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. Reudi 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

List of Figures

Abstract

A large scale study of BXD genetic reference population metabolome, proteome, transcriptome, genome and phenome was undertaken to determine factors involved in metabolic disease and Aging. This investigation include 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.

In other to reduce the complexity and increase reproducibility of the metabolomics protocol a pilot study of on 24 mouse livers was used to remove steps determine whether all of the homogenization and extracting extraction steps were truly necessary. Once the procedure was optimized 632 mice livers where subjected to proteomics, metabolomics and transcriptomics analysis. Although the statistical algorithms exist as easy to deploy packages in R, as effort to write them from scratch was made in order to ensure no defaults settings or erroneous variable assignments present in the resulting analysis leading us to find bugs in a published R package after validation.

Many differentially evident metabolites between the diet and age cohorts were discovered and were added to a list of a biomarker candidate s. 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. Lastly, 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 aforementioned pathways were again appended to a growing list of possible aging and metabolic biomarkers.

QTL analysis was able to find strongly regulated metabolites that had been previously found in the same mouse population. Additionally, three novel QTLs in central glucose metabolism, glutathione metabolism and amino acid metabolism were found. The protein data and transcript data were still being processed at the end of this

Machine Learning Algorithms were used to determine the most important discriminating factors between

Diet cohorts and Aging 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.

From this analysis, a few metabolites which have known knock out mice available and known inhibitors are added as a list of metabolites to follow up with in validation experiments due to occur at the beginning of next year.

Chapter 1

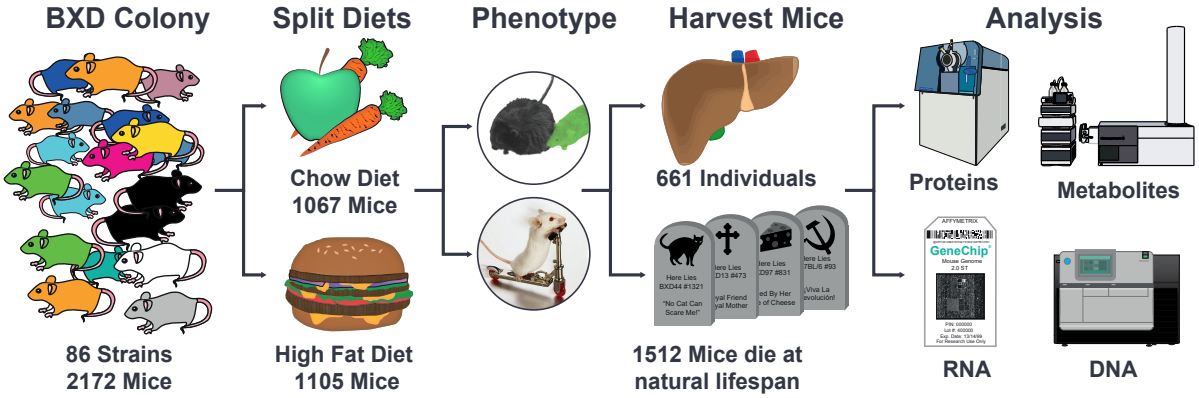
Introduction

The late 19th and early 20th centuries were a golden era of health innovation. Breakthroughs like germ theory, antibiotics, and widespread vaccination, as well as major public-health advances in sanitation and regulation, neutralized many long-leading causes of death. With life expectancy increasing rapidly in the the focus of medical innovation has shifted away from the eradicating of infection disease and more towards managing chronic age-related conditions. Aging or more specifically age-related degeneration is a risk factor for several of the worlds most prevalent diseases, including neurodegenerative disorders, cancer and cardiovascular disease.

Despite large efforts using model organisms the understanding of the pathways and molecules involved in the onset and progression of chronic diseases, how they interact with aging remains unclear to due the financial and logistical challenge of large longitudinal studies in mice. The development of longitudinal assessments of aging phenotypes in multiple model organisms, with attention to differences in sex, strain and diet composition, could accelerate our ability to better screen for interventions that would lead to people living longer and healthier lives.

The goal of targeting and treating the process of aging is to identify interventions that could extend the health span of individuals rather than their life-spans outright. Through a clinical perspective, it seems more

the goal is to extend health-span. The ideal would be to preserve all the faculties of a young person in their prime in an old person and maintain them for as long as possible, ideally well into a person's golden years.



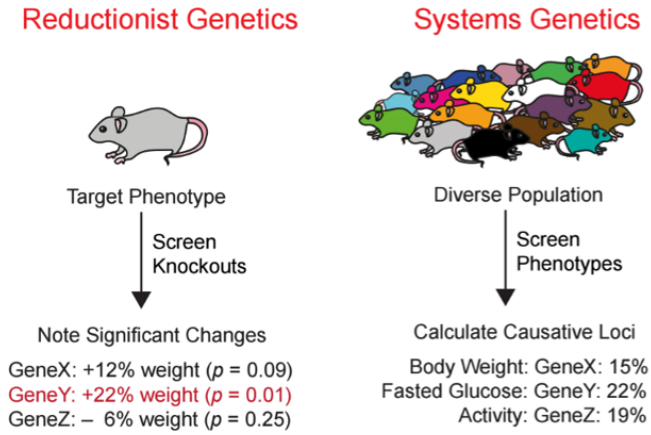
1.1 Project Outline

The majority of the mouse husbandry has already been performed at the time the thesis project began. Over a 180 phenotypes exist for all the mice included standard plasma metabolite, VO2 max .

As 600 samples are not trivial to The Optimization of the

Additionally the Proteomics required much longer ruin-times as there is a hour long chromatographic separation prior to an orthogonal proteomic analysis using the triple-TOF machine.

1.2 Gene-Function Paradigms



The reductionist approach in genetic involves identifying a gene of interest and altering through mean of

random and directed mutagenesis and observing the manifested effect of its absence or hyper expression. 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.

The classical forward genetics involve determining polymorphisms that modulate certain phenotypes and disease risks. This is the generalization of medial inheritance analysis applied to complex traits. Certain genes are alter through direct editing, silencing or mutagenesis with the commensurate phenotype used to induct the function of the gene.

most reductive genetics, knockout or massive transgenic over expression such errors are rare, in REAL bioligcal, because may be fatal especially like complex disease, like diabietes (diabieites) can be mono genic ever though its exptremely not "useless but acutyally answers different quetsion P53 kn jownout effect too large, reverse genetics

1.3 Reverse and Forward genetics

There are two different study designs that can be used to determine the function of a gene. With a reverse genetics design, a specific gene is distrupted usually through a knock out or through the use of viral promoter that cause super over-expression of genes in order to determien the down-stream effects. The process of disruption or alteration can either be targeted specifically as in the case of gene silencing or homologous recombination or can rely on non-targeted random disruptions (e.g., chemical mutagenesis, transposon mediated mutagenesis) followed by screening a library of individuals for lesions at a specific location. for small rapidly prolifferating organisms like *C. Elegans* or *Drosophilla Melanogaster*, random mutategesis is a tenable strategy to generate a large library of variant to screen.

Variants help us understand the 'normal.' Variation can be measured at many scales from macro (body size, morphology) to different levels of micro variation (crude protein profiles to DNA sequence variation). Forward genetics refers to a process where studies are initiated to determine the genetic underpinnings of observable phenotypic variation. In many cases the observable variation has been induced using a DNA damaging agent (mutagen) but also may be naturally occurring. The investigator eventually ends up sequencing the gene or genes thought to be involved (Figure 1).

With the advent of whole genome sequencing many researchers are now in a very different position. They have access to all of the gene sequences within a given organism and would like to know their function. So, instead of going from phenotype to sequence as in forward genetics, reverse genetics works in the opposite direction a gene sequence is known, but its exact function is uncertain. In reverse genetics, a specific gene or gene product is disrupted or modified and then the phenotype is measured (Figure 1). Here we will overview

some of the techniques for reverse genetics with a special emphasis on the TILLING (Targeting Induced Local Lesions IN Genomes) technique which is being utilized on plant pathogens in the genus Phytophthora.

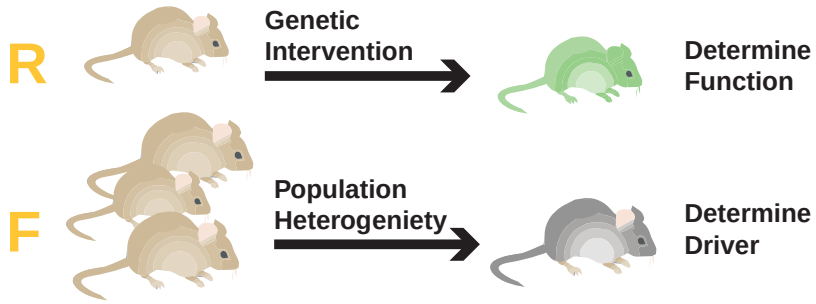


Figure 1.1: Forward and Reverse Genetic Paradigms

1.4 Systems Approach to Complex Trait Analysis

The causal relationship between genetic polymorphism within a species and the phenotypic differences observed between individuals is of fundamental biological interest. The ability to predict genetic risk factors for human disease and important traits like growth rate and yield in plants in the agricultural sector require an understanding of both the specific loci that underlie a phenotype, and the genetic architecture of a trait. This relationship between phenotype and genotype has been of major interest at least since Mendel postulated the existence of internal factors that are passed on to the next generation(CITE BIO TEXTBOOK).

Systems genetics is an approach to understand the flow of biological information that underlies complex traits. It uses a range of experimental and statistical methods to quantitative and integrate intermediate phenotypes, such as transcript, protein or metabolite levels, in populations that vary for traits of interest[citation].

Systems genetics studies have provided the first global view of the molecular architecture of complex traits and are useful for the identification of genes, pathways and networks that underlie common human diseases. Given the urgent need to understand how the thousands of loci that have been identified in genome-wide association studies contribute to disease susceptibility, systems genetics is likely to become an increasingly important approach to understanding both biology and disease[citation].

Recombinant inbred strains A set of inbred strains that is generally produced by crossing two parental inbred strains and then inbreeding random inter-cross progeny; they provide a permanent resource for examining the segregation of traits that differ between the parental strains[citation].

Complex traits are determined by large numbers of genetic and environmental factors as well as their interactions. Identifying the contributing genes and quantifying their effects in the context of one or multiple

environments is of key importance in the development of improved breeding strategies in livestock, the identification of therapeutic targets for animal and human disease, and the understanding of how natural and artificial selection shape the genomes of animals and humans.

INSERT TRANSITION TO HERITABLITY WHY TRAIN HAS TO BE HERTIBALE TO BE DETECTED

1.5 Heritability

A fundamental question in biology is whether a presenting trait is inherited from parents or the result of environmental factors. Heritability (h^2) is an attempt to quantify the relationship between genetic and the environment in the determination of a complex trait phenotype. The concept quantifies the portion of variation within an observed population that can be explained by genetic differences. Formally, Heritability is defined as the portion of phenotypic variation V_p that is due to variation in genetic values V_g. It is a value that ranges from [0.0 , 1.0], with the low heritability meaning there is a low probability phenotypes in the offspring will resemble the parents, and high heritablity values [0.7 , 1.0] stating, the phenotype observed in the offspring are very similar to the parental phenotypes.

In order to determine heritablity one must create 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, 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.

In complex traits, many loci interact with each other through dominance and epistatic relationships in different combinations in the population creating a distribution of observed phenotypes. If either there is no variation in the genome, there can be no variation in their expression due to genes and Heritability is zero. Whether the result is because of genes going to fixation (allele is be lost completely from the population) or because of genetic similarities as in twins is difficult disentangle from whether the gene is FINISH THOUGHT

* GENES DOES ALWAYS MEAN GENOME*

Stochastic manifestation of environmental factor , especially long term disease for chronic disease Genetically identical mice exposed to the same environmental conditions can show significant variation in molecular content and marked differences in phenotypic characteristics due to the stocastic onset of biological trains.

1.6 Why Model Organism Population facilitate Complex Trait Analysis

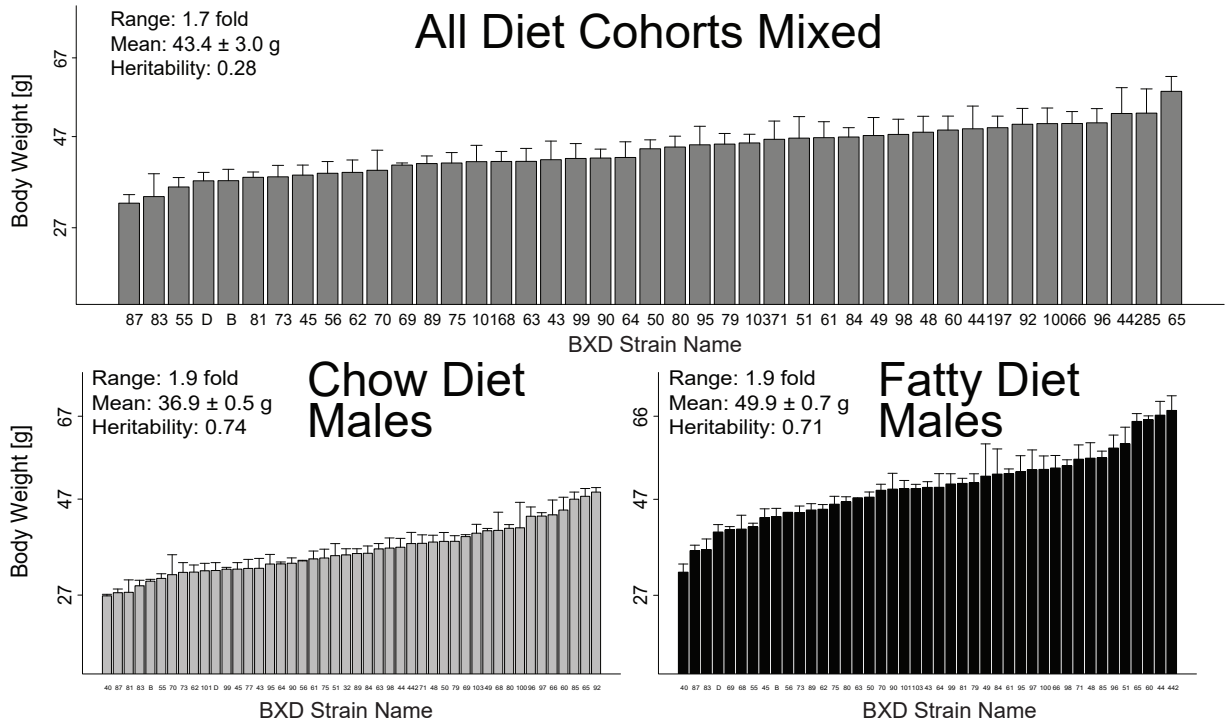


Figure 1.2: Heritability of Body Weight and fixed and mixed Diet population

For consideration, BXD mouse weight has an observed heritability of 0.74. What this means is that while there is an influence exerted by environmental factors (nutritional intake) in the weight of an individual, the major portion of the influence is exerted by the genes. More importantly, it really tells us that the major influence in explaining the *differences* between individuals is accounted for in this fashion. Note that it tells us nothing about what gave rise to the particular height for any particular individual, but rather what explains the differences between individuals within a particular population.

Looking at the in the diagram, we have a uniform environment in the Chow and High fat cases where the nutritional intake and housing are identical so therefore effect of the variation in body weight is more likely due to the effect of genes (heritability = 70+%). In the same mice if we look at the high fat cohort instead of the chow diet cohort we are again, observing the effect of genetics as they interact with the environment(diet) with a large portion of the variability due to genetic factors.

This gives rise to the condition that we can have high heritability within groups,substantial variation between groups, but **no genetic** difference between the groups. In both cases, the environment is essentially held constant, so the variation in weight is solely due to the genes, hence 71% and 74% heritability. However,

if we now combined both diet so that there were also environmental differences between the groups, then the heritability will be the 0.28 value indicated previously because the variation between the groups is not accounted for solely by the genes, but also by the environmental effects that produce the weight.

This is why model organism populations are invaluable to the study of genetic driver and is the reason it so so complex trait analysis is so difficult in Human and wild population. Humans already have several additional layers of redundancy and regulation as compared to mice. If one was to compound the effect of mixed diets, family and a myriad of unknown contributing factors, determine genetic drivers for phenotypes becomes quite difficult.

What is Not Heritability

The exact meaning and interpretation of heritability are sometimes erroneously thought as the portion of a phenotype that is genetic CITE Visscher,2008. Rather it is the proportion of phenotypic variance that is due to genetic factors. Moreover, heritability is a population parameter and, therefore, it depends on population-specific factors, such as allele frequencies, the effects of gene variants, and variation due to environmental factors. As such, applying the concept of heritability to individuals is not appropriate as individuals do not vary.

As illustrated in the example of mouse weights, heritablity does not necessarily predict the value of heritability in other populations or other species. Within the two diet cohorts, although values of observed heritability were similar they were not the same, despite the use of mice with identical genetic backgrounds CITE Visscher,2008.

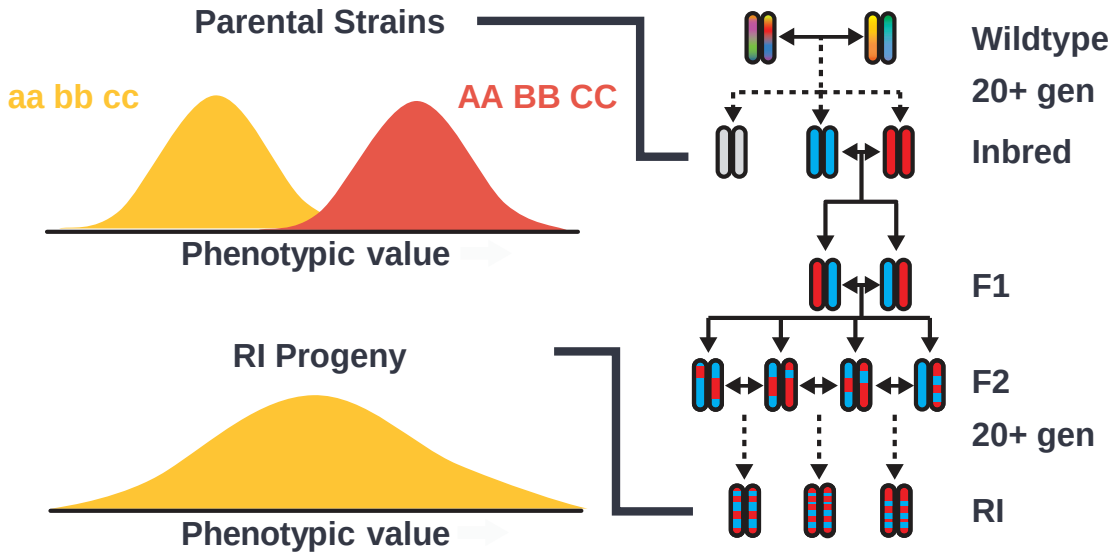
Heritable trait is heritable, if its core physiology and behavioral trait distinction

Applications of heritability estimation are broad and cross a range of disciplines, from evolutionary biology to agriculture to human medicine. In humans, estimation of heritability has been applied to diseases and behavioral phenotypes (e.g., IQ), and it has helped establish that a substantial proportion of variation in risk for many disorders, like schizophrenia, autism, and attention deficit/hyperactivity disorder, is genetic in origin. Despite these advances, studying complex trait and their heritablity in humans is significantly complicated due to the lack of conditional control and as such, this study uses a recombinant imbred population of mice.

