nature reviews. Genetics*, 4(11):911–6.

Agilent Technologies (2017). Agilent — 6550 iFunnel Q-TOF LC/MS system video.

Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P., and Dorrestein, P. C. (2017).

Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry*, 1(7):s41570–017.

Andreux, P., Williams, E., Koutnikova, H., Houtkooper, R., Champy, M.-F., Henry, H., Schoonjans, K., Williams, R., and Auwerx, J. (2012). Systems Genetics of Metabolism: The Use of the BXD Murine Reference Panel for Multiscalar Integration of Traits. *Cell*, 150(6):1287–1299.

Anton, G., Wilson, R., Yu, Z.-h., Prehn, C., Zukunft, S., Adamski, J., Heier, M., Meisinger, C., Römisch-Margl, W., Wang-Sattler, R., Hveem, K., Wolfenbuttel, B., Peters, A., Kastenmüller, G., and Waldenberger, M. (2015). Pre-Analytical Sample Quality: Metabolite Ratios as an Intrinsic Marker for Prolonged Room Temperature Exposure of Serum Samples. *PLOS ONE*, 10(3):e0121495.

Armanios, M., de Cabo, R., Mannick, J., Partridge, L., van Deursen, J., and Villeda, S. (2015). Translational strategies in aging and age-related disease. *Nature Medicine*, 21(12):1395–1399.

Bailey, A. J., Paul, R. G., and Knott, L. (1998). Mechanisms of maturation and ageing of collagen. *Mechanisms of ageing and development*, 106(1-2):1–56.

Bateman, N. W., Goulding, S. P., Shulman, N. J., Gadok, A. K., Szumlinski, K. K., MacCoss, M. J., and Wu, C. C. (2014). Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). *Molecular & cellular proteomics : MCP*, 13(1):329–38.

Beard, D. A., Liang, S.-d., Qian, H., Palsson, B., Barabási, A.-L., and Tramper, J. (2002). Energy balance for analysis of complex metabolic networks. *Biophysical journal*, 83(1):79–86.

Becker, D. J. and Lowe, J. B. (2003). Fucose: biosynthesis and biological function in mammals. *Glycobiology*, 13(7):41R–53R.

Ben Bolstad, R. (2003). Software for affy data analysis.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–93.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Broman, K. W. and Sen, S. (2009). A guide to qtl mapping with r/qtl. volume 46. Springer.

Bruderer, R., Bernhardt, O. M., Gandhi, T., and Reiter, L. (2016). High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics*, 16(15-16):2246–56.

Caligioni, C. S. (2009). Assessing reproductive status/stages in mice. *Current protocols in neuroscience*, Appendix 4:Appendix 4I.

Cani, P., Delzenne, N., Cloarec, O., Coen, M., Tang, H., Gieger, C., Chang, D., Milburn, M. V., Gall, W. E., Weinberger, K. M., Mewes, H.-W., de Angelis, M. H., Wichmann, H.-E., Kronenberg, F., Adamski, J., and Illig, T. (2009). The role of the gut microbiota in energy metabolism and metabolic disease. *Curr Pharm Des*, 15(11):e13953.

Civelek, M. and Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature reviews. Genetics*, 15(1):34–48.

Coen, M. (2010). A metabonomic approach for mechanistic exploration of pre-clinical toxicology. *Toxicology*, 278(3):326–40.

Collaborative Cross Consortium, C. C. (2012). The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics*, 190(2):389–401.

Cortez, P. and Maintainer, . (2016). Data Mining Classification and Regression Methods.

Czyz, W., Morahan, J. M., Ebers, G. C., and Ramagopalan, S. V. (2012). Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. *BMC medicine*, 10:93.

De Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217):1251.

Domalain, V., Hubert-Roux, M., Tognetti, V., Joubert, L., Lange, C. M., Rouden, J., and Afonso, C. (2014). Enantiomeric differentiation of aromatic amino acids using traveling wave ion mobility-mass spectrometry. *Chemical Science*, 5(8):3234–3239.

Dumas, M.-E. (2012). Metabolome 2.0: quantitative genetics and network biology of metabolic phenotypes. *Molecular BioSystems*, 8(10):2494.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *BIOINFORMATICS*, 18(1):105–110.

Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207.

Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS*, 13(1):22–24.

Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M. J., and Rinner, O. (2012). Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics*, 12(8):1111–21.

Fakhro, K. A., Staudt, M. R., Ramstetter, M. D., Robay, A., Malek, J. A., Badii, R., Al-Marri, A. A.-N., Khalil, C. A., Al-Shakaki, A., Chidiac, O., Stadler, D., Zirie, M., Jayyousi, A., Salit, J., Mezey, J. G., Crystal, R. G., and Rodriguez-Flores, J. L. (2016). The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Human Genome Variation*, 3:16016.

Falconer, D. S. and Mackay, T. F. C. (1996a). Introduction to quantitative genetics.

Falconer, D. S. D. S. and Mackay, T. F. C. (1996b). *Introduction to quantitative genetics*. Longman.

Fang, M., Ivanisevic, J., Benton, H. P., Johnson, C. H., Patti, G. J., Hoang, L. T., Uritboonthai, W., Kurczy, M. E., and Siuzdak, G. (2015). Thermal Degradation of Small Molecules: A Global Metabolomic Investigation. *Analytical Chemistry*, 87(21):10935–10941.

Fernstrom, J. D. (1983). Role of precursor availability in control of monoamine biosynthesis in brain. *Physiological reviews*, 63(2):484–546.

FGCZ (2017).

Gerhard Adam (2012). What Is Heritability?

Ghose, R., Omoluabi, O., Gandhi, A., Shah, P., Strohacker, K., Carpenter, K. C., McFarlin, B., and Guo, T. (2011). Role of high-fat diet in regulation of gene expression of drug metabolizing enzymes and transporters. *Life sciences*, 89(1-2):57–64.

Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*, 11(6):O111–016717.

Glish, G. L. and Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. *Nature reviews. Drug discovery*, 2(2):140.

Goncalves, R. L., Bunik, V. I., and Brand, M. D. (2016). Production of superoxide/hydrogen peroxide by the mitochondrial 2-oxoadipate dehydrogenase complex. *Free Radical Biology and Medicine*, 91:247–255.

Goodner, K. L., Milgram, K. E., Williams, K. R., Watson, C. H., and Eyler, J. R. (1998). Quantitation of ion abundances in fourier transform ion cyclotron resonance mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 9(11):1204–1212.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian journal of internal medicine*, 4(2):627–35.

Hartwell, L. H., Hopfield, J. J., Leibler, S., Murray, A. W., et al. (1999). From molecular to modular cell biology. *Nature*, 402(6761):C47.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Random Forests. pages 587–604. Springer, New York, NY.

Haynes, C. A., Allegood, J. C., Park, H., and Sullards, M. C. (2009). Sphingolipidomics: methods for the comprehensive analysis of sphingolipids. *Journal of Chromatography B*, 877(26):2696–2708.

Higashino, K., Tsukada, K., and Lieberman, I. (1965). Saccharopine, a product of lysine breakdown by mammalian liver. *Biochemical and Biophysical Research Communications*, 20(3):285–290.

Hu, C., van der Heijden, R., Wang, M., van der Greef, J., Hankemeier, T., and Xu, G. (2009). Analytical strategies in lipidomics and applications in disease biomarker discovery. *Journal of Chromatography B*, 877(26):2836–2846.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4):e15.

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

Jenne, C. N. and Kubes, P. (2013). Immune surveillance by the liver. *Nature Immunology*, 14(10):996–1006.

Joe Pinsker (2013). Why We Live 40 Years Longer Today Than We Did in 1880 - The Atlantic.

Johnson, A. D. and O'Donnell, C. J. (2009). An open access database of genome-wide association results. *BMC medical genetics*, 10:6.

Johnson, C. H., Ivanisevic, J., and Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–459.

Kafri, R., Levy, M., and Pilpel, Y. (2006). The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 103(31):11653–8.

Kalueff, A., Minasyan, A., Keisala, T., Shah, Z., and Tuohimaa, P. (2006). Hair barbering in mice: implications for neurobehavioural research. *Behavioural processes*, 71(1):8–15.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361.

Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008). Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536.

Kissen, R., Øverby, A., Winge, P., and Bones, A. M. (2016). Allyl-isothiocyanate treatment induces a complex transcriptional reprogramming including heat stress, oxidative stress and plant defence responses in Arabidopsis thaliana. *BMC Genomics*, 17(1):740.

Korte, A. and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9:29.

Krupp, M., Marquardt, J. U., Sahin, U., Galle, P. R., Castle, J., and Teufel, A. (2012). RNA-Seq Atlasa reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, 28(8):1184–1185.

Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., and Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nature methods*, 5(10):873–875.

Lange, V., Picotti, P., Domon, B., and Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology*, 4:222.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739.

Li, R., Yu, K., Hatanaka, T., and Hildebrand, D. F. (2010). Vernonia DGATs increase accumulation of epoxy fatty acids in oil. *Plant Biotechnology Journal*, 8(2):184–195.

Liaw, A., Wiener, M., and Andy Liaw, M. (2015). Breiman and Cutler's Random Forests for Classification and Regression Description Classification and regression based on a forest of trees using random inputs.

Liu, Y., Hüttenhain, R., Surinova, S., Gillet, L. C., Mouritsen, J., Brunner, R., Navarro, P., and Aebersold, R. (2013). Quantitative measurements of ¡i¿N¡/i¿ -linked glycoproteins in human plasma by SWATH-MS. *PROTEOMICS*, 13(8):1247–1256.

Lorenz, A. J., Hamblin, M. T., and Jannink, J.-L. (2010). Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PloS one*, 5(11):e14079.

Mackay, T. F. C., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565–577.

Mackenbach, J. P. (2007). Sanitation: pragmatism works. *BMJ*, 334(suppl_1):s17–s17.

Melinda B. Tierney and Kurt H. Lamour (2005). An Introduction to Reverse Genetic Tools for Investigating Gene Function.

Miller, M. B. and Tang, Y.-W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, 22(4):611–33.

Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., and Prunotto, M. (2017). Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery*, 16(8):531–543.

Mulligan, M. K., Mozhui, K., Prins, P., and Williams, R. W. (2017). GeneNetwork: A Toolbox for Systems Genetics. In *Methods in molecular biology (Clifton, N.J.)*, volume 1488, pages 75–120.

Mushtaq, M. Y., Choi, Y. H., Verpoorte, R., and Wilson, E. G. (2014). Extraction for Metabolomics: Access to The Metabolome. *Phytochemical Analysis*, 25(4):291–306.

Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology*, 7:548.

Paltiel, L., Rønningen, K. S., Meltzer, H. M., Baker, S. V., and Hoppin, J. A. (2008). Evaluation of Freeze Thaw Cycles on stored plasma in the Biobank of the Norwegian Mother and Child Cohort Study. *Cell preservation technology*, 6(3):223–230.

Peter Kelmenson (2015). Oh no, my mice are balding!

Picotti, P. and Aebersold, R. (2012). Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature methods*, 9(6):555–566.

Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009). Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & Cellular Proteomics*, 8(11):2405–2417.

Richards, A. L., Merrill, A. E., and Coon, J. J. (2015). Proteome sequencing goes deep. *Current opinion in chemical biology*, 24:11–7.

Rio, D. C., Ares, M., Hannon, G. J., and Nilsen, T. W. (2010). Purification of RNA using TRIzol (TRI reagent). *Cold Spring Harbor protocols*, 2010(6):pdb.prot5439.

Robinson, O. J., Sahakian, B. J., Stoy, N., Egerton, M., Christofides, J., Stone, T., Darlington, L., Diksic, M., Houle, S., and Meyer, J. H. (2009). Acute tryptophan depletion evokes negative mood in healthy females who have previously experienced concurrent negative mood and tryptophan depletion. *Psychopharmacology*, 205(2):227–235.

Roepstorff, P. and Fohlman, J. (1984). Letter to the editors. *Biological Mass Spectrometry*, 11(11):601–601.

Röst, H. L., Aebersold, R., and Schubert, O. T. (2017). Automated SWATH Data Analysis Using Targeted Extraction of Ion Chromatograms. In *Methods in molecular biology (Clifton, N.J.)*, volume 1550, pages 289–307.

Röst, H. L., Aebersold, R., and Schubert, O. T. (2017). Automated swath data analysis using targeted extraction of ion chromatograms. *Proteomics: Methods and Protocols*, pages 289–307.

Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L., and Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology*, 32(3):219–223.

Saa, P. A. and Nielsen, L. K. (2016). Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. 6:29635.

Schaechter, J. D. and Wurtman, R. J. (1990). Serotonin release varies with brain tryptophan levels. *Brain Research*, 532(1-2):203–210.

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11:213.

Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., Lam, H., Amodei, D., Mallick, P., MacLean, B., and Aebersold, R. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*, 10(3):426–441.

Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G., and Aebersold, R. (2017). Quantitative proteomics: challenges and opportunities in basic and applied research. *Nature Protocols*, 12(7):1289–1294.

Sell, D. R., Strauch, C. M., Shen, W., and Monnier, V. M. (2007). 2-aminoadipic acid is a marker of protein carbonyl oxidation in the aging human skin: effects of diabetes, renal failure and sepsis. *The Biochemical journal*, 404(2):269–77.

Shawe-Taylor, J. and Sun, S. (2011). A review of optimization methodologies in support vector machines. *Neurocomputing*, 74(17):3609–3618.

Shimizu, S. (2014). Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98.

Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I. (2011). iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Molecular & cellular proteomics : MCP*, 10(12):M111.007690.

Smith, O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. *Ther Drug Monit*, 27.

Solier, C. and Langen, H. (2014). Antibody-based proteomics and biomarker research-Current status and limitations. *PROTEOMICS*, 14(6):774–783.

Song, Y.-Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130–5.

Stadtman, E. R. (2004). Role of oxidant species in aging. *Current medicinal chemistry*, 11(9):1105–12.

Stanford Medicine (2017). Blood tests — stanford health care.

Steen, H. and Mann, M. (2004). The abc's (and xyz's) of peptide sequencing. *Nature reviews. Molecular cell biology*, 5(9):699.

Telenti, A., Pierce, L. C. T., Biggs, W. H., di Iulio, J., Wong, E. H. M., Fabani, M. M., Kirkness, E. F., Moustafa, A., Shah, N., Xie, C., Brewerton, S. C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B. A., Och, F. J., Turpaz, Y., and Venter, J. C. (2016). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42):11901–11906.

Traut, T. W. (1994). Physiological concentrations of purines and pyrimidines. *Molecular and cellular biochemistry*, 140(1):1–22.

Välikangas, T., Suomi, T., Elo, L. L., ML, B., YD, C., AN, D., and ER, D. (2016). A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics*, 86:bbw095.

van Eunen, K., Bouwman, J., Daran-Lapujade, P., Postmus, J., Canelas, A. B., Mensonides, F. I. C., Orij, R., Tuzun, I., van den Brink, J., Smits, G. J., van Gulik, W. M., Brul, S., Heijnen, J. J., de Winde, J. H., de Mattos, M. J. T., Kettner, C., Nielsen, J., Westerhoff, H. V., and Bakker, B. M. (2010). Measuring enzyme activities under standardized in vivo-like conditions for systems biology. *The FEBS journal*, 277(3):749–60.

Venable, J. D., Meng-Qiu, D., Wohlschlegel, J., Dillin, A., and Yates, J. R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods*, 1(1):39.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Voet, D. and Voet, J. G. (2011). *Biochemistry*. John Wiley & Sons.

Walther, D. M. and Mann, M. (2011). Accurate Quantification of More Than 4000 Mouse Tissue Proteins Reveals Minimal Proteome Changes During Aging. *Molecular & Cellular Proteomics*, 10(2):M110.004523.

Wang, X., Pandey, A. K., Mulligan, M. K., Williams, E. G., Mozhui, K., Li, Z., Jovaisaite, V., Quarles, L. D., Xiao, Z., Huang, J., et al. (2016). Joint mouse-human phenome-wide association to test gene function and disease risk. *Nature communications*, 7.

Warden, C. H. and Fisler, J. S. (2008). Comparisons of diets used in animal models of high-fat feeding. *Cell metabolism*, 7(4):277.