ADD LINK ABOUT HERITBALITY = MAPPABLE TRAIT

1.7 What is QTL Analysis?

A Quantitative Trait Locus (QTL) analysis is a statistical method that links is aimied at linking continuous phenotypic trait measurements and genotypic molecular markers, in attempt to explain the variation observed to the genetic differences.



This BXD population study contains plethora of rich phenotype data and metabolites at the time of writing.

1.8 QTL vs. GWAS

Forward genetics, in which many individuals that differ in genotype are screened for phenotypes of interest, has been a hugely powerful tool to address such questions. In general, the raw genetic differences being screened are obtained either by mutagenesis or sampled from a natural population. Any phenotypic differences identified are connected back to the underlying causative loci via various mapping approaches including Quantitative Trait Locus (QTL) mapping and Genome-Wide Association Studies (GWAS).

QTL mapping is method to identify regions of the genome that co-segregate with a given trait either in F2 populations or Recombinant Inbred (RI) population. The key components of the cholesterol metabolism pathway in BXD mice have been dissected in this way (E. G. Williams et al., 2016). Despite this success, QTL mapping suffers from two fundamental limitations; only allelic diversity that segregates between the parents of the particular F2 cross or within the RI population can be assayed (R. W. Williams and E. G. Williams, 2017), and second, the amount of recombination that occurs during the creation of the RI population places

a limit on the mapping resolution. Resolution can be dramatically improved with several generations of intercrossing when establishing the RI population, e.g. advanced intercross RI. Additionally, allelic diversity within a mapping population can be increased up to a limit by intercrossing multiple genetically diverse founder before establishing the RIs, e.g. the collaborative cross RI which have 8 founder mice compared to the 2 in BXD.

Nevertheless, the allele frequencies and combinations present in any such lab population will differ from those in the natural population (R. W. Williams and E. G. Williams, 2017). For many applications this does not present a problem, but it does confound the analysis of epistasis for example, and offers only a limited view of the functional diversity present within the natural population(CITE MAGE).

GWAS overcome the two main limitations of QTL analysis mentioned above, but introduce several other drawbacks as a trade-off. The basic approach in GWAS is to evaluate the association between each genotyped marker and a phenotype of interest that has been scored across a large number of individuals. This approach was pioneered in human genetics [citation], with nearly 1,500 published human GWAS to date [citation]. GWAS are now routinely applied in a range of model organisms including such as mice [citation], and to non-model systems including crops [citation] and cattle [citation]. Generally, after identifying a phenotype of interest, GWAS can serve as a foundation experiment by providing insights into the genetic architecture of the trait, allowing informed choice of parents for QTL analysis, and suggesting candidates for mutagenesis and transgenics. Thus, GWAS are often complementary to QTL mapping and, when conducted together, they mitigate each others limitations[citation].

HOW TO VALIDATE QTL

INTRO TO QTL

collaborates 6 mus muslus domenticus and 2 from other sub sepcies too much varaition in this stock actually poses breeklding problems eventhough they have all these variants it causes problem. Ergo more varioation not always better

1.9 RI Mouse Population for GxE

GxE interaction are the effects on phenotype are partially determined by the interplay of genetics with diet, environmental stressors, age, pathogens, drug exposure, and differences in social interactions. Mice and other inbred and isogenic model organisms are extremely well suited to evaluate complex experimental effects in the context of QTL mapping. The ability to impose well-controlled perturbations across large cohorts is among the strongest motivations to use model organisms. This kind of design is already the most common

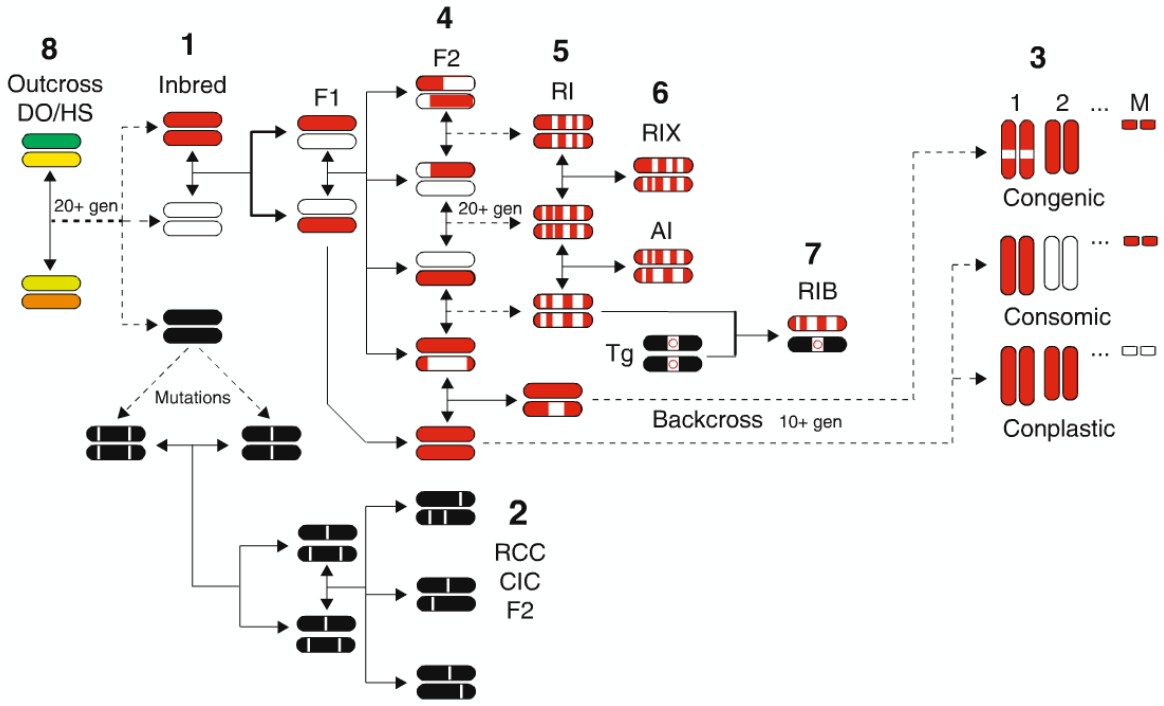


Figure 1.3: Produciton of RI stock and validation of QTL results in congenic lines

and critical in agricultural genetics.

As discussed about the most important disadvantage of conventional RI strains and other standard two-parent crosses is that they segregate for only a fraction of all known polymorphisms. For example, the BXD family segregates for a total of $\tilde{5}.2$ million sequence variants about 44 % of common variants among standard inbred strains [citation]. Some stretches of the genome will be almost completely identical by descent [citation] and these regions will not normally contribute much to trait variance. This disadvantage however may also be viewed as an advantage when trying to dissect a QTL, since the load of polymorphisms within an interval will be about sixfold lower than that of the corresponding interval in the collaborative cross batches, and thus the number of viable candidate genes may be much reduced.Phenotypes that map into these genetic blindspots can be particularly easy to map to QTNs [Li et al.].

1.10 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 strech 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.

1.11 Study Design

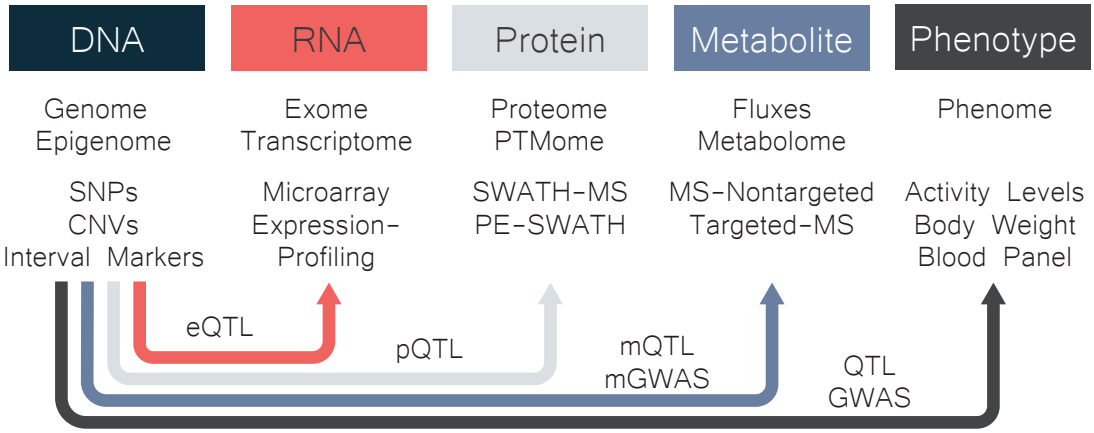
1.11.1 Components of Chow and High Fat diet

Regular chow is composed of agricultural byproducts, such as ground wheat, corn, or oats, alfalfa and soybean meals, a protein source such as fish, and vegetable oil and is supplemented with minerals and vitamins. Thus, chow is a high fiber diet containing complex carbohydrates, with fats from a variety of vegetable sources. Chow is inexpensive to manufacture and is palatable to rodents[citation] In contrast, defined high-fat diets consist of amino acid supplemented casein, cornstarch, maltodextrose or sucrose, and soybean oil or lard, also supplemented with minerals and vitamins. Fiber is often provided by cellulose. Chow and defined diets may exert significant separate and independent unintended effects on the measured metabolites, proteins and transcripts.

1.11.2 Study Design: Mouse Sex

Among the differences between the sexes in mice, one of the most pronounced contrasts to humans is the rapid estrous cycle mice experience. In humans, the reproductive cycle, called the menstrual cycle, lasts approximately 28 days, in rodents this cycle, called the estrous cycle, lasts approximately 4-5 days. Although this 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. Estural cycle data is not included in the phenotypic observation of the mice and as a result cannot be reliably excluded.

In the previous large BXD mouse experiment undertaken in the Aebersold, Auwerx and Zamboni lab, only male mice were used due to their larger size and lack of estrous cycle. The expansive literature of mice physiology is however significantly biased, using only mice for the aforementioned reasons. Female mice however, analogous to female human have longer life span on average and it was seen as pertinent to identifying anti-aging mechanisms. This study to maximize the number of novel aging related features, a majority mice population was used.



Male mice mice are extremely territorial and will even identical twins to the death if housed in the same cage. Evolutionary biology paradigm of . To keep males separates however then shortens their lifespan as these social animals,

Barbering is a characteristic social interaction among C57BL/6 and C57BL-related mice (C57BL/10, C57BR, C57L, C58, and C57BL congenic strains) which can also be seen in females. It is an expression of dominance. The dominant mice physically nibble or pluck fur and whiskers from their cage mates.

Female mice are known to lived shorted when they given birth to pup

Mouse don't breed after 1 year, go through menapause?

Mixing gneder is akin ot mixing dits, half the power of the study. Had been 2012 first study in cell shows that there are many difference between the sexes, draw back buit traits studied are very different, resrouce intensite to do each specific study

1.12 Why Multiple Omics

With the advent of 1000s of human genome and significant additional SNP variants mapping becoming available to public researchers Genome-wide association studies (GWAS) have become an important tool for evaluating the association between common genetic variants and risk of disease. Through signification efforts, thousands of disease associated SNPs have been found(A. D. Johnson and O'Donnell, 2009). The results of many GWA studies however, are not followed with investigations of the underlying mechanisms and explain only a limited amount of heritability. Additionally, most SNP variants are associated with small increased risk probabalities of disease and thus have weak predictive value by themselves without any

mechanistic insights.

With maturation in omics technologies, orthogonal high-throughput approaches can be used for mechanistic investigation of disease associations. Integration across different genomics, proteomics, transcriptomics and metabolomics can comprehensively determine the consequences and relationships between genes and protein expression and metabolite concentrations, and their phenotypic repercussions. Owing to the fact that transcript, protein and metabolites levels have only a modest correlation with each other, and that metabolites can be further modified by enzymatic processes and can originate from and be modified by various internal and external stimuli, good experimental design and optimization are necessary to increase the resolving power of these investigations.

Chapter 2

Metabolomics

2.1 Introduction to Non-Targeted Metabolomics

The metabolism is the sum of all the biochemical reaction that occur in an organism. Metabolites are shuttled through multiple enzymatically catalyzed reactions in the fundamental processes of energy production, growth and integrates 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 studying using multiple targeted and non-targeted techniques.

The most common analytical techniques used to identify the global consortium of metabolites present in a cells or tissue are NMR and untargeted mass spectrometry based. Additional chromatographic separation steps can be used to differentiate between specific set of compounds inseparable with a single mass spectrometery step. The goal of non-targeted metabolomics is to detect and quantify as many metabolites from a single extracted sample without prior knowledge of the specific composition of the metababolome. However, 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. Therefore, metabolomics analyses are always focused on a part of a given metabolome, for exmaple small polar molecules which are the focus of this study or the lipids extracted from a biological sample. Such an inevitable focus is the reason for the rather independently evolving disciplines of small polar molecules metabolomics, lipidomics, glycomics and other metabolome related approaches (Aksenov et al., 2017). Although the types of metabolites recovered in different experiments may be different, they result in complex data that require computational tools to process the raw signal from the mass spectrometer into peaks that can be assigned a normalized intensity and a chemical annotation, to determine 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.

In contrast to un-targeted MS-metabolites, targeted metabolites operate on prior knowledge and hypotheses and quantify a molecules of specific properties of masses. Chromatography is optimized for the extraction and separation of target metabolites and pathways such as the hexose sugars in glycolosis or lipids with equals masses but differences in the locations of their unsaturations. Targeted analysis can therefore be used as a follow up to untargeted metabolomics in order to validate ones hypotheses or quantify specific functional isobars 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. 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 due to its technical difficulty and very expensive.

2.1.1 Metabolomics Methods

Metabolomics combines analytical chemistry, platform technology, MS with sophisticated data analysis for deconvoluting dense . It involves the application of advanced analytical tools to profile the diverse metabolic complement of a given biofluid or tissue. Metabolomics 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, and combination with advanced and high-throughput separation techniques can reduce the complexity of metabolite separation, while MS-based techniques require a sample preparation step that can cause metabolite loss.

The types of samples that can be analysed using metabolomics are wide-ranging and include tissues, cells and biofluids. Tissue analysis, in particular, is perhaps the most powerful approach for studying localized and specific responses to stimuli and pathogenesis, yielding explicit biochemical information about the mechanisms of disease. Traditionally, tissue analysis involves extraction of the complete tissue material into a liquid form, from which the metabolite changes are averaged across the different cell types and regions of the analysed organ. In addition to this total tissue analysis, subregional, cellular and even subcellular metabolite profiles can provide further insight into structure-to-function relationships; this is particularly valuable in the case of heterogeneous tissues such as brain and cancers17. Simultaneous sampling of arterial blood (entering the organ) and venous blood (draining the organ), followed by paired analysis, can also have value in the

investigation of tissue metabolic activity16. This paired arteriovenous approach provides information about the metabolite uptake and release patterns across the tissue of interest and therefore gives insight into tissue metabostasis. The power of this paired analysis allows for the measurement of metabolite arteriovenous differences or ratios and offers a compelling compromise with sampling effort, compared to the traditional approach of venous blood analysis.

During the past few years, substantial progress has been made in metabolomic analysis by improving instrument performance, experimental design and sample preparation, ultimately facilitating broader analytical capabilities. Moreover, the surge in new chemoinformatic (computational approaches for handling chemical information) and bioinformatic (computational approaches for handling biological information) tools has provided extensive support for data acquisition, analysis and integration. This has greatly enhanced our ability to identify metabolites in various samples and allowed us to correlate these metabolites with particular phenotypes, thus establishing useful biomarkers that are indicative of particular physiological states or aberrations. The ultimate challenge now is to move beyond simply identifying metabolites and using them as biomarkers, and to start establishing the direct physiological roles of metabolites and their involvement in metabolic networks, as well as determining how changes in their levels are implicated in different phenotypic outcomes. This Innovation article focuses on how this most relevant hurdle for metabolomics can be overcome. We describe how advances in technologies that are used in metabolite identification and analysis, experimental design and pathway mapping are helping us to gain more meaningful data, revealing important nodes for further investigation. We also discuss how this information, when combined with traditional biological methods, can enable us to ascertain molecular mechanisms and begin to infer biological causality.

Summary of Metabolic Data Acquisition & Analysis Timeline

Mouse livers are pulverized into granules in liquid nitrogen. Each liver is weighed for sample normalization by weight and small polar molecule extraction is used to extract samples for mass spectrometer analysis

Sample Preparation Extraction

Experimental Optimization

Metabolites are highly diverse unlike proteins and successful detection of complex chemistry requires optimization of experimental protocol

Raw Data Acquisition and Conversion

Take the raw machine level intensities stored in proprietary formats and convert them to open source mzXML or mzML format to allow preprocessing using Matlab

Preprocessing

Detection of ion features is across of the data set generated is done in Matlab which detects, quantifies and aligns peaks throughout the samples

Centroiding and Quality Control

The aligned peaks are centroided, and corrected for bias from the instrument or batch effects. Data imputation may be performed to extract data from low intensity peaks.

Annotation

Annotate molecules that have been detected using search engines of libraries of known molecules and adducts in Online databases

Univariate Statistics

Determine whether ion are differentially represented in different Age and Diet cohorts. Parametric and Non-parametric test for significant differences employed

Multivariate Statistics

In order to extract information from the set with thousands of ion measurements: Clustering, classification and Dimension reduction and visualization used to summarize data and determine relationships within

Metabolic Network and Enrichment

Using Metabolic Network analysis to investigate correlation networks, various effects on pathways in order to determine structural an causal relations in the metabolic data

QTL Mapping, Biomarker detection MultiOmics Integration

In a refined set of target pathways performed QTL and biomarker analysis using multiomics data sets in order to enrich and validate observations in the metabolic level

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 homogenous,comprising mostly of hepatacytes which can be effectively ground to a powder in N_{2L} which enable rapid extractions. The power if liver cells are 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 pulverized tissue.

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 through a combination of the MeOH and ACN in the extraction solution denaturing the proteins in the sample and low temperatures in long extractions or short extraction times in high temperature extractions 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 so small scale pilot studies were used to determine whether the metabolite extraction protocol used in the previous paper(E. G. 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. Moreover, differential detection of key discriminant metabolites (such as N6-methyl lysine found in significantly higher quantities in chow diet mouse metabolites) and respective known QTLs which were observed in previous publications of the BXD mouse data were used to benchmark our current protocol (E. G. Williams et al., 2016; Wu et al., 2014).

1. Pilot Study 1 - the first pilot study is used to determine the difference between hot and cold metabolite protocol and exraction 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 converge.

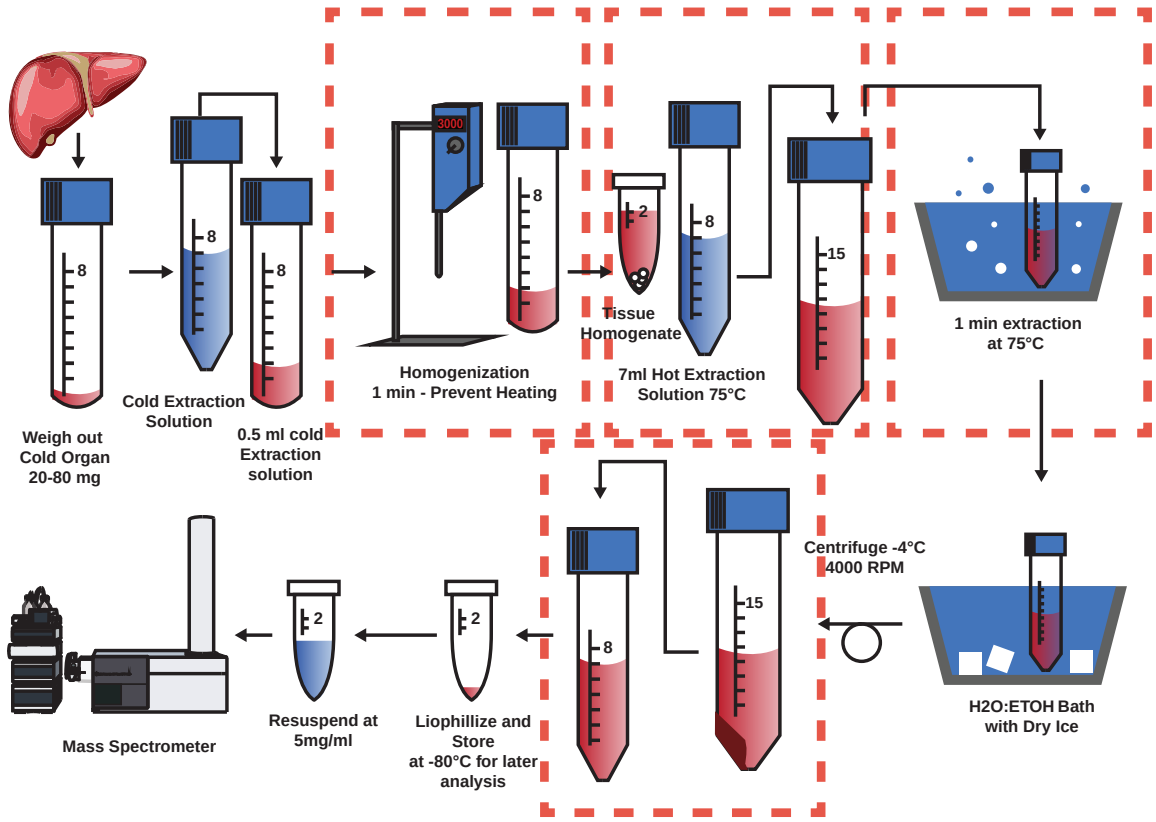


Figure 2.1: Hot Polar Metabolite Extraction Protocol

Hot Extraction:

In a hot extraction protocol, liver samples kept at -20° on dry ice are weighed into cell culture or falcon tubes. The tubes must be than 10cm long to allow for the homogenization head to read the bottom of the tube. 0.5mL of extraction solution is added to the samples. The extraction solution is sufficiently micelle disrupting and dissolves cellular membranes. A homogenization step using a laboratory-grade blender is included in the protocol to further lyse cells 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 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:

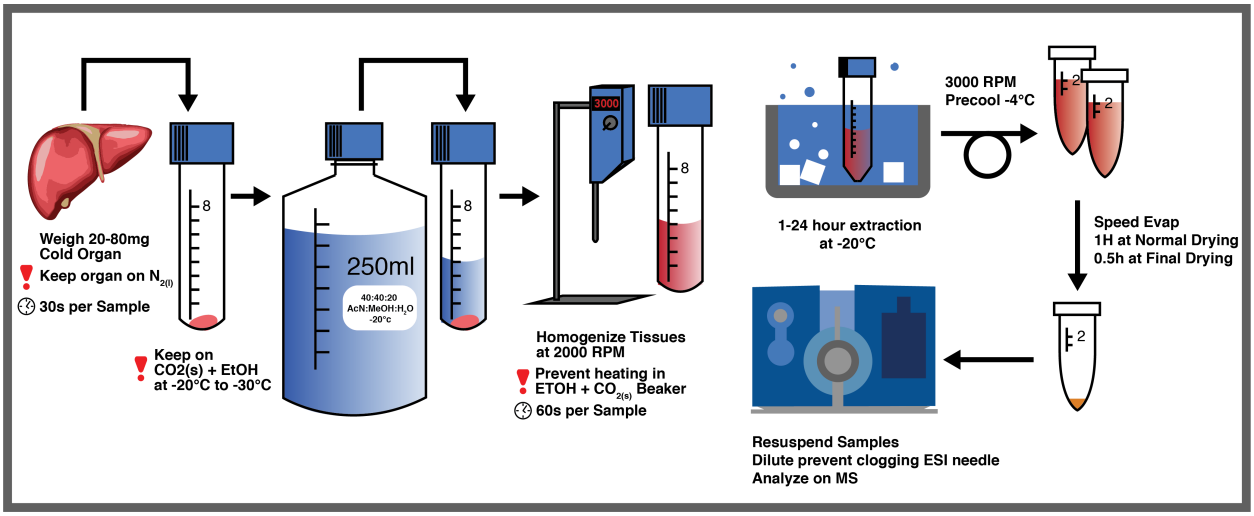


Figure 2.2: Cold Polar Metabolite Extraction Protocol

The cold extraction 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 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, the protocol described in (E. G. Williams et al., 2016) an additional homogenization step with a laboratory mill ball or laboratory-grade blender is called for after the cell lyses/quenching and extraction solution are 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 and the logistically complexity of performed ultrasonication on 600 samples precluded this homogenization.

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 introduce 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 the figure on the next page. In the third column, the Hot:H1 column compares the intensities seen in the standard

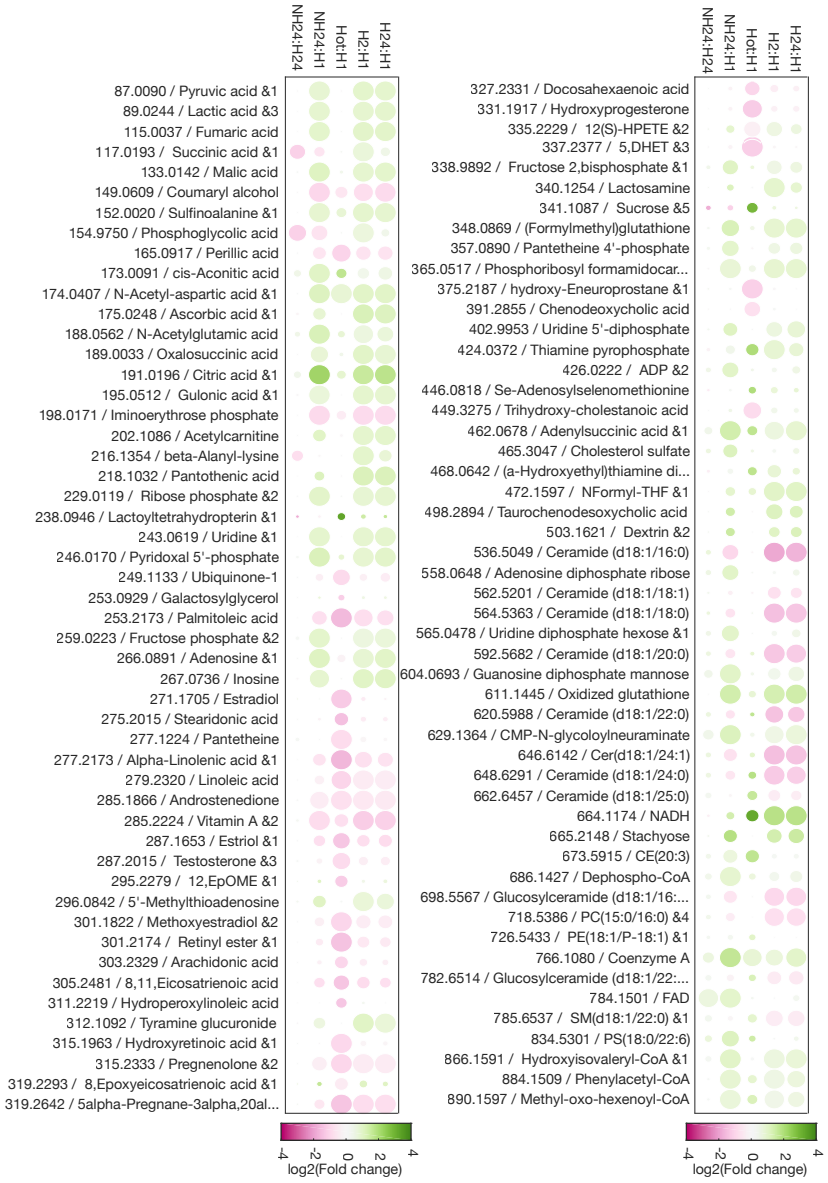


Figure 2.3: Metabolites Log_2 foldchanges between different extraction conditions. Hot - indicates the hot extraction used in the Science paper, H1 in the cold extraction with a 1 hour extraction time, H24 also the cold extraction but with a 24 hour extraction time, NH24 is the cold extraction with 24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

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. Almost all of the highlighted metabolites in the volcano-plot 2.4 are thus not or primary interest. In conclusion the use of either hot or cold metabolites extractions does not hugely bias the metabolites intensities.

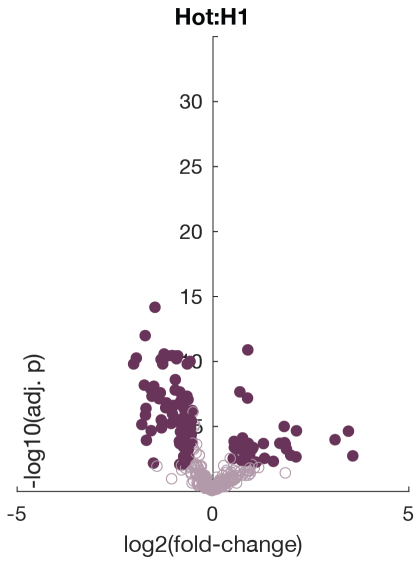


Figure 2.4: Volcano Plot of P-Values and Log2 Fold Changes seen between Hot Extraction protocol and Standard Cold Extraction Protocol

Time Dependence of Cold Extraction

Results indicate the homogenization is not necessary to sufficiently extract polar metabolites from the cells. Additionally, cold and hot extraction both perform similarly in terms of the coverage of metabolites that are extracted, however the variation in the spectra is much larger in the cold extraction due to the longer processing times in which degradation products accumulate. In the figure below all annotation are displayed on

2.4.1 Extraction Time Performance

2.4.2 Effect of Homogenization on Performance

The results from the bubble plot show that homogenization does hot significantly alter the metabolites of interest in the study. This step is thus removed from the protocol

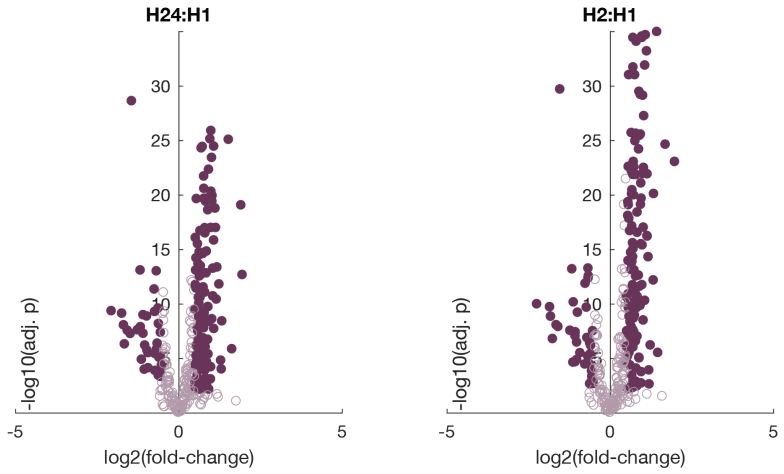


Figure 2.5: Volcano Plot of P-Values and Log2 Fold Changes seen between Standard Cold Extraction Protocols with an additional 1hour and 24 hour extraction period as compared to the standard cold extraction protocol

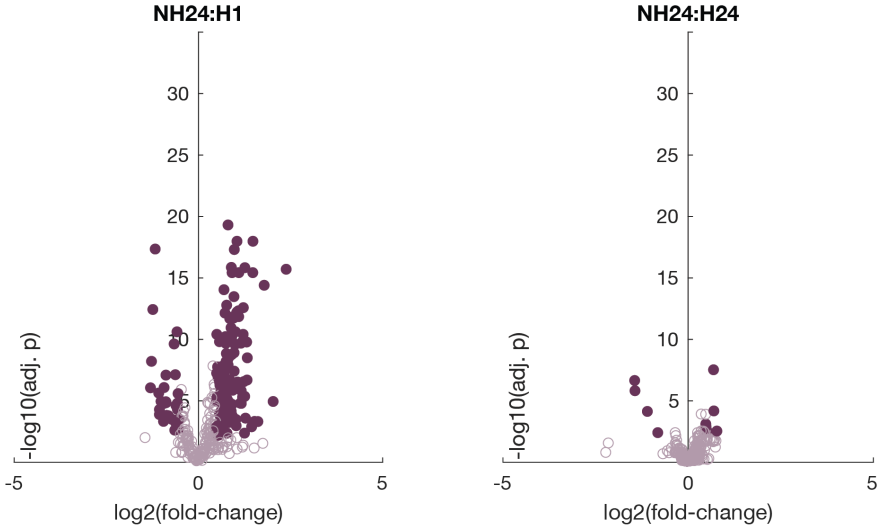
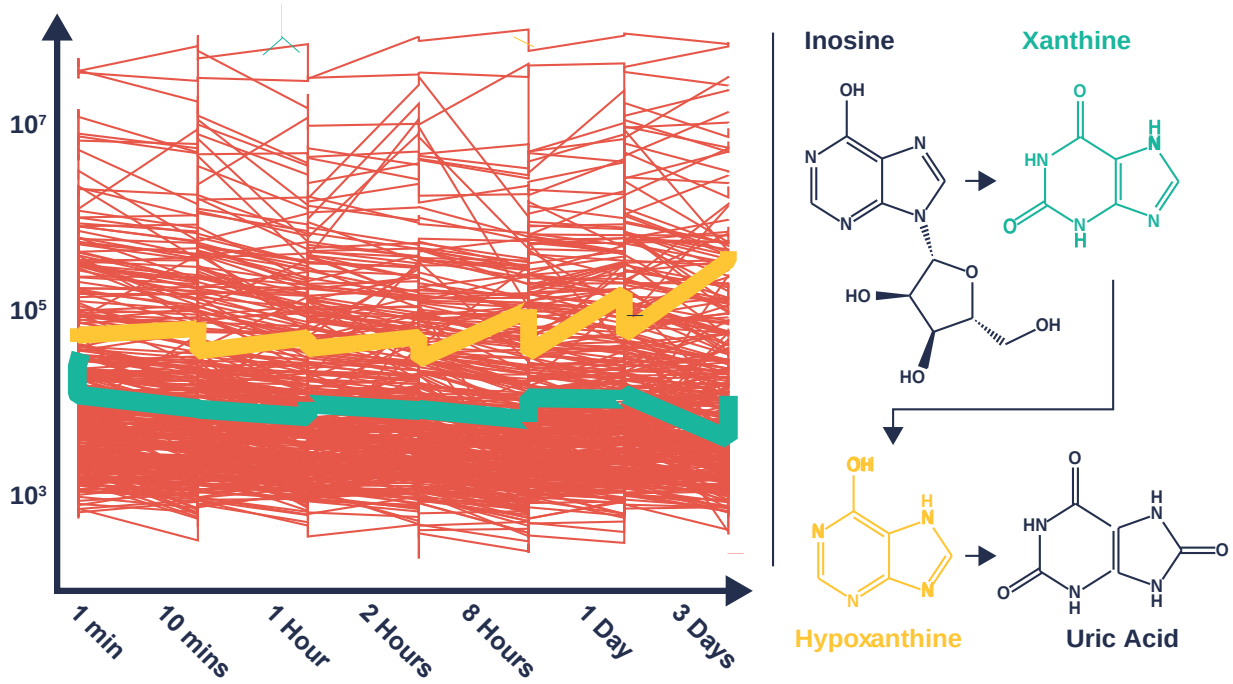


Figure 2.6: Volcano Plot of P-Values and Log2 Fold Changes seen between Standard Cold Extraction Protocols with an additional 1hour and 24 hour extraction perdion as compared to the standard cold extraction protocol

2.4.3 Extraction Times

2.4.4 Effect of Freeze Thaw Cycles

2.4.5 Pilot Study Results



2.5 Raw Data conditioning and Analysis

2.5.1 Centroiding Raw Spectral Data Analysis

Mass spectral signal in its rawest form is the continuous sampling at a given time interval of current at at the end of the flight chamber in a time of flight mass spectrometer. This continuous response is called raw scan data, profile mode data, or continuum data, depending on the particular MS instrumentation and vendor. What one actually obtains from any physical mass spectrometer with its finite resolving power, mass accuracy, and linear dynamic range is always a continuous response curve similar to the one shown in the 4 bottom panel of the figure below.

Similar to optical instrumentation where lines may broaden with incorrect optical arrangement and lens degradations, the complete profile of a peak shape function provides a numerical representation of the ion dispersion or aberration in the mass spectrometer, including spatial and velocity dispersion inside the ion source. 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 TOFMS instruments used in this experiment, the peak shape function show fast intensity spike with a gradual decay of the signal. This is due to the arrival isotopically pure species followed by lower abundance isomers with heavier isotopes of carbon, nitrogen and oxygen. Additionally, as molecules fly through the flight path, trace amount of gas many introduce a spread in the

velocity and energy of ion in a stochastic manner. As a result the peak function of an instrument has a unique signature in every instrument in a specific set of conditions which include the aformented effects in the mass analyzer(flight path), and voltages on all of the ion optics and is known as the MS Tune.

The FWHM defines the mass spectral resolution or resolving power, whereas the amount of shift in the position of the peak shape function defines mass accuracy. In the figure below, well defined defined peaks can be seen in the lower M/Z range highlighted in the boxes on the left side of the figure. The peaks and strong and can be aptly extracted by modeling them with a combination of gaussian functions. However high resolution of the instrument used is seen in the spectra on the right. However, the signal in the high m/z regime is quite low as seen in the centroided spectra shown in figure 2.1.A, and requires a high resolving power instrument to detect single peaks in the convoluted grass. Figure 2.1C show

To get sub-dalton unit mass resolution system, this calibration is accomplished through the measurement of internal or external calibration standards whose elemental compositions are known. This uncertainty from the MS peak shape can, however, be calibrated out and removed through a novel and comprehensive calibration process that involves not just m/z, as is the case for all conventional MS calibration, but more importantly the peak shape.