Wiese, T. J., Dunlap, J. A., and Yorek, M. A. (1997). Effect of l-fucose and d-glucose concentration on l-fucoprotein metabolism in human Hep G2 cells and changes in fucosyltransferase and α-l-fucosidase activity in liver of diabetic rats. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1335(1-2):61–72.

Williams, E. and Auwerx, J. (2015). The Convergence of Systems and Reductionist Approaches in Complex Trait Analysis. *Cell*, 162(1):23–32.

Williams, E. G. (2014). A Systems Approach to Identify Genetic and Environmental Regulators of Metabolism. *doi.org*, pages –.

Williams, E. G., Wu, Y., Jha, P., Dubuis, S., Blattmann, P., Argmann, C. A., Houten, S. M., Amariuta, T., Wolski, W., Zamboni, N., Aebersold, R., and Auwerx, J. (2016). Systems proteomics of liver mitochondria function. *Science*, 352(6291):aad0189–aad0189.

Williams, R. W. and Williams, E. G. (2017). Resources for Systems Genetics. pages 3–29. Humana Press, New York, NY.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., and Scalbert, A. (2013). HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al. (2012). Hmdb 3.0the human metabolome database in 2013. *Nucleic acids research*, 41(D1):D801–D807.

Wu, G., Fang, Y.-Z., Yang, S., Lupton, J. R., and Turner, N. D. (2004). Glutathione metabolism and its implications for health. *The Journal of nutrition*, 134(3):489–92.

Wu, Y., Williams, E. G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S. M., Argmann, C. A., Faridi, P., Wolski, W., Kutalik, Z., et al. (2014). Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell*, 158(6):1415–1430.

WURTMAN, R. J. and ANTON-TAY, F. (1969). The Mammalian Pineal as a Neuroendocrine Transducer. In *Proceedings of the 1968 Laurentian Hormone Conference*, pages 493–522.

Xia, J. and Wishart, D. S. (2010a). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(Web Server):W71–W77.

Xia, J. and Wishart, D. S. (2010b). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(Web Server):W71–W77.

Xia, J., Wishart, D. S., Xia, J., and Wishart, D. S. (2016). Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. In *Current Protocols in Bioinformatics*, pages 1–14. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Yuan, R., Peters, L. L., and Paigen, B. (2011). Mice as a mammalian model for research on the genetics of aging. *ILAR journal*, 52(1):4–15.

Zamboni, N., Fendt, S.-M., Rühl, M., and Sauer, U. (2009). 13c-based metabolic flux analysis. *Nature protocols*, 4(6):878.

Zhang, J. D., Hatje, K., Sturm, G., Broger, C., Ebeling, M., Burtin, M., Terzi, F., Pomposiello, S. I., and Badi, L. (2017). Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genomics*, 18(1):277.

Zlatkis, A. and Liebich, H. M. (1971). Profile of volatile metabolites in human urine. *Clinical chemistry*, 17(7):592–4.

Zucker, I. and Beery, A. K. (2010). Males still dominate animal studies. *Nature*, 465(7299):690–690.

Acknowledgments

I would like to thank Reudi(who is an amazing boss, and incredibly humble for how much he has accomplished) and all of the members of the Aebersold lab for being so welcoming, warm and having quick answers to tricky questions whenever I needed them. I guess that is the advantage of having a few dozen world class scientists sitting within a minutes walk from you. The lengthy, enlightening, discussions we had talking through the challenges of this thesis were a delight in their own right.

In the same vein I would also like to thank the metbolomics maestro Nicola Zamboni, my other project supervisor who once sat with me for 3, one hour sessions in a single week, enchanting me with harrowing tales of exploring the frontier of shotgun metbolomics!. He was an absolutely delight to work with.

I would never have applied for this thesis if it was not for the encouragement from professors in the BSSE like Dr. Panke and Dr. Borgwardt who support students going out of their comfort zone to tackle new challenges.

I would also like to personally thank Anuli, Yansheng, Ari, Theodora, Peng and Rodlfo for always having great advise, healthy amount of sarcasm and being such a delight to share my days and messy work benches with. Additionally, a thanks to my two French buddies, Etian my gregarious Canadian friend who somehow can distill a whole field of research into a few succinct ideas and Audrey who is as charming as she is hard working and always willing to tell me the hard truth.

A heartfelt acknowledgment also goes out to Aroosa Ijaz, who stayed up many nights looking through my writing and helping with the monotonous metabolite extractions that were such a success in this thesis. She was a momentous support throughout this thesis.

My family, my support system, are also well deserving of an acknowledgment. They have been there with me the whole time, always a skype call away. My father and mother for believing in my ludicrous decision to come to Switzerland and explore a new continent even-though I had a PhD position in locked down in Canada. My sister Zuha who is my rolemodel for someone with true perseverance and someone who really made me believe I could do anything I set my mind to and my brother who I couldn't be prouder of for following in my foot steps. My friends Kevin, Amy, Shivalai, Nabeel, Peter, Carly, Graeme, Karan, Lape, Chanel, Zack, Vincent, Dr. Elizabeth Mayer and Dr. Scott Hopkins, almost like family, were another shining reason I give it my best everyday. Alongside my nuclear family and Canadian friends, I have to thank my new family in Basel, my friends whose support has been fantastic and with whom I have shared countless unforgettable memories. When my family in Canada and Pakistan was too far to hold out a helping hand they are always there for me. I thank these people with my .

Finally, I would like to thank my direct supervisor Evan Williams. Despite all the disadvantages of being a person from Memphis, he is one of the most humble, intelligent and clever people I have ever met. Aside from learning about science, Evan has taught me multitudes about how to navigate a successful career in academia. How not to waste time going after questions that would be too expensive and take too long to answer when there are many promising lines of inquiry right ahead. He's taught me that knowing how to code, or learning fancy statistic doesn't mean anything unless you can gleam insights from the complex, stochastic noisy systems we aim to study. Lastly I'd like to thank him for offering me a position and encouraging me in a time I thought I would not continue to at ETH. I had an incredible time under his tutelage and I have learned more in these 8 months than I did sleeping through 4 years of lectures in college.

This has truly been an ephemeral and yet transformative experience.