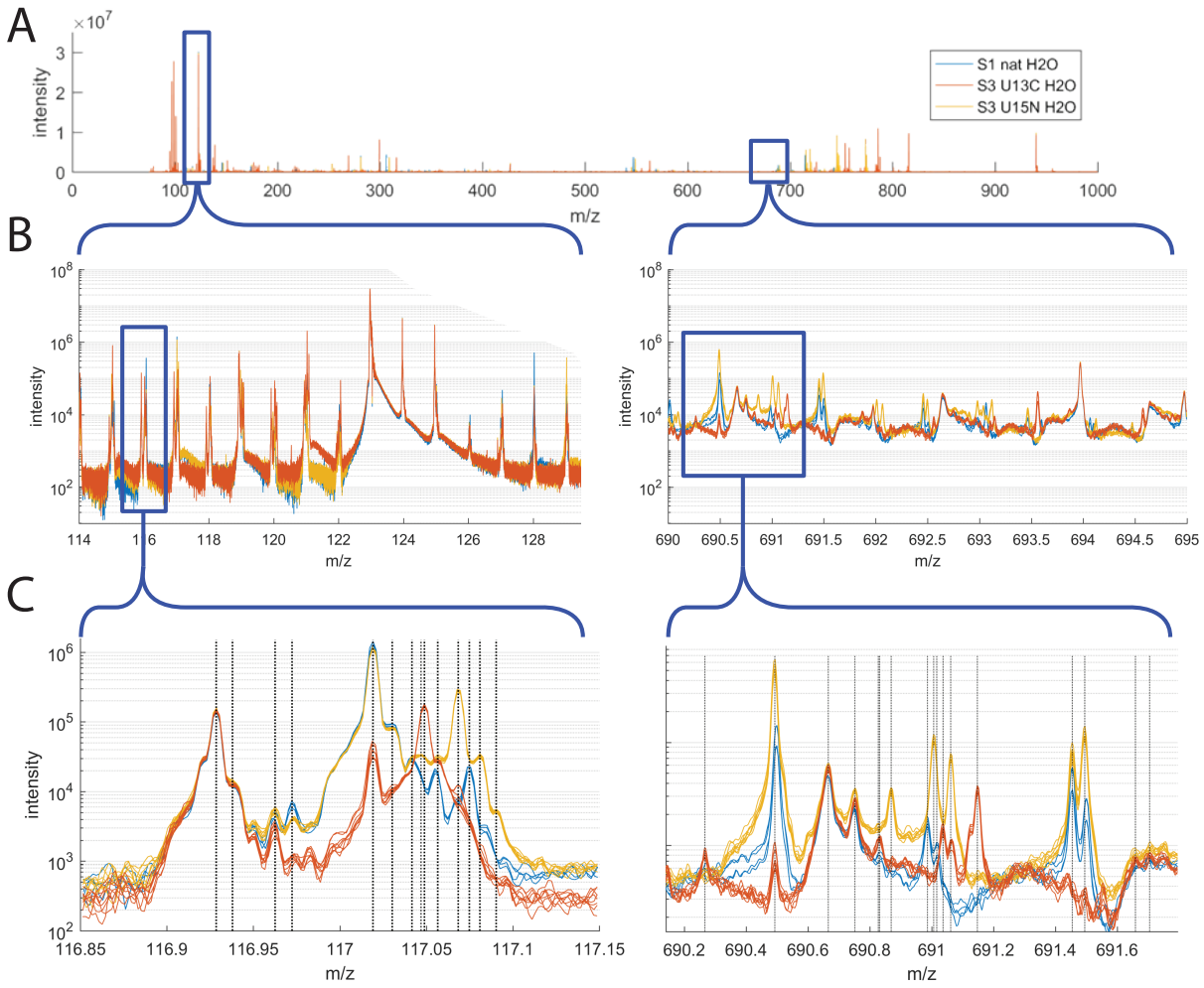
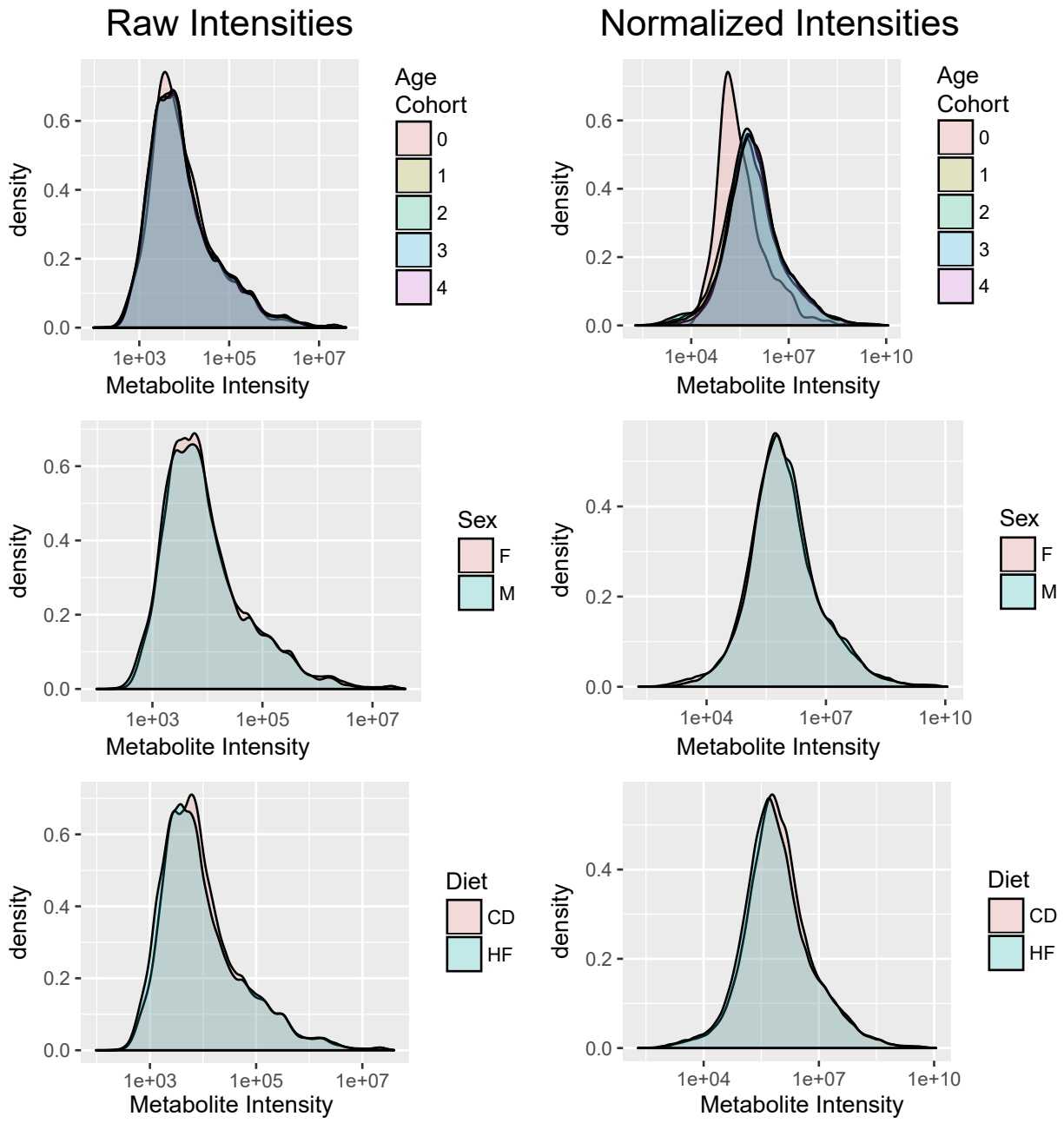


Figure 2.7: A.
B.
C.

All samples were run with two technical replicates which were monitored on the LC to ensure robust and reproducible spraying. High sodium and low total volumes in some samples yield inconsistency chromatograms but the majority of samples were seen to be robust. In the electro-spray 19 000 featured were detected, 4000 of which could be annotated with some certainty, the majority of which however remains complex high molecular weight material.


Raw Intensities
Normalized Intensities
density
Metabolite Intensity
1e+03
1e+05
1e+07
1e+04
1e+10
0.0
0.2
0.4
0.6
Age Cohort
0
1
2
3
4
Sex
F
M
Diet
CD
HF

2.5.2 Annotating Centroided Data

2.5.3 Normalization of the Raw Data

During analog to digital signal conversion detector and pre-amplifier electronic noise is recorded and appears in the data, as does any ugliness in the pulse shape. Absolute quantification is very difficult because the detector gain has a first order effect on spectrum peak heights. Detector analogue gain is a highly non-linear function of excitation voltage, subject to drift and difficult to measure unless extra hardware is present, such as a Faraday cup. In practice an internal standard is necessary (a mass peak to normalize against). In these metabolic study no house keeping metabolites are available for normalization nor are the samples spiked with an internal standard, as a results relative signal intensities are compared that than absolute counts.

Data handling tasks in metabolomics can be roughly divided into two steps: data processing and data analysis. The data processing step consists of low-level processing of raw data with signal processing methods and combining data between measurements. These tasks transform the raw data into format that is easy to use in the subsequent data analysis steps. The data analysis stage includes tasks for analysis and interpretation of processed data. This typically includes multivariate analyses such as clustering of metabolic profiles or discovering important differences between groups of samples. In proteomics, a similar classification has been proposed, while further dividing the data processing stage into low-level and mid-level

If you have variables that always get positive numbers, such as relative metabolite gene and transcription concentration , and that show much more variation with higher values (heteroscedasticity), a log-normal distribution might be a clearly better description of the data than a normal distribution. In such cases I would log-transform before doing PCA. Log-transforming that kind of variables makes the distributions more normally distributed, stabilizes the variances, but also makes a model multiplative on the raw scale instead of additive. That is of course the case for all types of linear models, such as t-tests or multiple regression.

2.5.4 Data Analysis

2.6 Differential Metabolites

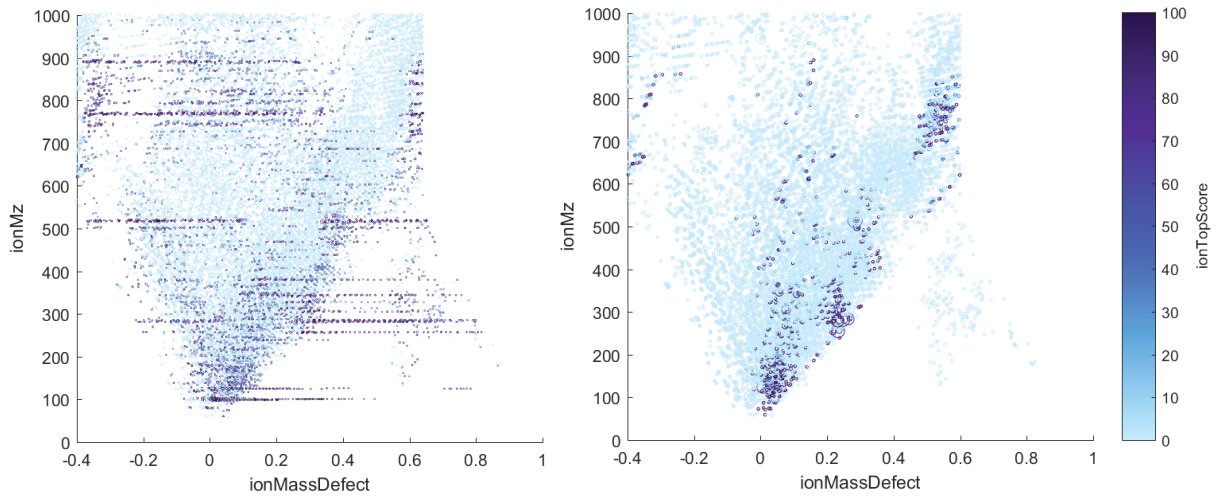


Figure 2.8: Left: All anntated peak annotated automatically. Right: Ringing peaks and impossible mass combination filtered

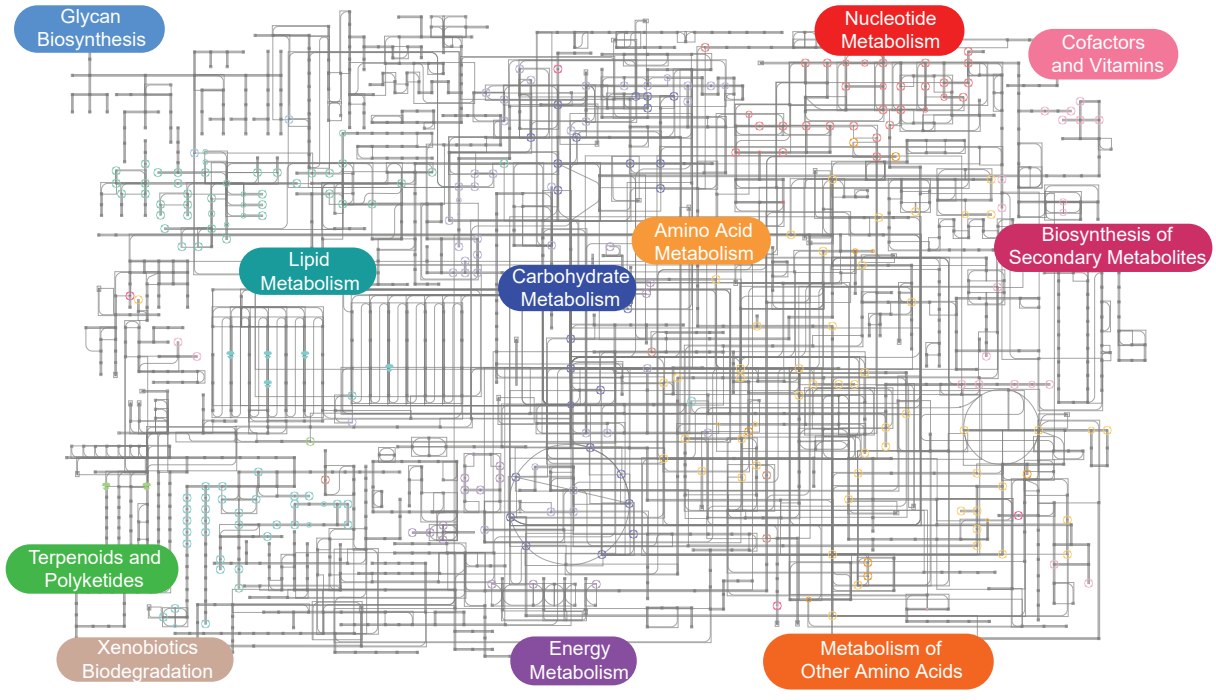
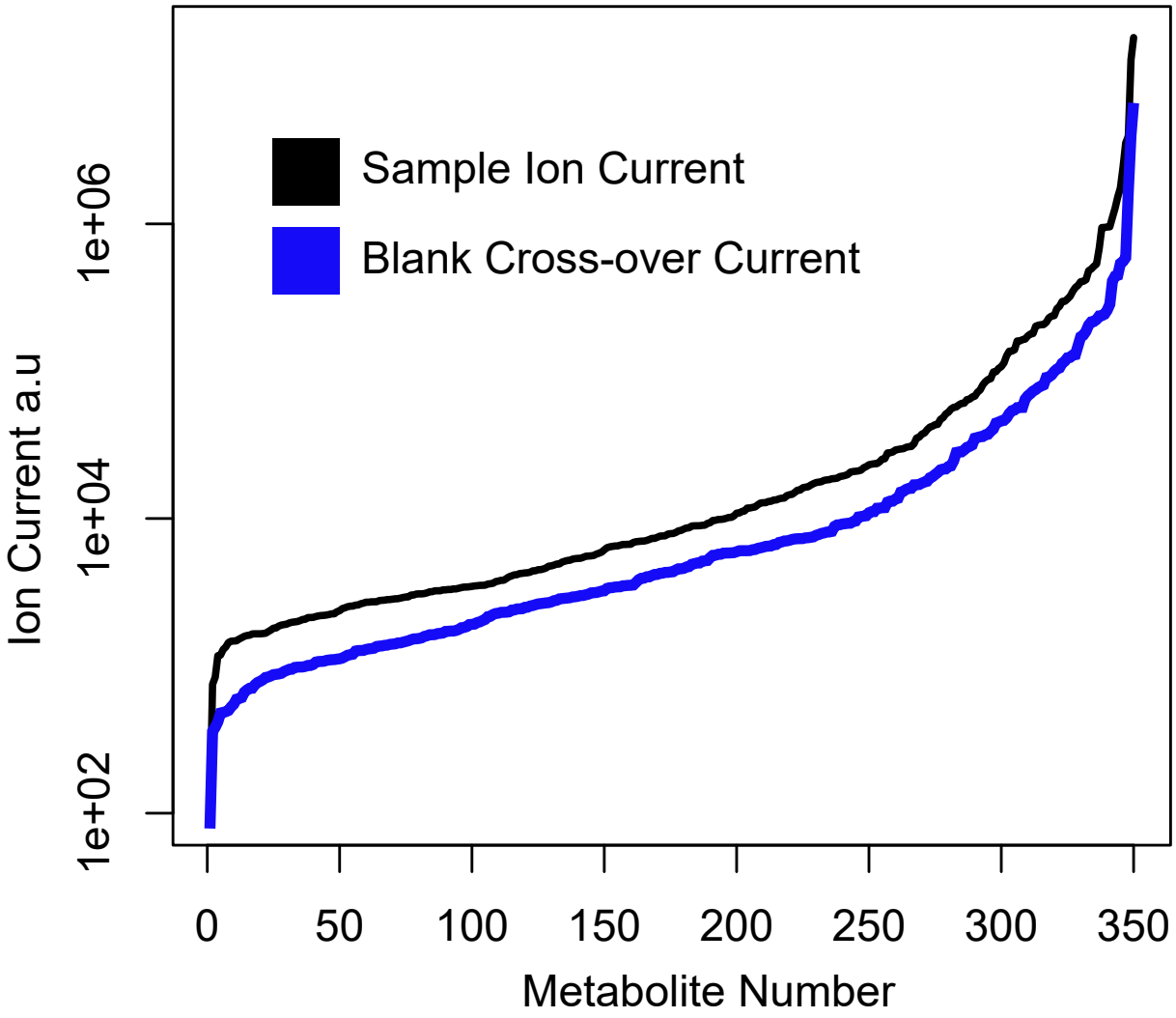
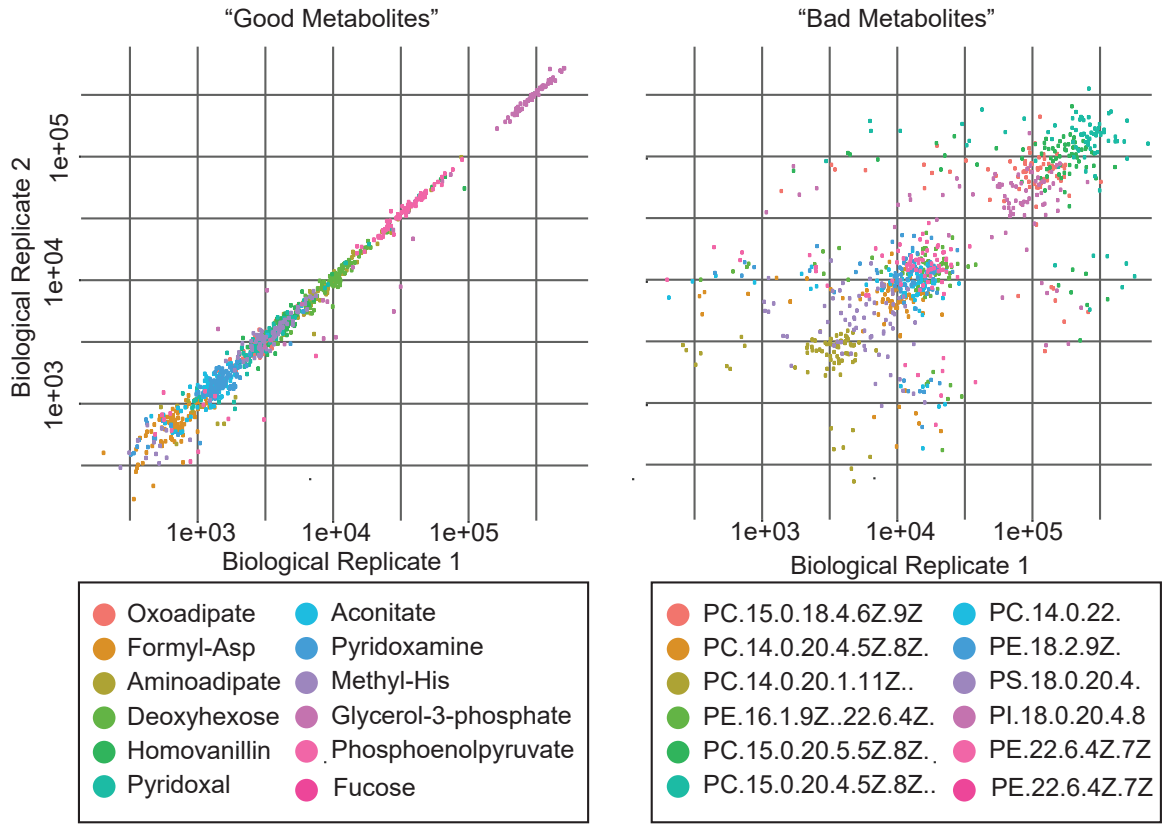