Chapter 6

Appendix

6.A Metabolite Extraction Kinetics

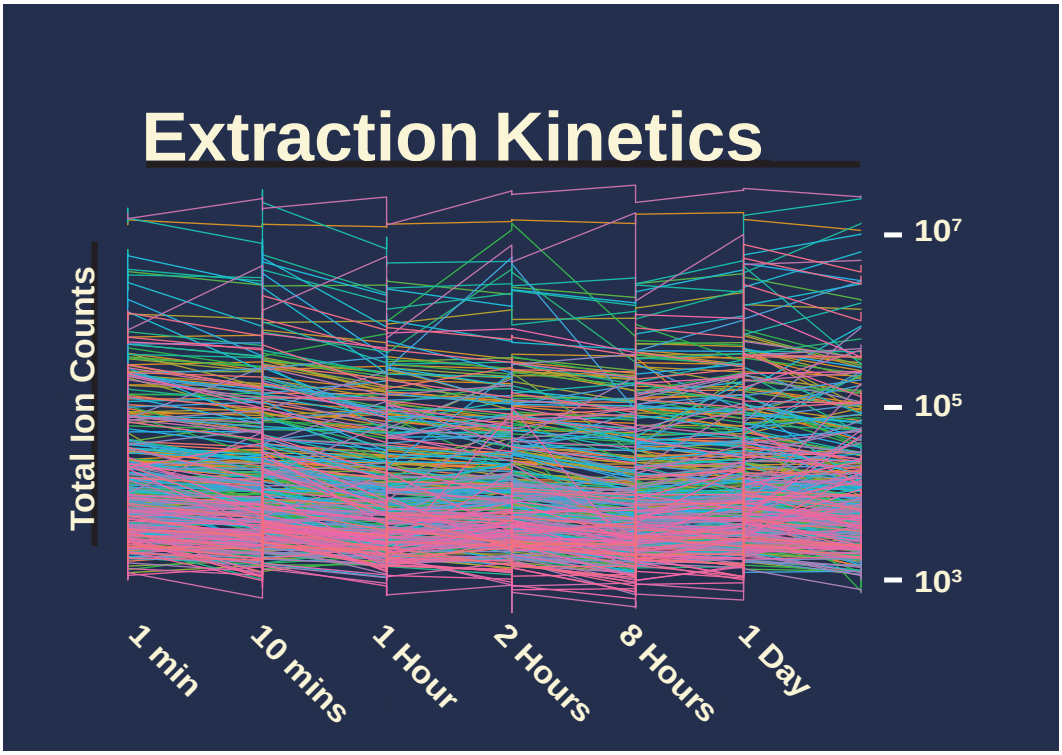
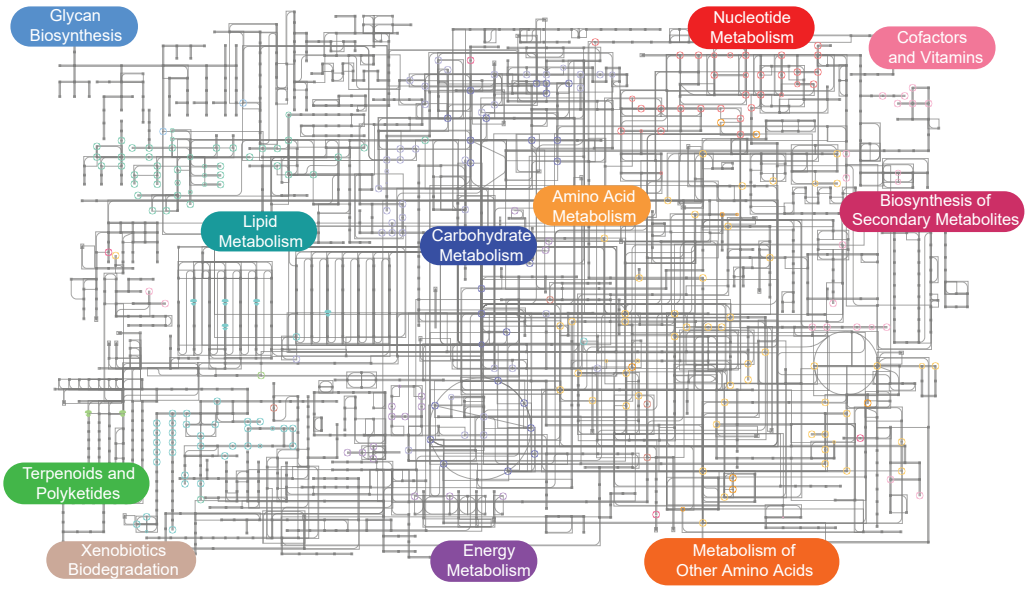


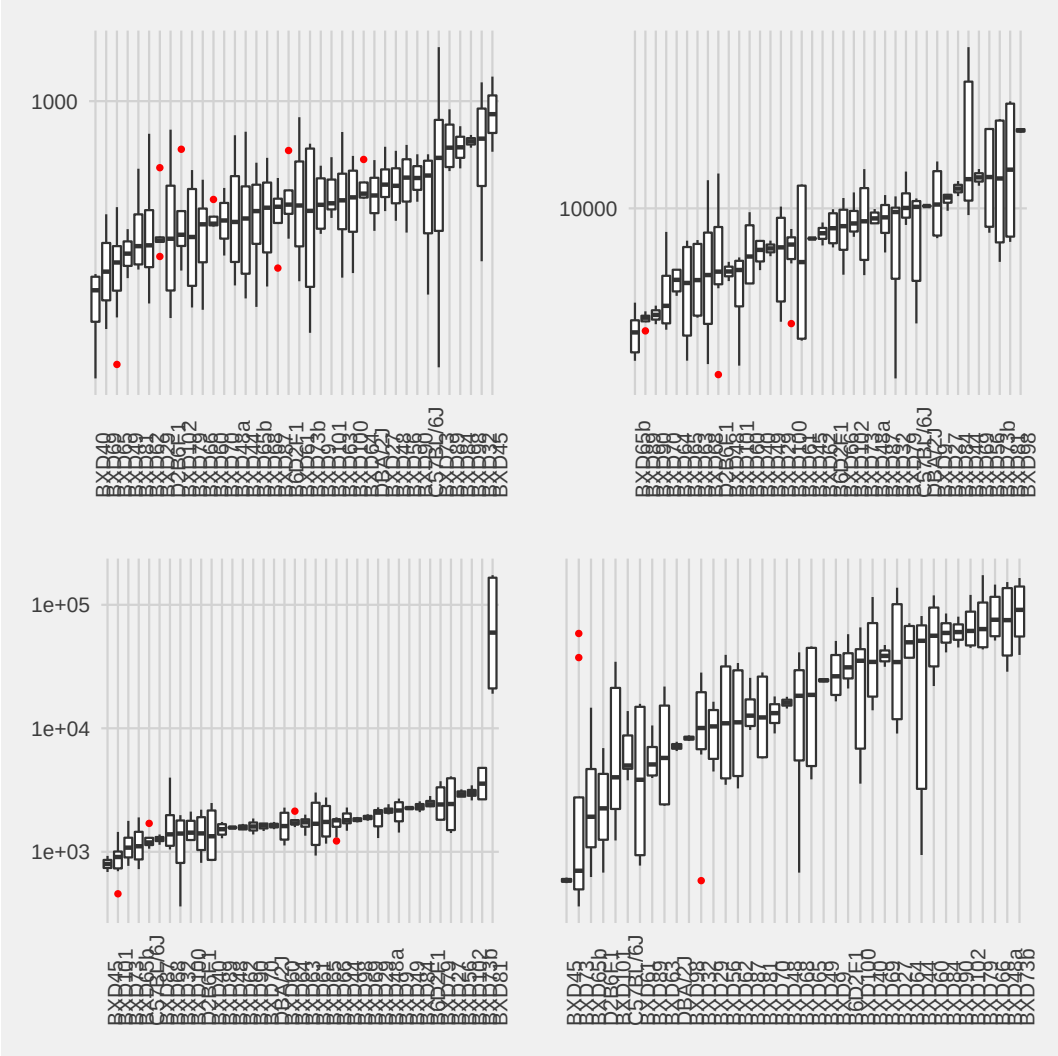
Figure 6.1: The extraction results of all detected metabolites in Pilot study 2 can be seen. On average there is little effect on the metabolites with respect to time. Very short extractions (1 and 10 minutes) produce more variable metabolites intensities than longer extraction. An exception is the 3 day extractions in which a significant amounts of degredation products begin to accumulate.

6.B Metabolite Coverage

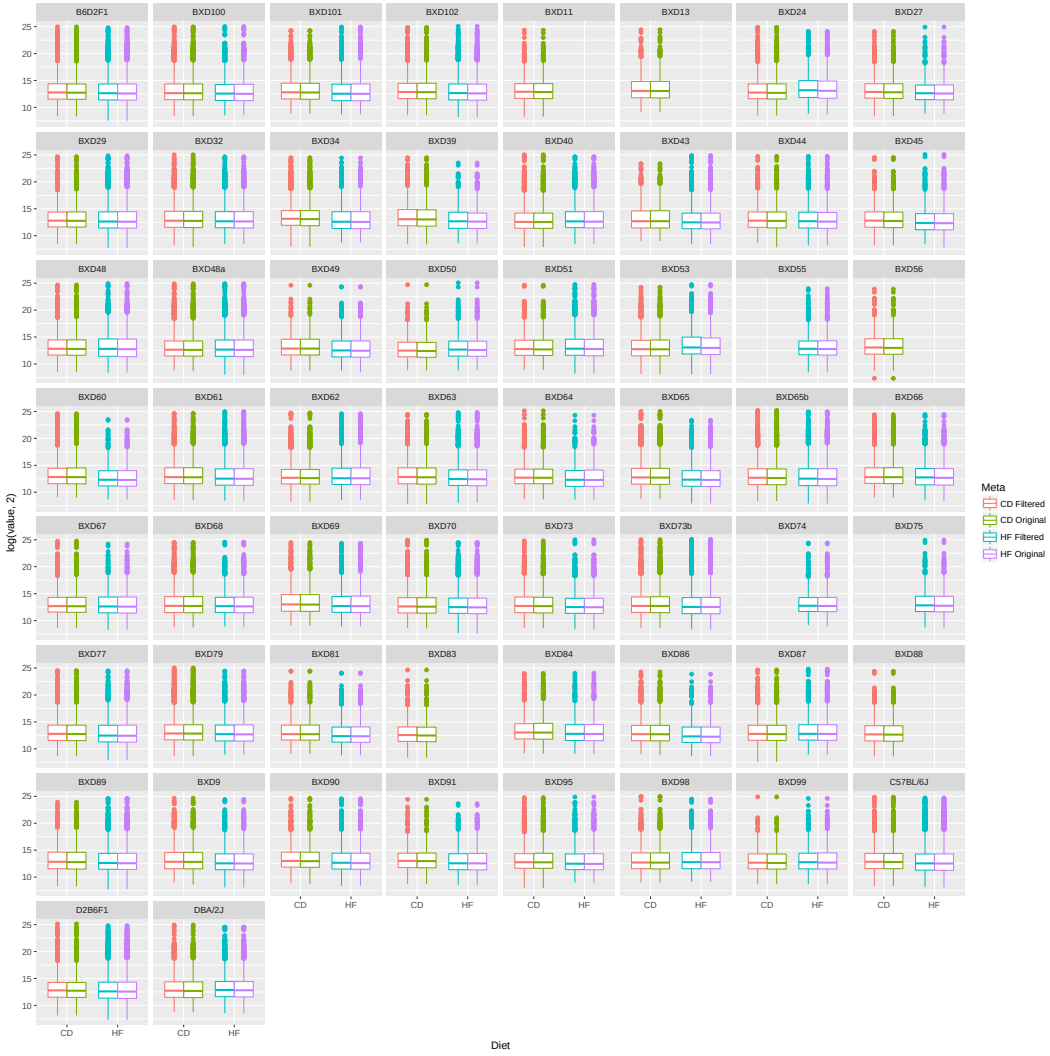


captionMetabolome Coverage in Full Run 1

6.C Metabolites Intensities



6.D Effects Normalization on Metabolite Data



6.E QTL Results

QTL Results CD Mice

Marker	Phenotype	Metabolite Name	chr	pos	lod	pval	Diet	Age
rs49388131	HMDB14900	Irinotecan	7	74.34081	9.799634	0	CD	All
rs3693161	HMDB00673	Linoleic acid	1	160.2851	8.938082	0	CD	All
rs31523971	HMDB00201	L-Acetylcarnitine	3	38.48307	8.937312	0	CD	Young
rs31523971	HMDB00150	Gluconolactone	3	38.48307	8.937312	0	CD	Young
rs30698864	HMDB00210	Pantothenic acid	3	36.54333	8.596278	0	CD	Young
rs27455097	HMDB07860	PA(16:0/18:2(9Z,12Z))	2	114.3171	8.586125	0	CD	Young
rs3151914	HMDB06802	O-Phospho-4-hydroxy-L-threonine	3	36.59936	8.586124	0.01	CD	Young
rs3151914	HMDB00638	Dodecanoic acid	3	36.59936	8.586124	0.01	CD	Young
rs31640914	HMDB00895	Acetylcholine	3	23.43916	8.571553	0	CD	All
rs13476709	HMDB00036	Taurocholic acid	2	114.1972	8.474607	0	CD	Young
rs3151914	HMDB00012	Deoxyuridine	3	36.59936	8.395378	0	CD	Young
rs27455097	HMDB00214	Ornithine	2	114.3171	8.300737	0	CD	Young
rs27455097	HMDB00996	3-Sul noalanine	2	114.3171	8.300733	0	CD	Young
rs33299937	HMDB08825	PE(14:0/16:1(9Z))	2	5.653007	8.26945	0	CD	All
rs3714242	HMDB08856	PE(14:1(9Z)/15:0)	12	47.31295	8.237837	0	CD	All
rs30698864	HMDB00929	L-Tryptophan	3	36.54333	8.213883	0	CD	Young

QTL Results HF Mice

Marker	Phenotype	Metabolite Name	chr	pos	lod	pval	Diet	Age
rs30199004	HMDB06791	Melibiitol	10	86.46702	9.17565	0	HF	Old
rs29624362	HMDB01358	Retinal	2	180.1137	9.109887	0	HF	All
rs32071323	HMDB11750	Dihydroxyacetone Phosphate Acyl Ester	5	111.3495	8.716479	0	HF	All
rs49386605	HMDB11676	D-Xylono-1,5-lactone	6	92.88526	8.424436	0	HF	All
rs47865777	HMDB01487	NADH	7	3.078244	8.34451	0.01	HF	All
rs36251697	HMDB04662	S-(Hydroxymethyl)glutathione	1	3.812265	8.290696	0	HF	All
rs29982107	HMDB11676	D-Xylono-1,5-lactone	9	94.90475	8.197366	0	HF	Young
rs50120872	HMDB10383	LysoPC(16:1(9Z))	12	34.60364	8.14489	0	HF	Old
rs49386605	HMDB10343	1-(alpha-Methyl-4-(2-methylpropyl)benzeneacetate)-beta-D-Glucopyranuronic acid	6	92.88526	8.119253	0	HF	All
rs50180998	HMDB00026	Ureidopropionic acid	12	30.1821	8.101939	0	HF	Old
rs32098314	HMDB00821	Phenylacetylglycine	15	87.98895	8.05834	0	HF	Young
rs13480777	HMDB00673	Linoleic acid	10	116.134	8.050887	0.01	HF	Young
rs258367496	HMDB04662	S-(Hydroxymethyl)glutathione	1	3.659804	7.910335	0	HF	Young
rs4198085	HMDB59612	7-Methylguanosine 5'-phosphate	16	67.77714	7.866325	0	HF	Young
rs32098314	HMDB00904	Citrulline	15	87.98895	7.856642	0	HF	Young