Figure 2.9: This map shows the


Sample Ion Current
Blank Cross-over Current
Ion Current a.u
1e+02
1e+04
1e+06
0
50
100
150
200
250
300
350
Metabolite Number



2.7 Analysis of Metabolic Data

2.7.1 PCA

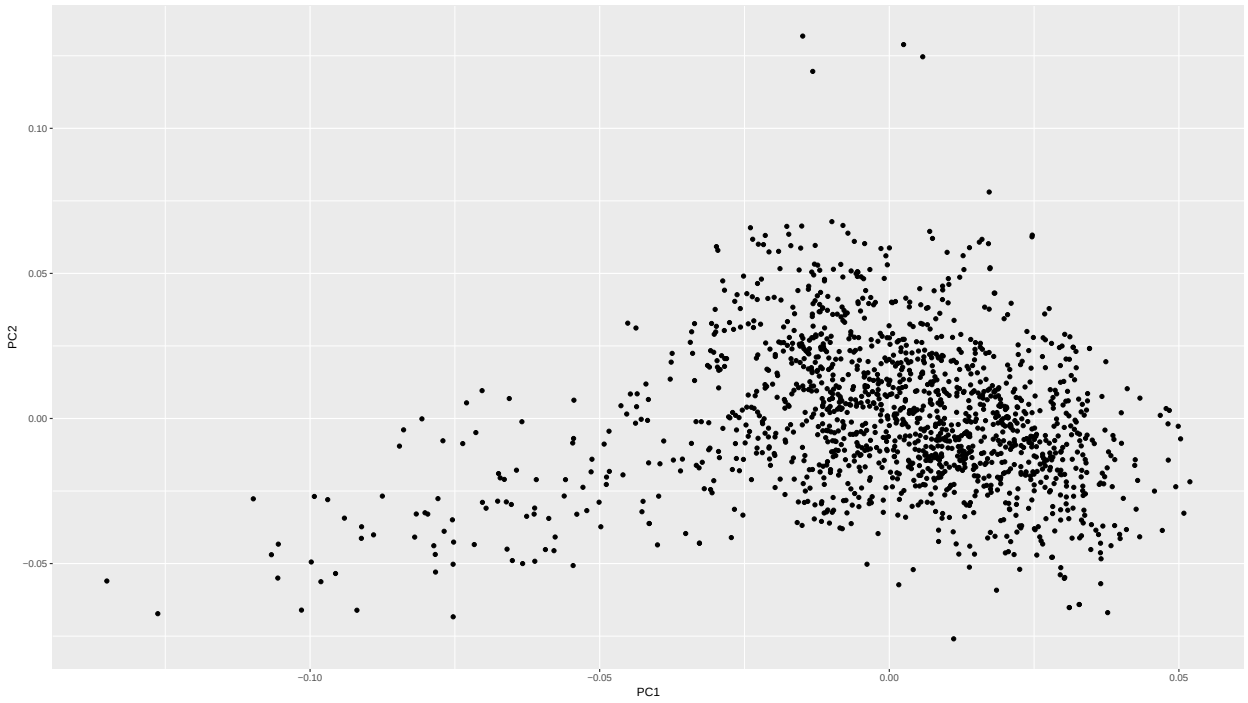
PCA constructs orthogonal,mutually uncorrelated,linear combinations that, explains as much common variation as possible, in a desending order. PCA can be done based on the covariance matrix as well as the correlation matrix as scaling the data matrix such that all variables have zero mean and unit variance, makes the two approaches identical.

2.7.2 Clustering

http://www.sciencedirect.com/science/article/pii/S0031320311003517

Algorithm:

1. Start with clusters, each containing only a single observation.



2. Find the nearest pair of distinct clusters, say and . Let =, remove and decrease the number of clusters by one.

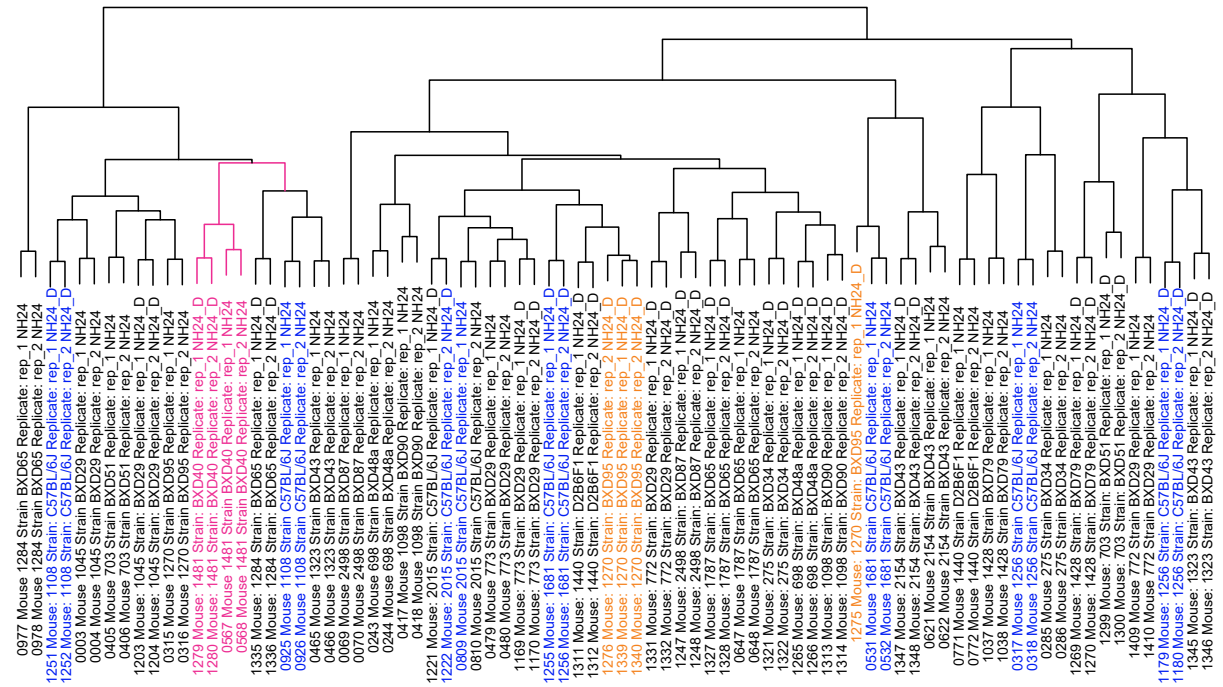
3. If the number of clusters equals 1 then stop, else go to step 3.

Observations that are grouped together at some point cannot be separated anymore later. By cutting the tree at a certain height, one obtains a number of clusters. Results depend on how we measure distances between observations and between clusters and as such Euclidean, Manhattan and

Hierarchical clustering, k-means and PAM based on the raw data may not be sensible when the variables are on very different scales. Metabolites with the largest range has the most weight and might dominate the analysis. As such, the variables can be are scaled are standaradized using the

The k-means algorithm is parameterized by the value k. the algorithm begins by creating k centroids. It then iterates between an assign step (where each sample is assigned to its closest centroid) and an update step (where each centroid is updated to become the mean of all the samples that are assigned to it. This iteration continues until some stopping criteria is met; for example, if no sample is re-assigned to a different centroid. The k-means algorithm makes a number of assumptions about the data, the most notable assumption is that the data is 'spherical,' see how to understand the drawbacks of K-means for a detailed discussion.

Agglomerative hierarchical clustering, instead, builds clusters incrementally, producing a dendogram. As the picture below shows, the algorithm begins by assigning each sample to its own cluster (top level). At each step, the two clusters that are the most similar are merged; the algorithm continues until all of the clusters



have been merged. Unlike k-means, specify a k parameter does not need to be supplied: once the dendogram has been produced, the tree can be cut at the level which is most interpretable.

https://math.stackexchange.com/questions/128255/what-is-the-correct-definition-of-minkowski-distance

The way the bootstrapping increasing the ability to determine the way the

2.7.3 Good QTL

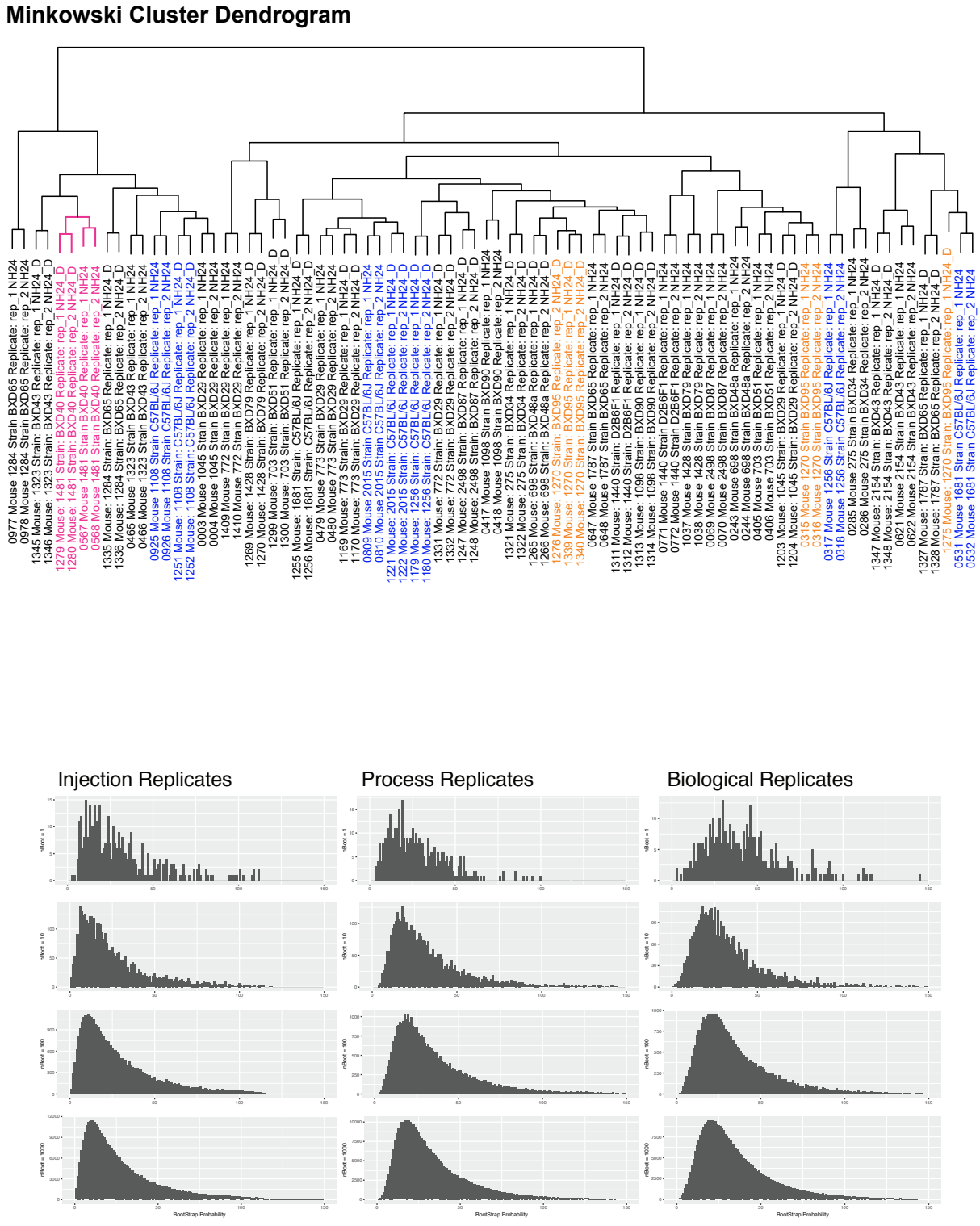


Figure 2.10: Boot Strap distribution of the CV between injections, Process Replicates and Biological Replicates

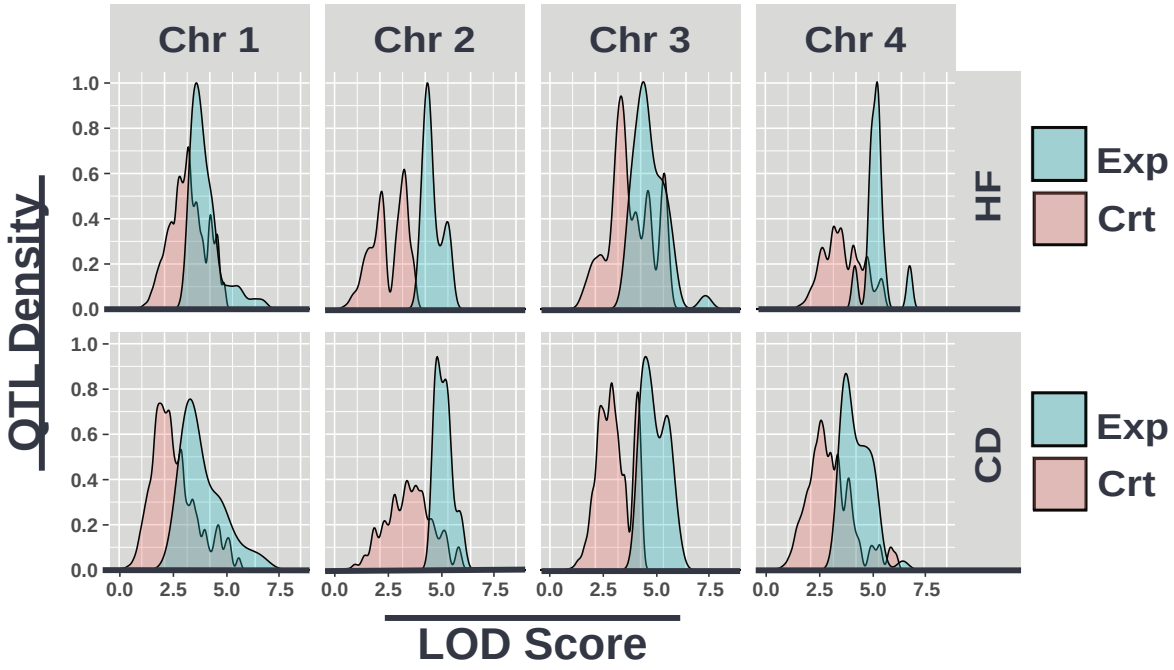
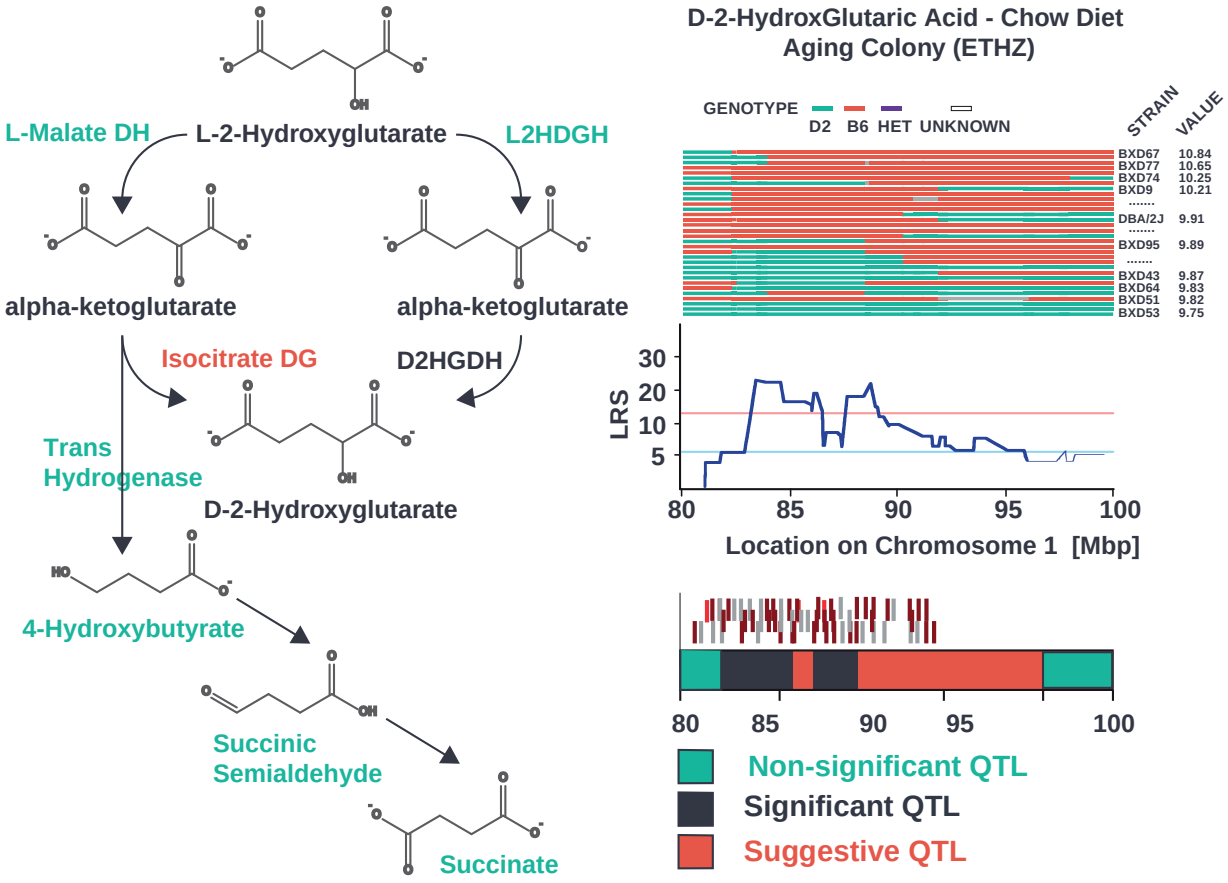

Chr 1
Chr 2
Chr 3
Chr 4
HF
CD
Exp
Crt
Exp
Crt
QTL Density
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5
LOD Score

Figure 2.11


L-Malate DH
L-2-Hydroxyglutarate
L2HDGH
alpha-ketoglutarate
alpha-ketoglutarate
Isocitrate DG
D2HGDH
Trans
Hydrogenase
D-2-Hydroxyglutarate
4-Hydroxybutyrate
Succinic
Semialdehyde
Succinate
D-2-HydroxGlutaric Acid - Chow Diet
Aging Colony (ETHZ)
GENOTYPE
D2
B6
HET
UNKNOWN
STRAIN
VALUE
BXD67 10.84
BXD77 10.65
BXD74 10.25
BXD9 10.21
DBA/2J 9.91
BXD95 9.89
BXD43 9.87
BXD64 9.83
BXD51 9.82
BXD53 9.75
LRS
30
20
10
5
80
85
90
95
100
Location on Chromosome 1 [Mbp]
Non-significant QTL
Significant QTL
Suggestive QTL

Chapter 3

Metabolite Set Analysis

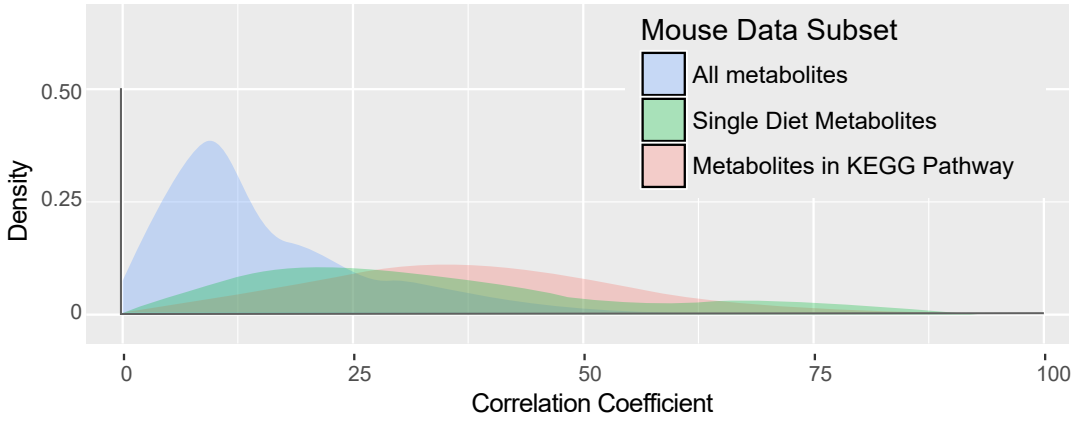
3.1 Introduction

Metabolite set enrichment analysis is used to identify patterns and in the changes of metabolites in a biologically meaningful way. Using prior knowledge to determine changes of metabolite concentration withing predefined pathways.