Figure 6.2: The highest LOD Score QTL for HF and CD segregated mice

6.F Appendix: Summary Statistics of Metabolite, Protein and Transcript Data

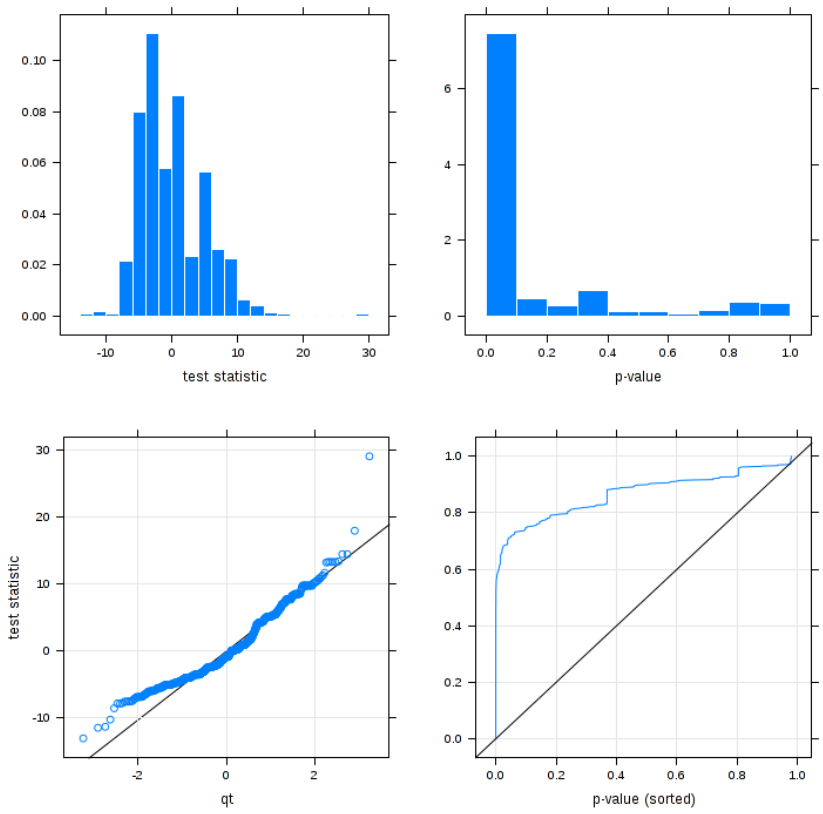


Figure 6.3: Summary Statistic for the metabolite data

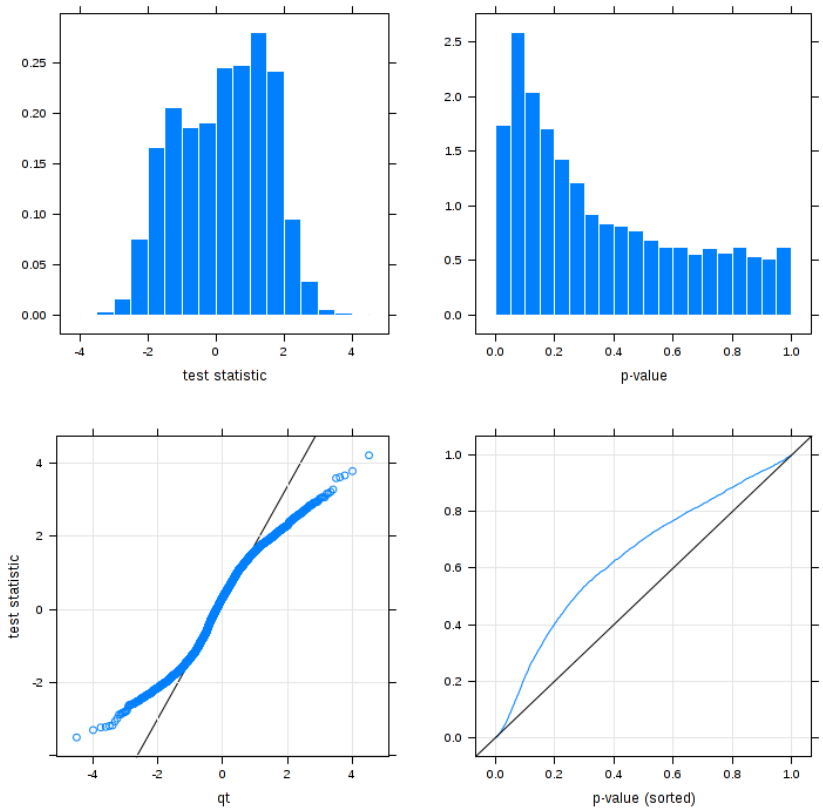


Figure 6.4: summary statics for the proteomics data

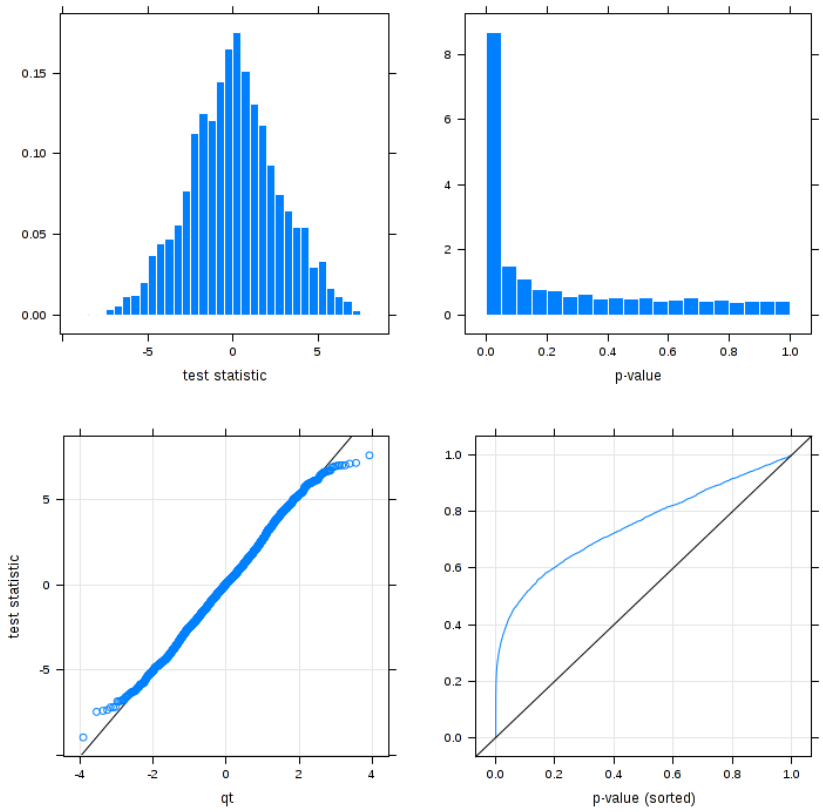


Figure 6.5: summary statistics for the transcriptomics data

6.G Appendix: MSEA Metabolite Pathway Enrichment Tables

Full Run 1- Age Regression

Pathway Name	Match Status	p	-log(p)	Holm p	FDR	Impact
Porphyrin and chlorophyll metabolism	8/27	2.1807E-7	15.338	1.5701E-5	1.5701E-5	0.29701
Taurine and hypotaurine metabolism	5/8	1.745E-6	13.259	1.2389E-4	6.2819E-5	0.99999
Drug metabolism - other enzymes	6/30	3.0013E-5	10.414	0.0021009	6.347E-4	0.14815
Citrate cycle (TCA cycle)	11/20	3.5261E-5	10.253	0.002433	6.347E-4	0.52411
Butanoate metabolism	12/22	4.5638E-5	9.9948	0.0031034	6.5719E-4	0.17392
Glyoxylate and dicarboxylate metabolism	10/18	7.582E-5	9.4872	0.0050799	9.0984E-4	0.58064
Pyrimidine metabolism	21/41	1.7284E-4	8.6631	0.011407	0.0015111	0.64309
Pyruvate metabolism	9/23	1.9779E-4	8.5283	0.012857	0.0015111	0.45792
Sphingolipid metabolism	6/21	2.0143E-4	8.5101	0.012891	0.0015111	0.43108
D-Glutamine and D-glutamate metabolism	5/5	2.0988E-4	8.469	0.013222	0.0015111	1.0
Purine metabolism	37/68	2.7107E-4	8.2131	0.016806	0.0017743	0.76648
Alanine, aspartate and glutamate metabolism	14/24	4.0487E-4	7.8119	0.024697	0.0022712	0.84598
Primary bile acid biosynthesis	28/46	4.1008E-4	7.7991	0.024697	0.0022712	0.33532
Glutathione metabolism	6/26	6.1633E-4	7.3917	0.036363	0.0031697	0.47328
Nitrogen metabolism	5/9	8.1624E-4	7.1108	0.047342	0.003918	0.0
beta-Alanine metabolism	8/17	0.0011935	6.7309	0.068027	0.0053706	0.79629
Steroid hormone biosynthesis	32/72	0.0014634	6.527	0.081948	0.0061977	0.30607
Histidine metabolism	9/15	0.0017148	6.3684	0.094315	0.0062754	0.46775
Synthesis and degradation of ketone bodies	2/5	0.0017504	6.3479	0.09452	0.0062754	0.6
Tryptophan metabolism	9/40	0.0018041	6.3177	0.095615	0.0062754	0.59635
Arginine and proline metabolism	20/44	0.0018303	6.3033	0.095615	0.0062754	0.50347
Phenylalanine, tyrosine and tryptophan biosynthesis	3/4	0.0019261	6.2523	0.098232	0.0063036	1.0
Glycine, serine and threonine metabolism	16/31	0.0024309	6.0195	0.12154	0.0076098	0.74853
Glycerophospholipid metabolism	11/30	0.0026672	5.9267	0.13069	0.0080017	0.67254
Phenylalanine metabolism	6/11	0.0027818	5.8847	0.13352	0.0080115	0.53704
Glycolysis or Gluconeogenesis	14/26	0.0034894	5.658	0.164	0.0096631	0.63724
Metabolism of xenobiotics by cytochrome P450	3/39	0.0040465	5.5099	0.18614	0.010791	0.0
Valine, leucine and isoleucine degradation	11/38	0.0047414	5.3514	0.21336	0.011956	0.11705
Tyrosine metabolism	14/44	0.0048975	5.319	0.21549	0.011956	0.49607
Ubiquinone and other terpenoid-quinone biosynthesis	2/3	0.0049816	5.302	0.21549	0.011956	1.0
Propanoate metabolism	8/20	0.007094	4.9485	0.29795	0.016476	0.0
Pantothenate and CoA biosynthesis	10/15	0.0079674	4.8324	0.32667	0.017927	0.61225
Aminoacyl-tRNA biosynthesis	18/69	0.0099983	4.6053	0.39993	0.021814	0.12903
Pentose and glucuronate interconversions	13/16	0.010403	4.5657	0.40572	0.02203	0.86667
Amino sugar and nucleotide sugar metabolism	25/37	0.010742	4.5336	0.40821	0.022099	0.72244
Terpenoid backbone biosynthesis	1/15	0.012701	4.3661	0.46995	0.025403	0.18817
Nicotinate and nicotinamide metabolism	3/13	0.013673	4.2923	0.49223	0.026607	0.44643
Cysteine and methionine metabolism	10/27	0.01529	4.1806	0.53515	0.028971	0.3854
Linoleic acid metabolism	4/6	0.018287	4.0016	0.62176	0.033761	1.0
Valine, leucine and isoleucine biosynthesis	7/11	0.023354	3.757	0.77069	0.042038	0.99999

Full Run 1 – Diet

Pathway Name	Match Status	p	-log(p)	Holm p	FDR	Impact
Pyrimidine metabolism	21/41	6.4775E-134	306.68	4.6638E-132	4.6638E-132	0.64309
Arachidonic acid metabolism	33/36	5.7657E-65	147.92	4.0936E-63	2.0756E-63	0.97927
Glycerophospholipid metabolism	11/30	2.5948E-44	100.36	1.8164E-42	6.2275E-43	0.67254
Alanine, aspartate and glutamate metabolism	14/24	1.1954E-36	82.715	8.2482E-35	2.1516E-35	0.84598
Taurine and hypotaurine metabolism	5/8	4.0960E-29	65.365	2.7853E-27	5.8984E-28	0.99999
Citrate cycle (TCA cycle)	11/20	6.4445E-29	64.912	4.3178E-27	7.7334E-28	0.52411
Steroid hormone biosynthesis	32/72	1.6476E-28	63.973	1.0875E-26	1.6947E-27	0.30607
Drug metabolism - other enzymes	6/30	5.0994E-28	62.843	3.3147E-26	4.5896E-27	0.14815
Linoleic acid metabolism	4/6	1.1105E-26	59.762	7.1069E-25	8.8836E-26	1.0
Retinol metabolism	13/16	2.9041E-24	54.196	1.8295E-22	2.091E-23	1.0
Lysine degradation	6/23	6.0631E-22	48.855	3.7592E-20	3.9686E-21	0.10295
Butanoate metabolism	12/22	6.6225E-21	46.464	4.0397E-19	3.9735E-20	0.17392
Glyoxylate and dicarboxylate metabolism	10/18	1.9247E-16	36.187	1.1548E-14	1.066E-15	0.58064
Synthesis and degradation of ketone bodies	2/5	2.4009E-16	35.965	1.4166E-14	1.2348E-15	0.6
N-Glycan biosynthesis	3/36	4.8088E-16	35.271	2.7891E-14	2.3082E-15	0.01801
Arginine and proline metabolism	20/44	2.2677E-14	31.417	1.2926E-12	9.6582E-14	0.50347
Tyrosine metabolism	14/44	2.2804E-14	31.412	1.2926E-12	9.6582E-14	0.49607
Ascorbate and aldarate metabolism	7/9	5.9687E-14	30.45	3.2828E-12	2.3875E-13	0.8
Propanoate metabolism	8/20	6.6643E-14	30.339	3.5987E-12	2.5254E-13	0.0
Histidine metabolism	9/15	1.928E-13	29.277	1.0218E-11	6.9408E-13	0.46775
D-Glutamine and D-glutamate metabolism	5/5	2.09E-13	29.196	1.0868E-11	7.1656E-13	1.0
Pyruvate metabolism	9/23	7.488E-12	25.618	3.8189E-10	2.4506E-11	0.45792
alpha-Linolenic acid metabolism	3/9	3.9471E-11	23.955	1.9736E-9	1.2356E-10	1.0
beta-Alanine metabolism	8/17	1.6837E-10	22.505	8.2503E-9	5.0512E-10	0.79629
Steroid biosynthesis	7/35	4.1894E-10	21.593	2.0109E-8	1.2065E-9	0.13485
Tryptophan metabolism	9/40	8.2033E-10	20.921	3.8556E-8	2.2717E-9	0.59635
Pentose and glucuronate interconversions	13/16	3.4626E-9	19.481	1.5928E-7	9.2336E-9	0.86667
Lysine biosynthesis	4/4	7.2068E-9	18.748	3.2431E-7	1.8057E-8	0.0
Cysteine and methionine metabolism	10/27	7.2729E-9	18.739	3.2431E-7	1.8057E-8	0.3854
Biosynthesis of unsaturated fatty acids	8/42	9.0115E-9	18.525	3.8749E-7	2.1628E-8	0.0
Glycine, serine and threonine metabolism	16/31	6.4192E-8	16.561	2.6961E-6	1.4909E-7	0.74853
Riboflavin metabolism	2/11	1.3372E-7	15.828	5.4826E-6	3.0087E-7	0.16667
Sphingolipid metabolism	6/21	1.6323E-7	15.628	6.529E-6	3.5613E-7	0.43108
Amino sugar and nucleotide sugar metabolism	25/37	1.9345E-7	15.458	7.5447E-6	4.0967E-7	0.72244
Nitrogen metabolism	5/9	3.1979E-7	14.956	1.2152E-5	6.5785E-7	0.0
Valine, leucine and isoleucine degradation	11/38	1.2273E-6	13.611	4.5408E-5	2.4545E-6	0.11705
Fatty acid metabolism	1/39	8.8567E-6	11.634	3.1884E-4	1.7235E-5	0.0
Porphyrin and chlorophyll metabolism	8/27	1.0812E-5	11.435	3.7842E-4	2.0486E-5	0.29701
Fructose and mannose metabolism	12/21	1.1137E-5	11.405	3.7866E-4	2.0561E-5	0.7061
Primary bile acid biosynthesis	28/46	1.4844E-5	11.118	4.8984E-4	2.6719E-5	0.33532