The classification is highly dependent on the quality of the prescribed ontologies, luckily *Mus Musculus* has an extended network of disease related and physiological networks within the metabolite analyst framework.

Higher level interpretations can be performed putting the metabolite concentrations in the context to their catalyzing enzymes and gene pathways.

The Global test algorithm is used on the back-end to power the statistical analysis in MSEA. In a dataset



where there are many factors observed for individuals and a single response, to determine which if the factors are associated with the response.

In comparison to the Log-Likelihood, the global test is not parameter invariant but gains an additional optimally such that at it is optimal in the neighborhood of the null hypothesis. In our case In which there are many more metabolites measured than individuals the likelihood ratio breaks down but the global test still functions often with good power.

3.2 limitations

3.3 Diet Related Metabolite Set Enrichment Analysis

3.4 Age Related Metabolite Set Enrichment Analysis

Chapter 4

Proteomics

Proteomics deals with structural and functional features of all the proteins in an organism. It is important to understand complex biological mechanisms including the mouses responses to stress tolerance. Age-related degeneration mechanisms involve stress perception, followed by signal transduction, which changes expression of stress-induced genes and proteins. Post-translational changes are also important in noise responses to abiotic stresses. A single gene can translate in several different proteins and a few genes can lead to a diverse proteome. Such inconsistency limits genomics and transcriptomic approaches more specifically, when post translational changes govern phenotype. Differential expression observed at the transcriptional (mRNA) level need not be translated into differential amounts of protein. To address this, several proteomic studies have been performed to understand abiotic stress tolerance mechanisms in Mice.

4.1 Introduction to MS Proteomics

What is DIA as an alternative to DDA or SRM. Ideally in the proteomic method you should be able to quantify a large set of protein, across multiple samples and have a consistent number of protein quantified, with high accuracy, reproducible and sensitive. Compared DDA has a reproducibility problem with fast scanning instruments, you can quantify many proteins but they are not as reproducible or reliable. SPRM a classical targeted proteomics is much high compared to DDA but the number of protein you can quantify is relatively discrete. With Shotgun, many protein can be determined, however there are many gaps between samples. With SRM there are a few number of protein quantified but with higher reliability

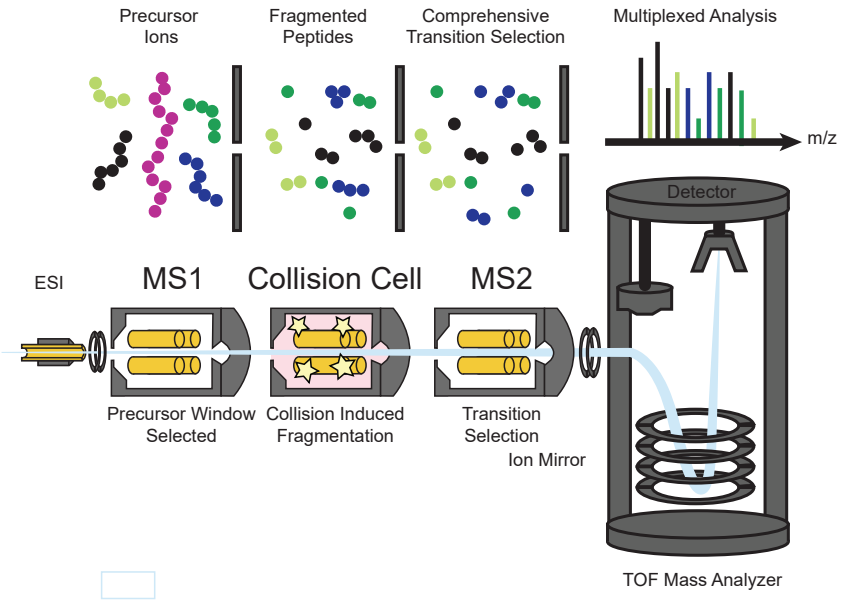


Figure 4.1: Schematic of AbSCiex 5600+ Triple TOF Mass Spectrometer

What is Data independent acquisitions

Data independent acquisition is a method of acquire mass spectrum data in which the signal intensities of the MS1 are not used to determine which peptides precusors are selected for further fragmentation(Venable et al., 2004). This is in contrast to Data Dependant mass spectrometric techniques in which fragment which show the highest intensity in the MS1 space as they elute off a column as selected my the machine in a duty cycle to be fragmented and further analyzed in the MS2 space. The reason this is unwanted, is because the signal intensities in the MS1 space can be high stochastic meaning proteins may not always be selected by the mass spectrometer between samples if they do not show the same high intensity peaks. Another key difference between DIA and DDA methods is that a SWATH machine runs on a fixed duty, where swath acquisition windows are programmed to shift. Normally with the wide precursor isolation windows, comprehensive sampling of the precursors in terms of their MS2 spectra. In order to do this, all fragments from precusors are analyzed together generating complex MS2 spectra that requires computational tools to deconvolute. Many DIA methods have been developed since the first mention in the Yates paper in 2004. All of the techniques that are nominally called DIA follow in 2012 faster instrument and smarter data.

In the figure below the red and blue peptides can be use to illustrate the SWATH sampling of the precursor space. Each consecutive window moving in the z-axis out of the page shows an MS2 spectra at subsequent time points. 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 an enlarged precursor isolation space of 5-20 m/z is used and selects both the red and blue peptides for

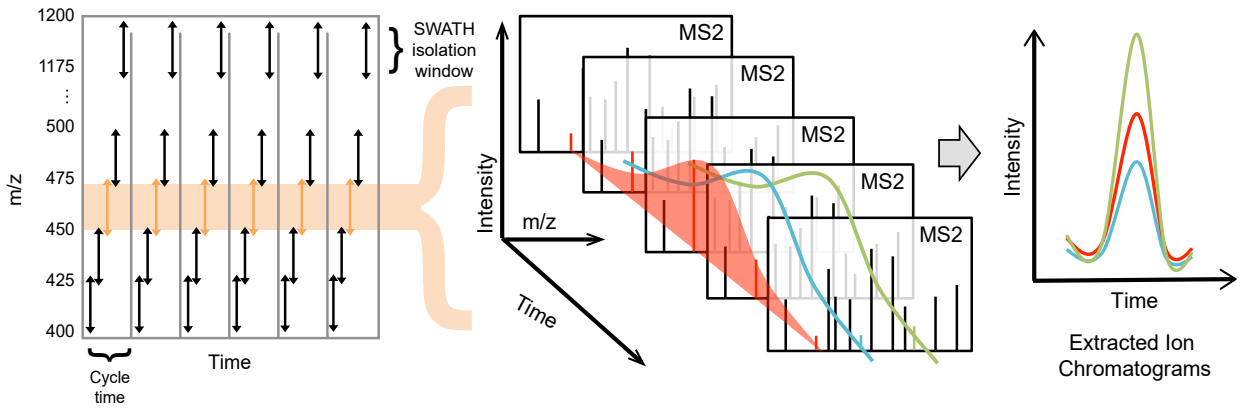


Figure 4.2: An Illustration of the SWATH-MS Duty cycle. (1) As peptides elute off an orthogonal chromatography column into the mass spectrometer, a window of mass ranges OR SWATHS are isolated and fragmented. (2) The ions that results from the fragmented peptides is recorded as a convoluted spectrum including fragments from the three **Red, Green and Blue** peptides. For each of 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.

fragmentation in the collision cell. Both peptides are co-isolated, co-fragmented and co-analyzed leading to the MS2 two windows seen in the figure. As a result, fragmentines from all the precursors are present in the final multiplexed data. The continuous monitoring of the peptides in both in terms of time and MS2 signal thus comes at the cost of added complexity in separating and quantifying the signal from different precursor peptides. When the instrument the is forced to fragment all the precursors in every duty cycle within the limit of detection, the result is a consistent quantification of all precursors in the sample.

DIA SWATH Operation

SWATH MS stands for serial windows

peptides elute off the column, a wide precursor isolated windows are selection and all of the fragments that can be found in a 5-25 MZ interaction, can be isolated and fragment. This yeileds a very complexes MS2 spectra. There is comprehensive samples within the window but highly convoluted int he MS2.

By determining all the fragmentariness than are known for a single peptides, the multiplexes recording of the fragmentines minutes can be deconvolutied.

in the figure below, an example of the data acquired from SWATH Acquisition can be seen. The x-axis represent the chromatographic dimension and the y-axis represents the mass charge space. The signal intensities of the peptides at each mass/charge unit at a given point in the elution shown in the grey and black scale. Each one of the black dots seen in this figure b

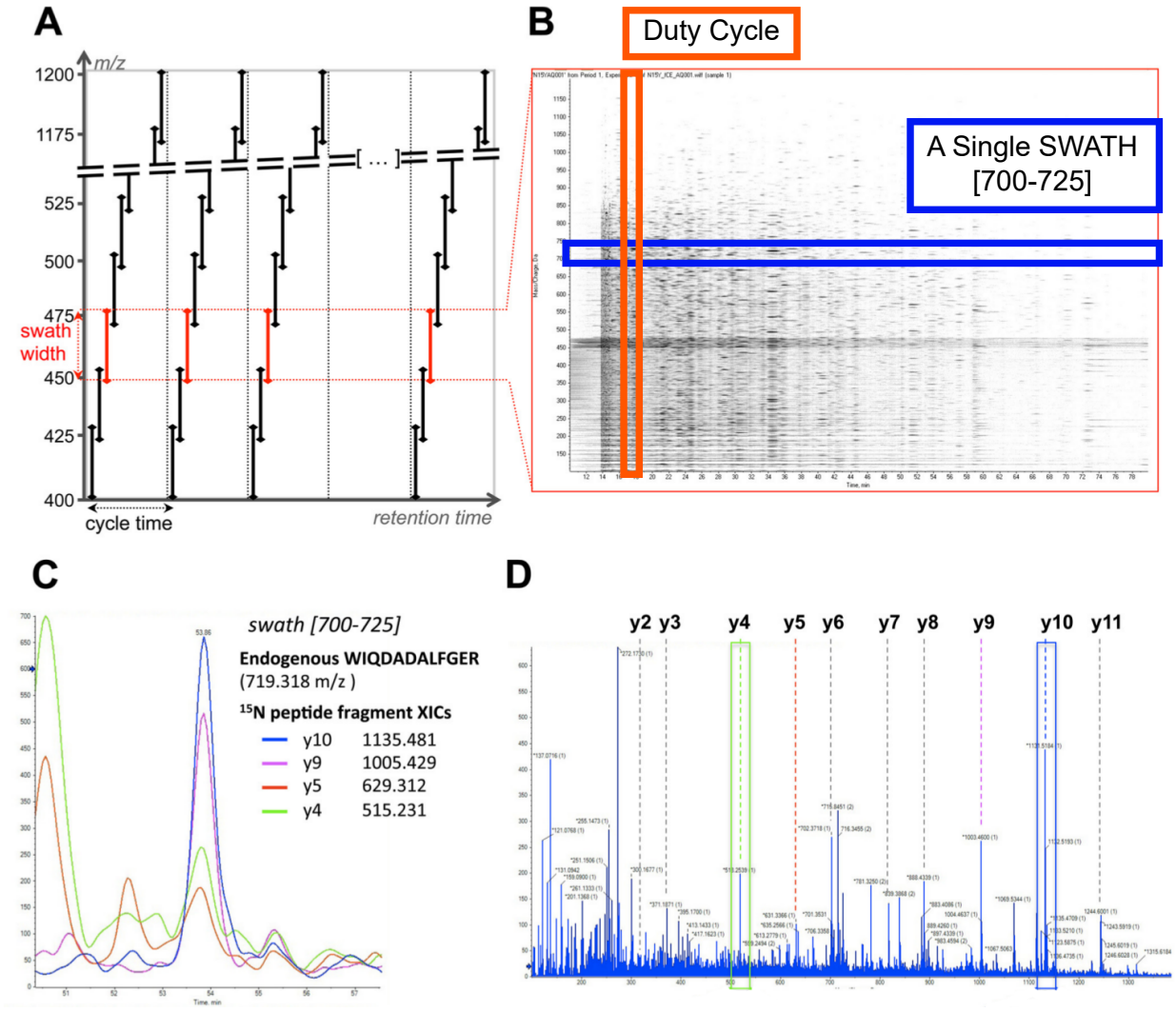


Figure 4.3: Number of RSUs that each vehicle has encountered

https://www.nature.com/nprot/journal/v10/n3/full/nprot.2015.015.html

4.1.1 Experimental Proteomics Protocol

Protocol Overview

In order to run shotgun and SWATH-MS (and indeed all "bottom-up proteomics" techniques), the total extracted proteins must be digested into shorter peptides. In this protocol, trypsin is to digest the protein into short amino acid chains, cleaved at lysine and arginine residues. We then clean the samples from any impurities such as lipids or salts that could affect the MS column. Take care that all liquid reagents, such as your water supply or acetone, are sufficiently pure for MS ("HPLC-grade"). Bi-distilled water is not likely to be sufficiently clean, as mass spectrometers are sensitive to contaminations, e.g., salts and detergents.

4.1.2 Reagents & Materials

- Water (HPLC-grade)
- Ammonium Bicarbonate (NH4HCO3)
- Acetonitrile (ACN) (HPLC-grade)
- Acetone (HPLC-grade)
- Methanol (HPLC-grade)
- Urea
- Potassium Hydroxide (KOH)
- Dithioethreitol (DTT)
- Indole-3-Acetic Acid (IAA)
- Trypsin (sequencing-grade)
- Formic Acid (FA)
- Indexed Retention Time peptides (iRT)
  - Important note if Spike-ins are being used:
    Spike in UPS1, Bovine Proteins, ETC. at 100 fmol concentration. You should pre-digest these and spike in the digested proteins. UPS1 is 6 µg - prepared this in volumes for 100 µg, but used 1/2 the trypsin concentration (1 µg) in twice the volume (1/4 the concentration, i.e. as if it was 25 µg, so still 4x concentration of whole samples)

Several buffers can be prepared in advanced and stored in sealed flasks for extended periods at room temperature (i.e. months). However, take careACN is volatile and will evaporate out of the mixtures, leading to decreasing ACN:H2O ratios if a flask is opened and used many times. This does not preclude using the same bottle for several days of experiments, nor preparing the bottles far in advance, but it is something to keep in mind. For stocks, it is recommended to prepare in advance:

- 0.1 M NH_4CO_3 in H_2O
- A high-concentration ACN:H_2O solution (8:2) + 0.1% FA
- A medium-concentration ACN:H_2O solution (5:5) + 0.1% FA
- A low-concentration ACN:H_2O solution (2:98) + 0.1% FA
- 0.1% FA solution in H_2O (1:999)

4.1.3 Equipment

- Centrifuge
- -20° Freezer

- Heated Shake Plate (up to 37°)
- Silica C_{18} Columns (e.g. MacroSpin Column from The Nest Group)
- 96 well plate (200 µL capacity)
- Vacuum Drying Centrifuge

Protocol

This protocol is approximately a three day process and is based on 50 µg of input protein. The lower and upper bounds of protein here will be limited by the capacity of the C18 columns used, and the amount of trypsin. This protocol is designed for 20300 µg of input protein. The exact reagent volumes used below are not sensitive, but the ratios are. The volumes should be kept relatively consistent across samples. The quantities here are planned for running 100 µg protein for each sample.

Day 1: Loading Controls and Acetone Precipitation

1. Add your batch correction loading control.
    - For the BXD study, a stock 20 pmol/µl of fetuin B, and 20 pmol/µl of alpha1-acid glycoprotein (AAG) was made.
    - This was 3.2 mg of fetuin B in 3.75 mL, and 1.74 mg of AAG in 3.75 mL, then mix it together for 10 pmol/µl of each in a 7.5 mL tube.
    - This was then frozen at -20° into aliquots of 520 µL/each; 5 µL will be added to each sample of 100 µg protein. The UPS1 standard may also be used.
2. Thaw samples on ice, then transfer 50 µg of protein to a new tube.
3. Add at least 6 volumes of cold HPLC-grade acetone (20°) to each sample to precipitate the protein. The next steps will be simpler if the acetone is ¡ 1.0 mL in volume.
4. Leave the samples in a regular 20° freezer and wait for a few hours (e.g. 424 hours; keep consistent within a study).

Day 2: Denaturing and Proteolysis

1. Centrifuge the samples at 20,000g for 10 minutes. Proteins should be well-fixed to the bottom of the tube. Remove acetone supernatant. (You can stop here and return to Day 2 much later, if you want.)
2. Prepare three fresh reagents:
    - 8 M urea + 0.1 M NH_4HCO_3 (add 8mL water, then 9.6g urea, then add 2 mL of stock 1 M NH_4CO_3; if necessary adjust to final volume of 20 mL)

- 360 mM DTT (DTT is 194 mg into 3.5 mL
- 800 mM IAA (800mM is 500 mg in 3.5 mL). IAA is light sensitive and should be kept and prepared in a low-light setting at all times and/or wrapped in aluminum foil.

3. Warm a shaking plate to 37°.
4. Using 85 µL of urea buffer (from step 5) for 100 µg of protein. Re-suspend samples by vortexing and sonicating. Then votex again.
5. Add 5 µL of 360 mM DTT buffer for the 100 µg of protein.
6. Vortex briefly, then incubate samples on the 37° shaking plate @ 400 rpm for 60 minutes. Take samples off and cool shaking plate to 25°.
7. Reduce light in the room as much as possible, then add 10 µL of 800 mM IAA for the 100 µg of protein.
8. Vortex briefly, then incubate samples on the 25° shaking plate for 45 minutes. Make sure that the samples are covered from light during this time (e.g. with aluminum foil).
9. Dilute samples with 0.1 M NH_4HCO_3 to a final urea concentration of 1.5 M (350 µL for every 100 µg of protein). Samples can be exposed to light.
10. Add sequencing-grade trypsin to the sample (add 4 µg trypsin for 100 µg protein; this is 8 µL at our standard trypsin concentration batch used). Need 17 tubes for a full 96 well plate (including accounting for error)
11. Warm shaking plate back up to 37°, then place samples here for 22 hours. Keep consistent. Put reasonably high speed (1000 rpm). Avoid going beyond 24 hours, as trypsin will start to self-digest which can create large peaks on mass spectrometry runs, obscuring the desired data.

Day 3: Column Cleaning and Final Sample Preparations

1. Activate the Silica C_{18} Columns with 180 µL of HPLC-grade methanol. (Note: double-check this with the protocol that comes with your C18 provider! We use NestGroup)
2. Centrifuge for 3 minutes at 1000g. Discard methanol.
3. Again, add 180 µL of HPLC-grade methanol.
4. Again, centrifuge for 3 minutes at 1000g. Discard methanol.
5. Wash with 180 µL of ACN:H2O 8:2 + 0.1% FA.
6. Centrifuge for 3 minutes at 1000g. Discard flow through.
7. Again, wash with 180 µL of ACN:H2O 8:2 + 0.1% FA.
8. Again, centrifuge for 3 minutes at 1000g. Discard flow through.
9. Prepare with 180 µL of ACN:H2O 2:98 + 0.1% FA.
10. Centrifuge for 3 minutes at 1000g. Discard flow through.
11. Take samples from the shaking plate (step 14) and centrifuge at 20,000g for 3 minutes. There should be no precipitate at the bottom. If there is, be careful to not pipette it in the next step. New step:

add 50 µL of 1% FA to make final buffer 0.1% FA.

12. Take 165 µL from the digested peptide samples and load them onto the C18 Columns.
13. Centrifuge at 1000g for 3 minutes.
14. Reload the outflow onto the column.
15. Again, centrifuge for 3 minutes at 1000g. Your peptides should be trapped in the column. Discard flow through.
16. Do step 27 two more times (to finish loading all 500 µL of sample that was digested)
17. Wash columns with 2% ACN
18. Centrifuge for 3 minutes at 1000g. Discard flow through.
19. Repeat the wash with ACN 2 more times.
20. Discard the old collection tube, add a new (and final) collection tube.
21. Add 150 µL of ACN:H2O 5:5 + 0.1% FA to elute the sample.
22. Centrifuge for 3 minutes at 1000g.
23. Add 150 µL of ACN:H2O 5:5 + 0.1% FA to elute the sample again
24. Again, centrifuge for 3 minutes at 1000g. Discard column.
25. Dry samples in a vacuum centrifuge. A warmed vacuum centrifuge to 37° will expedite this process.

If you do not plan on running your samples in the mass spectrometer immediately, stop at this step after the samples are dried, and freeze them at -80°.

Day 4: Mass Spectrometer Analysis

1. On the day you expect to start the mass spectrometry runs, re-suspend the dried samples with ACN:H2O 2:98 + 0.1% FA to a target concentration of around 2501000 ng/µl. (Your peptide quantity at the end will probably be 1/4 to 3/4 the input protein quantity, depending on experience and care.)
2. Vortex and sonicate to re-suspend the sample fully.
3. Centrifuge the samples at high speed (e.g. 20,000g) for 10 minutes to pull down any contaminants that may remain.
4. SPIKE IN YOUR DIGESTED & CLEANED PEPTIDE CONTROLS (e.g. UPS1) – these are the controls for different MS injections, not the controls for digestion differences (e.g. bovine).
5. Quantify your peptide concentrations.
6. Transfer some of each sample to the mass spectrometer sample tubes attempt to load approximately even concentrations across samples. The quantification data will be normalized afterwards, but it is better to start off with similar loadings.
7. If possible, run a few samples on a less sensitive mass spectrometer to ensure general protein quality and to check for any contaminations that would block the machine for the SWATH mass spectrometry

run. SEE STEP 46

8. Add 1 µL of the indexed retention time (iRT) peptides. This allows for correction across samples for small shifts in the mass-to-charge ratios measured.
9. Samples are now ready for injection in the mass spectrometer in either shotgun mode (for generating the library; additional fractionations are recommended) or SWATH mode (for quantifying the peptides; no fractionations are necessary). Inject as much protein as possible for the machine, in order to ensure that sufficient quantities of lowly expressed peptides can be measured. Low amounts (e.g. 100 ng) can be run, but fewer proteins will be quantified. Note that high amounts (e.g. ¿ 2 µg) may cause problems with certain machines.
10. For sample QC, take a handful of your samples (e.g. 10-15%) and go downstairs and load up the LTQ. You should first run a glufib, then you can run 2 of your samples, then run a beta-gal, etc. At the end, run 2 glufibs to clean out the system. Takes about 1 hour per sample. You can load as much as you want, but around 1 µg is usually good to go since that's what you'll run on the SWATH.

4.2 Results

4.3 Spectral Library Development

4.4 Quality Control

4.4

4.5 Biomarker Analysis

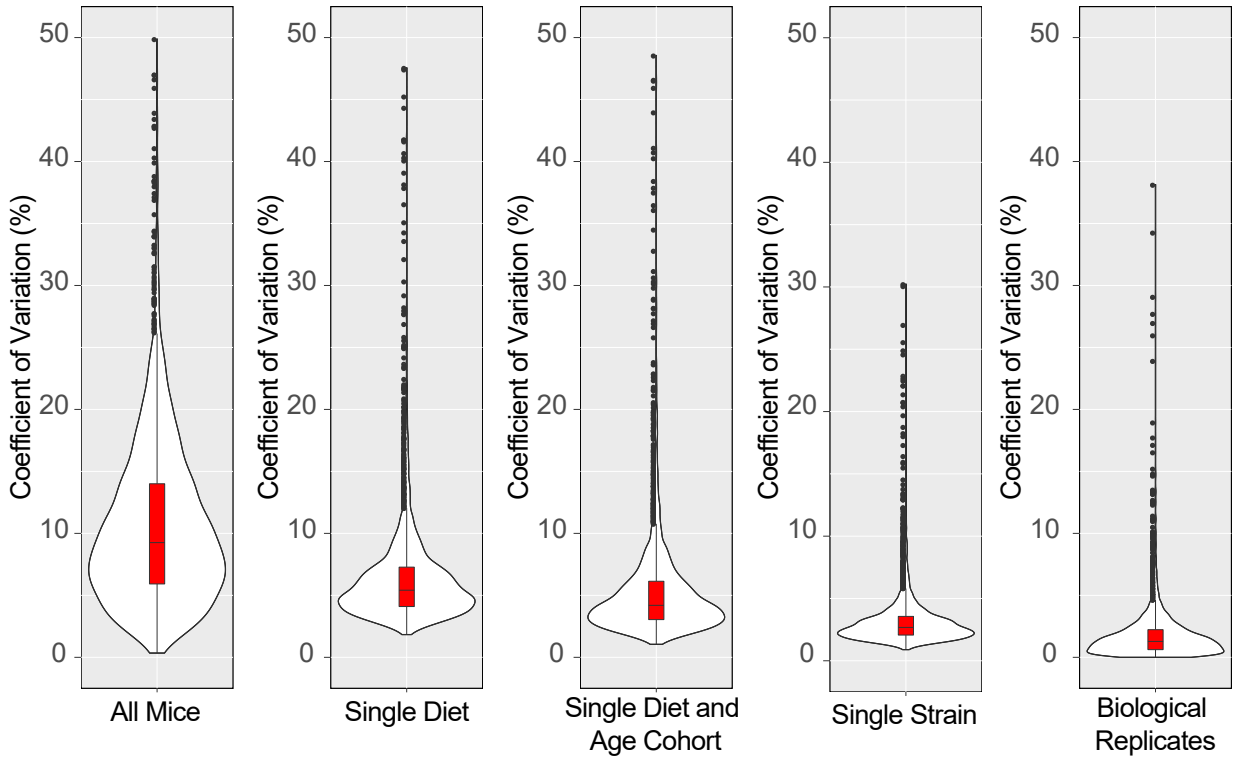
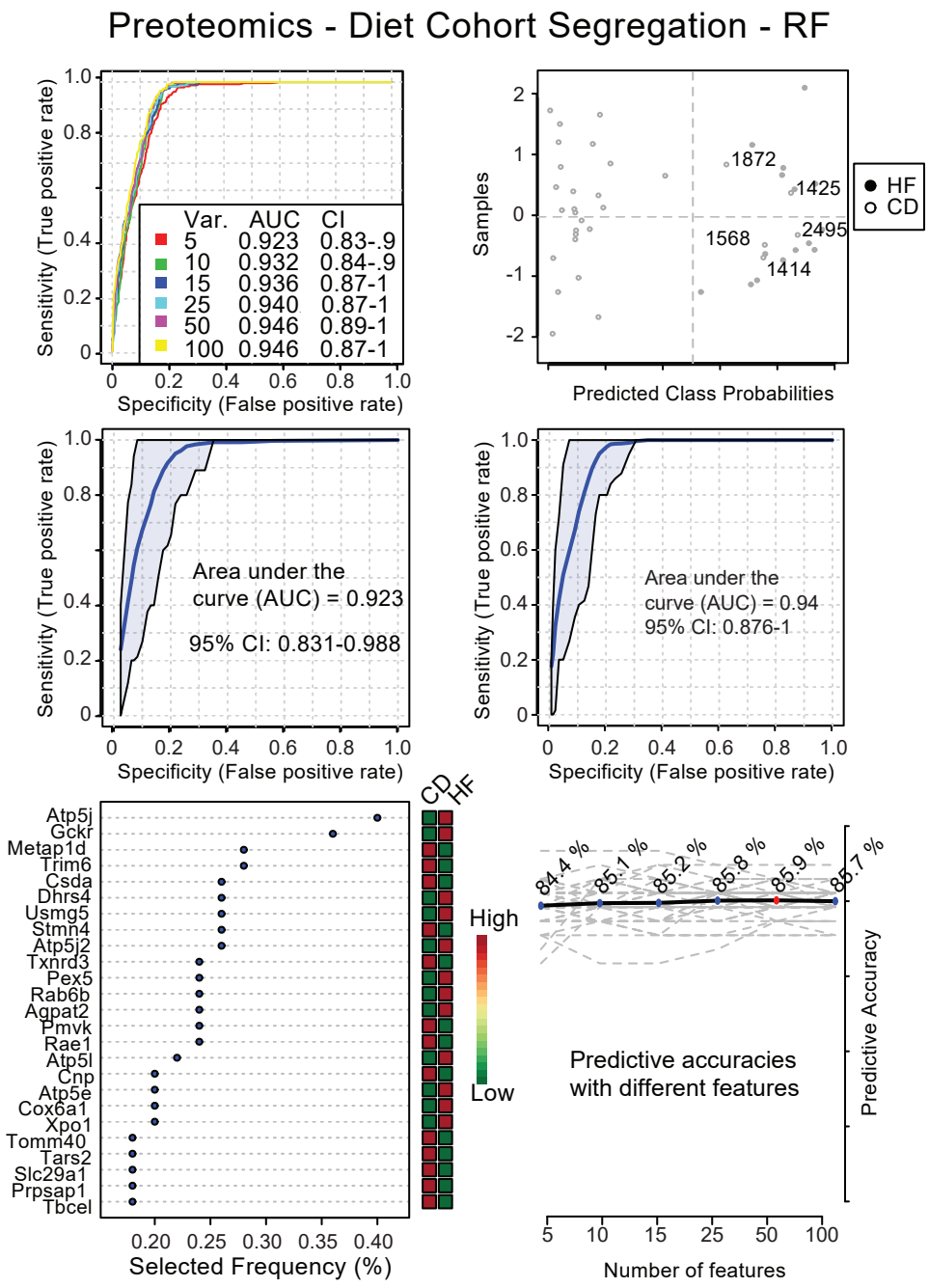
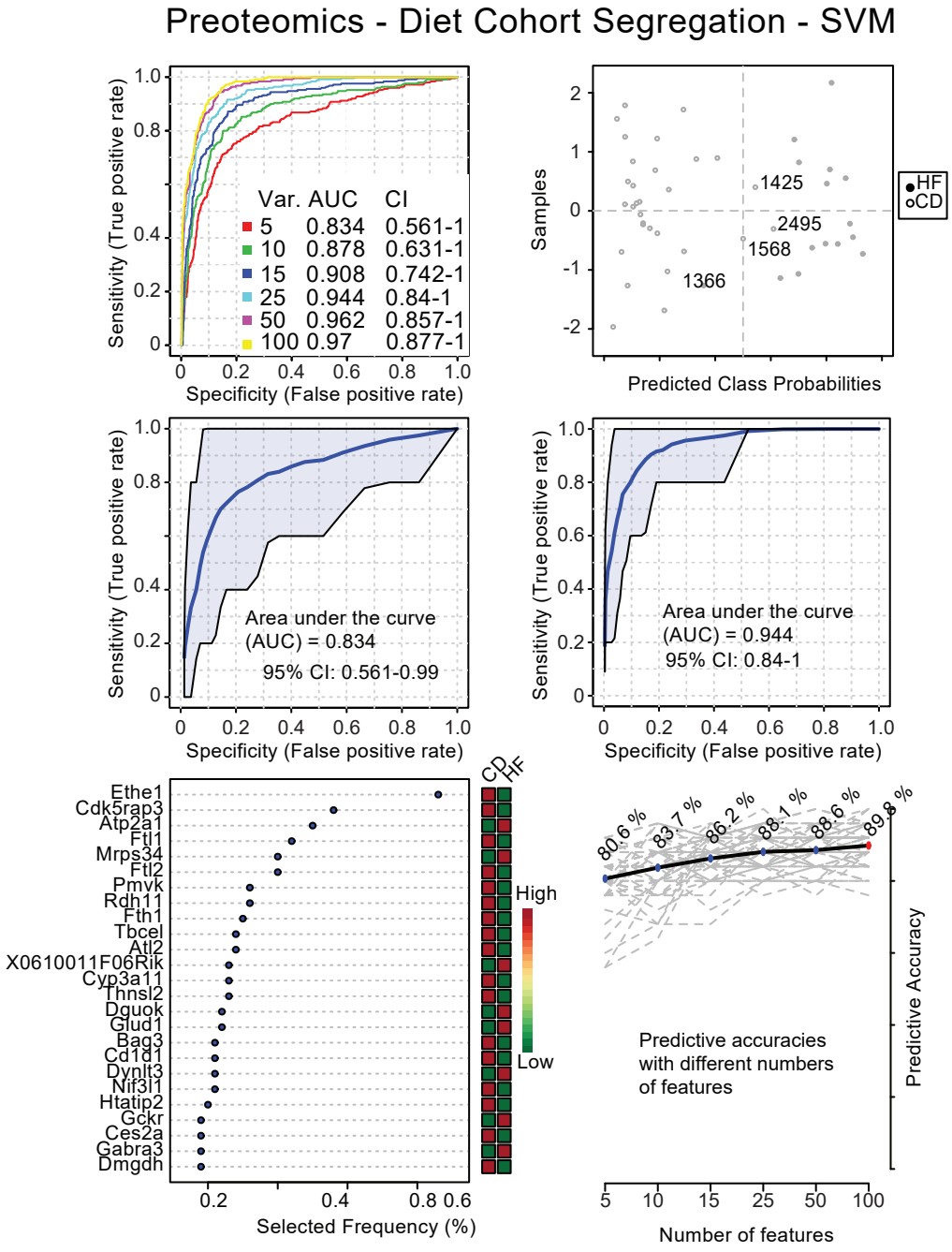


Figure 4.4: Coefficient of Variation of SWATH-MS Run between Mice measured on two days


Preoteomics - Diet Cohort Segregation - RF
Sensitivity (True positive rate)
Specificity (False positive rate)
Var. AUC CI
5 0.923 0.83-.9
10 0.932 0.84-.9
15 0.936 0.87-1
25 0.940 0.87-1
50 0.946 0.89-1
100 0.946 0.87-1
Samples
1872
1425
1568
2495
1414
HF
CD
Predicted Class Probabilities
Area under the curve (AUC) = 0.923
95% CI: 0.831-0.988
Area under the curve (AUC) = 0.94
95% CI: 0.876-1
Atp5j
Gckr
Metap1d
Trim6
Csda
Dhrs4
Usmg5
Stmn4
Atp5j2
Txnrd3
Pex5
Rab6b
Agpat2
Pmvk
Rae1
Atp5l
Cnp
Atp5e
Cox6a1
Xpo1
Tomm40
Tars2
Slc29a1
Prpsap1
Tbcel
Selected Frequency (%)
High
Low
84.4 %
85.1 %
85.2 %
85.8 %
85.9 %
85.7 %
Predictive accuracies with different features
Predictive Accuracy
Number of features


Preoteomics - Diet Cohort Segregation - SVM
Sensitivity (True positive rate)
Specificity (False positive rate)
Var. AUC CI
5 0.834 0.561-1
10 0.878 0.631-1
15 0.908 0.742-1
25 0.944 0.84-1
50 0.962 0.857-1
100 0.97 0.877-1
Samples
Predicted Class Probabilities
HF
CD
1425
2495
1568
1366
Area under the curve
(AUC) = 0.834
95% CI: 0.561-0.99
Area under the curve
(AUC) = 0.944
95% CI: 0.84-1
Ethe1
Cdk5rap3
Atp2a1
Ftl1
Mrps34
Ftl2
Pmvk
Rdh11
Fth1
Tbcel
Atl2
X0610011F06Rik
Cyp3a11
Thnsl2
Dguok
Glud1
Bag3
Cd1d1
Dynlt3
Nif3l1
Htatip2
Gckr
Ces2a
Gabra3
Dmgdh
High
Low
Selected Frequency (%)
80.6 %
83.7 %
86.2 %
88.1 %
88.6 %
89.8 %
Predictive accuracies
with different numbers
of features
Predictive Accuracy
Number of features

Chapter 5

Transcriptomics

5.1 Introduction to Microarray

DNA microarray are a technology that allows researchers to profile the expression and relative abundance of a large set of transcripts. A typical transcriptomics experiment on a micro array involves the immobilization of a library of coding and non-coding RNA sequences to the micro-array chip after they have been converted to cDNA. To determine the relative abundance of a transcript

5.1.1 Microarray Experimental Methods

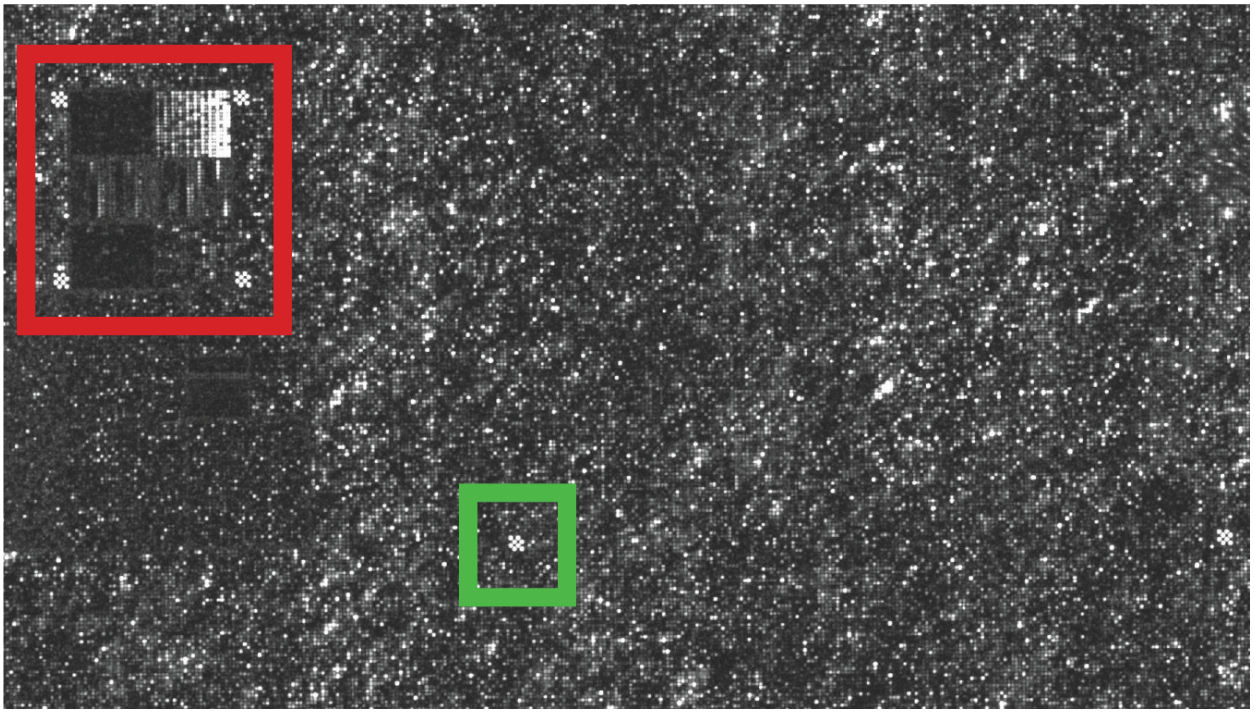
Liver tissues frozen in $N_(2l)$ were pulverized to a powder in a Motor and Pestle and shipped to the to the United states for precessing

5.1.2 Transcriptomics Data Processing

5.1.3 Data Extraction

The CDF (Chip Description File) contains information about the layout of the chip. There is one for each chip type. So for an experiment you normally have only one. A CDF file allows you to Link between probes and probesets Identify which probes are PM and which are MM Identify control probes.