Full Run 2 – Age

Pathway Name	Match Status	p	-log(p)	Holm p	FDR	Impact
GPI-anchor biosynthesis	2/14	1.8465E-7	15.505	1.3849E-5	1.3849E-5	0.0439
Inositol phosphate metabolism	18/28	6.859E-7	14.193	5.0756E-5	1.863E-5	0.71953
Terpenoid backbone biosynthesis	8/15	7.4518E-7	14.11	5.4398E-5	1.863E-5	0.72311
N-Glycan biosynthesis	3/36	1.2773E-6	13.571	9.1968E-5	2.395E-5	0.0924
Drug metabolism - other enzymes	17/30	2.8666E-6	12.762	2.0353E-4	4.2999E-5	0.48678
Pentose phosphate pathway	15/19	4.4171E-6	12.33	3.092E-4	5.5214E-5	0.59835
Glycolysis or Gluconeogenesis	17/26	5.9932E-6	12.025	4.1353E-4	6.4213E-5	0.74839
Fructose and mannose metabolism	14/21	1.1056E-5	11.413	7.518E-4	1.0365E-4	0.74861
Valine, leucine and isoleucine degradation	18/38	1.562E-5	11.067	0.0010465	1.3017E-4	0.42917
Cysteine and methionine metabolism	19/27	3.5382E-5	10.249	0.0023352	2.4098E-4	0.63993
Amino sugar and nucleotide sugar metabolism	29/37	3.9914E-5	10.129	0.0025944	2.4098E-4	0.73794
Pyruvate metabolism	12/23	4.0889E-5	10.105	0.0026169	2.4098E-4	0.6725
Glycerolipid metabolism	8/18	4.177E-5	10.083	0.0026315	2.4098E-4	0.53753
Synthesis and degradation of ketone bodies	3/5	5.7156E-5	9.7697	0.0035437	3.0619E-4	0.6
Vitamin B6 metabolism	8/9	9.5258E-5	9.2589	0.0058107	4.7629E-4	1.0
Tryptophan metabolism	30/40	1.2187E-4	9.0126	0.0073122	5.7126E-4	0.93713
Lysine degradation	11/23	1.3924E-4	8.8793	0.0082151	5.7969E-4	0.29413
Sphingolipid metabolism	14/21	1.4637E-4	8.8293	0.0084897	5.7969E-4	0.82708
Galactose metabolism	23/26	1.4685E-4	8.8261	0.0084897	5.7969E-4	0.94322
Histidine metabolism	12/15	2.3382E-4	8.3609	0.013094	8.1756E-4	0.61291
beta-Alanine metabolism	9/17	2.3829E-4	8.342	0.013106	8.1756E-4	0.79629
Glyoxylate and dicarboxylate metabolism	15/18	2.3982E-4	8.3356	0.013106	8.1756E-4	0.67742
Starch and sucrose metabolism	16/19	3.1076E-4	8.0765	0.016471	9.5284E-4	0.78464
Biotin metabolism	3/5	3.3078E-4	8.0141	0.0172	9.5284E-4	0.7
Valine, leucine and isoleucine biosynthesis	6/11	3.332E-4	8.0068	0.0172	9.5284E-4	0.99999
Purine metabolism	47/68	3.528E-4	7.9496	0.01764	9.5284E-4	0.81356
Drug metabolism - cytochrome P450	29/56	3.5481E-4	7.9439	0.01764	9.5284E-4	0.52144
Glycine, serine and threonine metabolism	20/31	3.5573E-4	7.9413	0.01764	9.5284E-4	0.85531
Pyrimidine metabolism	33/41	4.4007E-4	7.7286	0.020683	0.0011381	0.93805
Tyrosine metabolism	30/44	5.3169E-4	7.5395	0.024458	0.0012885	0.81085
Glutathione metabolism	13/26	5.3257E-4	7.5378	0.024458	0.0012885	0.68128
Propanoate metabolism	9/20	6.7815E-4	7.2961	0.029839	0.0015382	0.00862
Nicotinate and nicotinamide metabolism	11/13	6.8625E-4	7.2843	0.029839	0.0015382	0.79168
Ubiquinone and other terpenoid-quinone biosynthesis	2/3	6.9733E-4	7.2683	0.029839	0.0015382	1.0
Selenoamino acid metabolism	9/15	7.2642E-4	7.2274	0.029839	0.0015472	0.74312
Limonene and pinene degradation	2/8	7.4265E-4	7.2053	0.029839	0.0015472	0.0
Alanine, aspartate and glutamate metabolism	19/24	8.4061E-4	7.0814	0.032784	0.0017039	0.89028
Nitrogen metabolism	6/9	9.2074E-4	6.9903	0.034988	0.0018173	0.0
Fatty acid elongation in mitochondria	6/27	9.866E-4	6.9212	0.036504	0.0018973	0.33809
Cyanoamino acid metabolism	5/6	0.0011071	6.806	0.039857	0.0020759	0.0

Full Run 2 – Diet

Pathway Name	Match	p	-log(p)	Holm p	FDR	Impact
Pyrimidine metabolism	29/41	1.2428E-65	149.45	9.3219E-64	9.321E-64	0.90609
Biotin metabolism	2/5	3.1362E-59	134.71	2.3208E-57	1.176E-57	0.4
Drug metabolism - other enzymes	11/30	7.9414E-47	106.15	5.7972E-45	1.985E-45	0.3598
Glycerophospholipid metabolism	16/30	4.931E-42	95.113	3.5503E-40	9.245E-41	0.72038
Cyanoamino acid metabolism	5/6	1.0407E-20	46.012	7.3893E-19	1.561E-19	0.0
Citrate cycle (TCA cycle)	13/20	1.4148E-20	45.705	9.9039E-19	1.768E-19	0.62406
Alanine, aspartate and glutamate metabolism	17/24	7.5775E-20	44.027	5.2285E-18	8.118E-19	0.89028
Steroid biosynthesis	3/35	2.3150E-19	42.91	1.5742E-17	2.170E-18	0.04149
Butanoate metabolism	12/22	1.4436E-17	38.777	9.672E-16	1.203E-16	0.15943
Porphyrin and chlorophyll metabolism	7/27	6.8479E-14	30.312	4.5196E-12	5.135E-13	0.25681
Propanoate metabolism	9/20	8.2667E-12	25.519	5.3734E-10	5.636E-11	0.00862
Cysteine and methionine metabolism	17/27	3.7825E-11	23.998	2.4208E-9	2.364E-10	0.63993
Pyruvate metabolism	11/23	8.0716E-11	23.24	5.0851E-9	4.656E-10	0.6725
beta-Alanine metabolism	9/17	1.0597E-9	20.665	6.5701E-8	5.677E-9	0.79629
Taurine and hypotaurine metabolism	6/8	2.3689E-9	19.861	1.445E-7	1.1844E-8	0.99999
D-Glutamine and D-glutamate metabolism	5/5	1.6267E-8	17.934	9.7602E-7	7.6252E-8	1.0
Glyoxylate and dicarboxylate metabolism	14/18	2.2491E-8	17.61	1.327E-6	9.9226E-8	0.67742
Arginine and proline metabolism	30/44	3.6782E-8	17.118	2.1333E-6	1.4921E-7	0.66866
Pentose and glucuronate interconversions	14/16	3.7801E-8	17.091	2.1547E-6	1.4921E-7	0.73333
Nitrogen metabolism	5/9	4.2567E-8	16.972	2.3837E-6	1.5963E-7	0.0
Sphingolipid metabolism	6/21	6.2538E-8	16.587	3.4396E-6	2.2335E-7	0.49123
Drug metabolism - cytochrome P450	21/56	1.0341E-7	16.085	5.5843E-6	3.5255E-7	0.42144
Selenoamino acid metabolism	7/15	1.2038E-7	15.933	6.3799E-6	3.9253E-7	0.55046
Methane metabolism	4/9	1.9166E-7	15.468	9.9661E-6	5.9892E-7	0.4
Biosynthesis of unsaturated fatty acids	10/42	2.4266E-7	15.232	1.2375E-5	7.2797E-7	0.0
Histidine metabolism	11/15	4.1432E-7	14.697	2.0716E-5	1.1952E-6	0.61291
Fatty acid biosynthesis	6/43	4.7217E-7	14.566	2.3137E-5	1.3116E-6	0.02598
Purine metabolism	44/68	5.4531E-7	14.422	2.6175E-5	1.4607E-6	0.786
Fructose and mannose metabolism	14/21	6.1089E-7	14.308	2.8712E-5	1.5799E-6	0.74861
Lysine degradation	9/23	7.3856E-7	14.119	3.3974E-5	1.8464E-6	0.10295
Retinol metabolism	4/16	1.8995E-6	13.174	8.5477E-5	4.5955E-6	0.52096
Pentose phosphate pathway	13/19	2.2625E-6	12.999	9.9552E-5	5.3028E-6	0.53153
Glycine, serine and threonine metabolism	19/31	2.5021E-6	12.898	1.0759E-4	5.6867E-6	0.80883
Synthesis and degradation of ketone bodies	3/5	3.0321E-6	12.706	1.2735E-4	6.6884E-6	0.6
Ubiquinone and other terpenoid-quinone biosynthesis	1/3	4.5251E-6	12.306	1.8553E-4	9.6966E-6	0.0
Tryptophan metabolism	22/40	8.0554E-6	11.729	3.2222E-4	1.6782E-5	0.68088
Glutathione metabolism	9/26	8.6943E-6	11.653	3.3908E-4	1.7624E-5	0.67079
Glycolysis or Gluconeogenesis	16/26	1.3438E-5	11.217	5.1064E-4	2.6522E-5	0.6445
Amino sugar and nucleotide sugar metabolism	28/37	1.5671E-5	11.064	5.7983E-4	3.0136E-5	0.72038
Valine, leucine and isoleucine degradation	15/38	1.9582E-5	10.841	7.0497E-4	3.6717E-5	0.27135