The CEL file stores the results of the intensity calculations on the pixel values of the DAT file. This includes an intensity value, standard deviation of the intensity, the number of 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. The file stores the previously stated data for each feature on the probe array.

RMAExpress is a cross-platform program which provides methods for producing RMA expression values from Affymetrix CEL files. It simple program with a GUI interface aimed at Windows users. Implemented in C++. It has no dependencies on R. Generates RMA expression values. Requires text CEL and CDF files

5.1.4 Normalization

When running experiments that involve multiple high density oligonucleotide arrays, it is important to remove sources of variation between arrays of 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 situations.

Many traditional statical methodologies such as t-tests which will be performed on the microarray data afterwards are based on the assumption of normally distribution or at leart symmterically distributed data, with constant variance. if the assumptions are violated

http://dmrocke.ucdavis.edu/papers/ISMBTrans.pdf

Although we can use a LOWESS normalization, to estimate the error in the intensity and fit locally. However

is is preferred to have an error model in which our assumtion are made explicit.

5.1.5 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 of 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 of the abundance of the transcript in addition to noise that depends on the abundance and also noise that is indepedant or systematic noise. The parameter $b_k i$ summarizes abundance-dependent noise: which includes number of cells, hybridization efficiency, label efficiency. The parameter $a_k i$ Summareized 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 of 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 of 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 of 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 of variation as defined by $c = \dfrac{std}{mean}$

With this formulation, if we want to determine the relative abundance of 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 of both multiplicative and additive noise and in this case our linear expression for determine the true abundance of the transcript must be slightly altered. The additive noise term is also composed of a systematic error term a_i and sample specific but abundance-interdependent term $b_i\eta_{ki}$.

$$Y_k i = a_k i + b_k i \cdot x_k i \qquad a_{ki} = a_i + b_i \eta_k i \qquad b_k i = b_i \beta_k (1 + \epsilon_k i) \tag{5.1}$$

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 of the sources of noise in the microarray data from defined endogenous and exogenous sources in order for us to subtract it off, as indicated int he $Y_{ki} - a_i$ term.

5.1.6 Variance Stabilizing Normalization

Although the background noise has been subtracted a variance stabilization still needs to be performed on the data. This is because the coefficient of variation is not constant throughout the dataset. 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. 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 of data. Moreover, a log transformation provides good variance stabilization at high levels, but inflate the variance close to the detection threshold. Therefore, an arcsinh transformation is used instead as it behaves

like the log transformation asymptotically but is linear in the lowest intensity regions

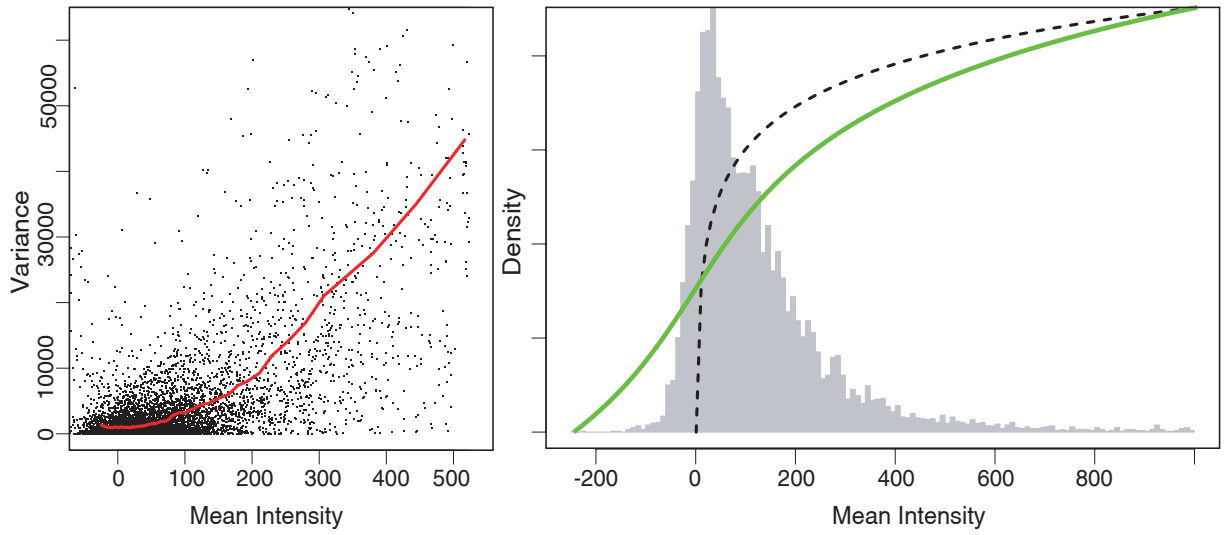


Figure 5.1: (Left) The variance and mean relationship found in experimentally produced microarrray data. The red line shows a plot of the v(u) described in at the end of section 2.3.6(right) The variance stabilization transformation performed on the data. The green line represents a log transformation, and the dotted line the arcsinh transformation which is hte preferred transformation method

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

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

5.1.7 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 of genes unchanged assumption in which the sample specific noise parameters can be more or less assumed to be the same across all transcripts.

$$b_i\beta_k = b\beta_k$$

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

5.1.8 Results

5.1.9 Quality Control

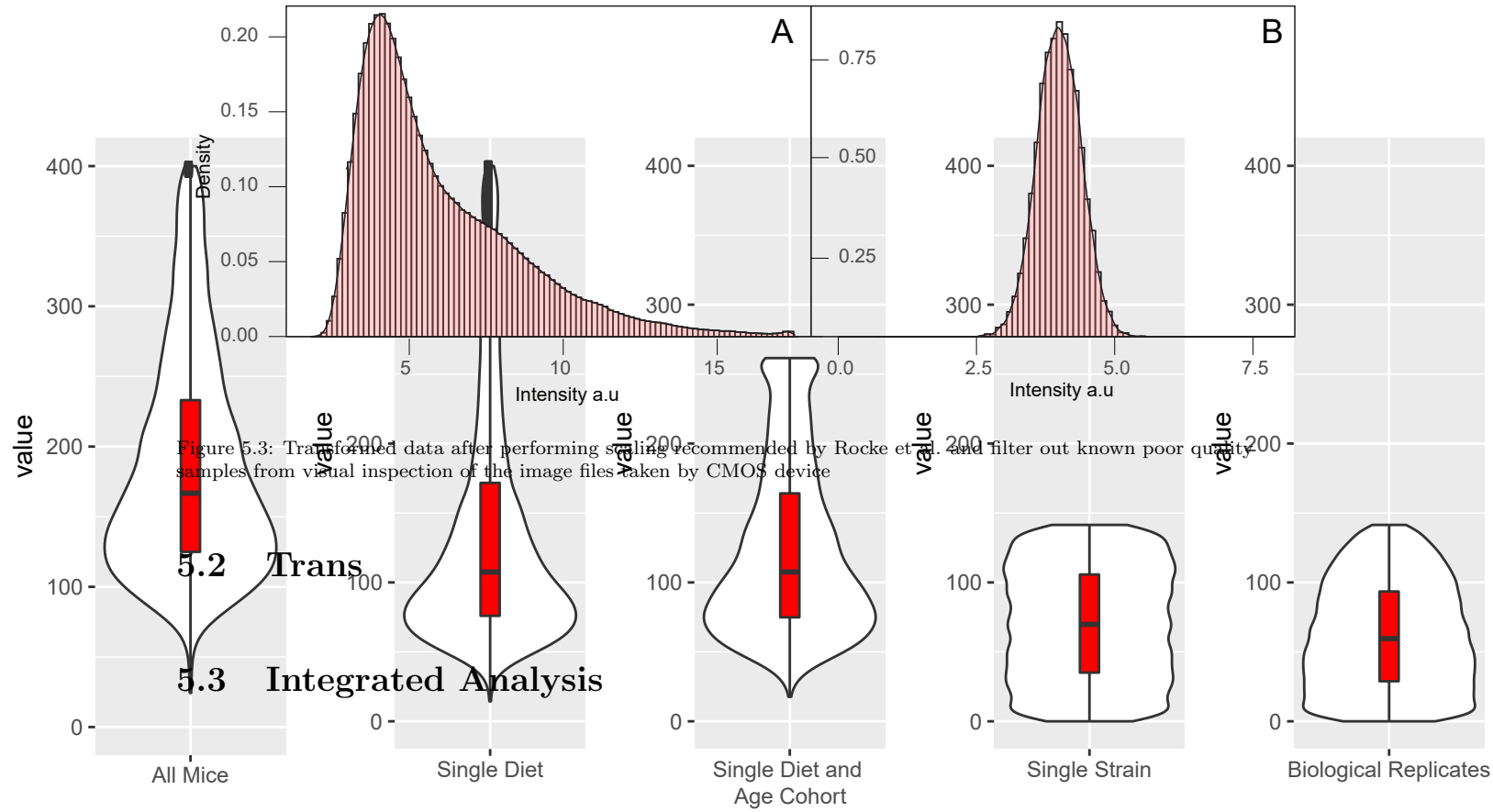


Figure 5.3: Transformed data after performing scaling recommended by Rocke et al. and filter out known poor quality samples from visual inspection of the image files taken by CMOS device

5.2 Trans

5.3 Integrated Analysis

RNA/Protein/Metabolite Network Analysis of Tryptophan Metabolism

All Fold Changes are the Log_2 ratio of HF/CD

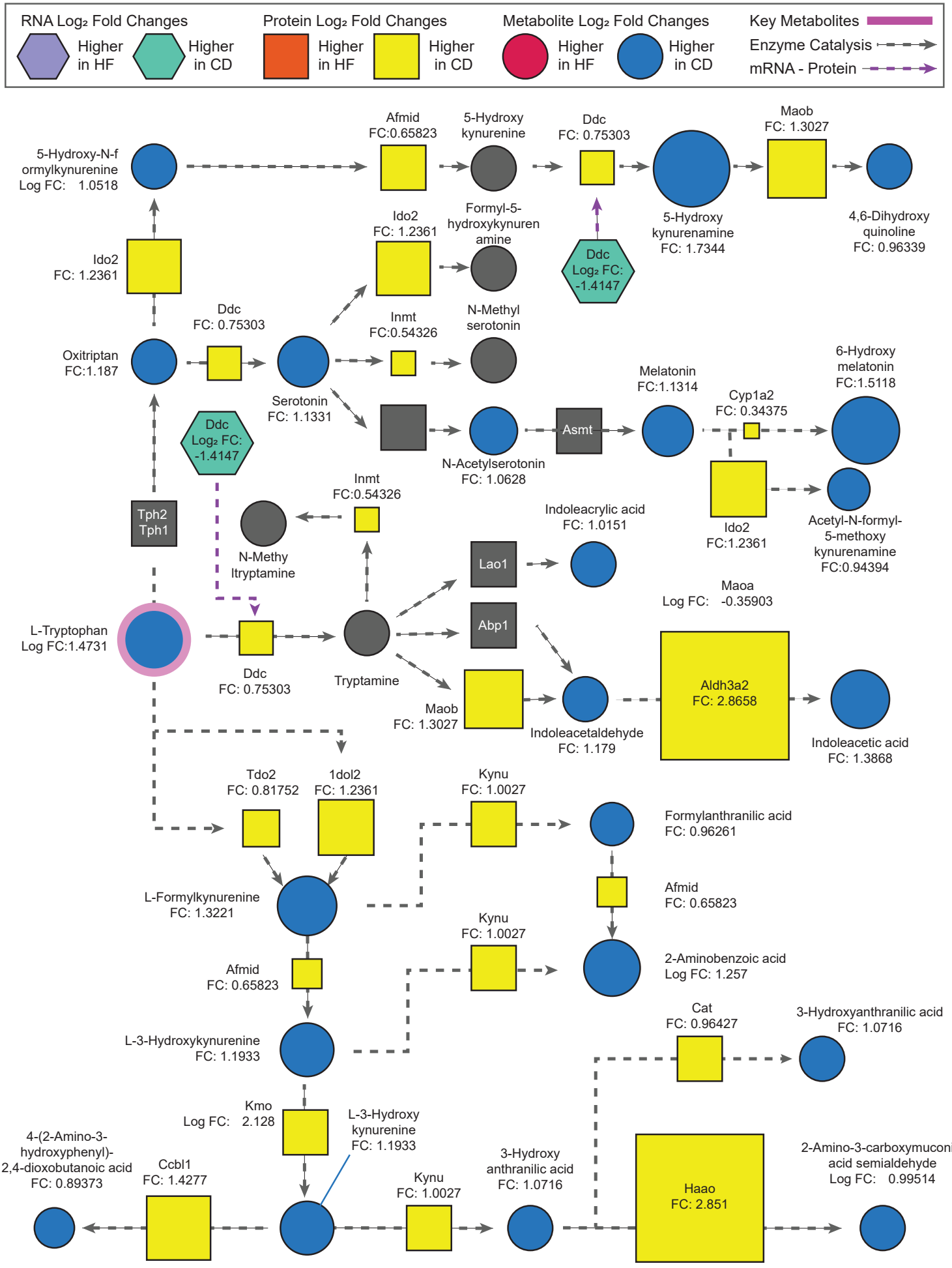


Figure 5.2: CV Analysis of the fist Micro array data

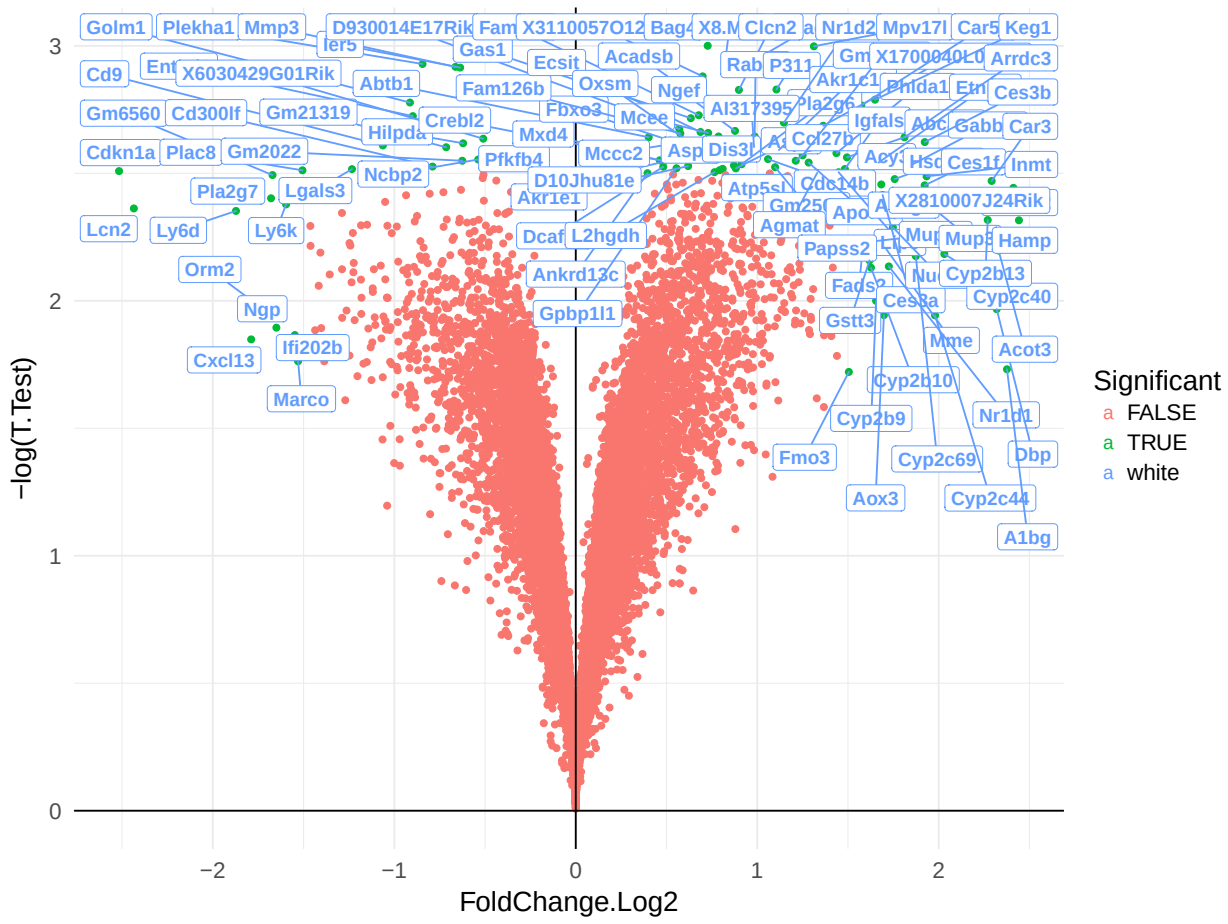
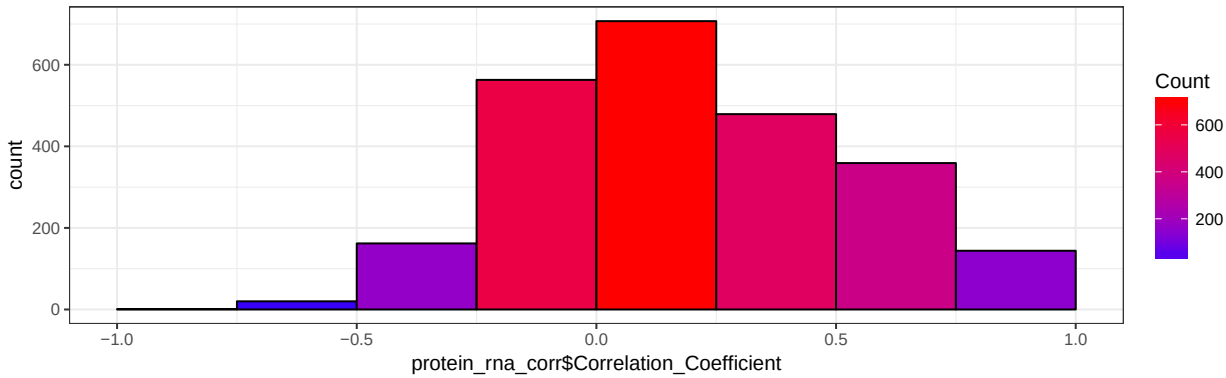


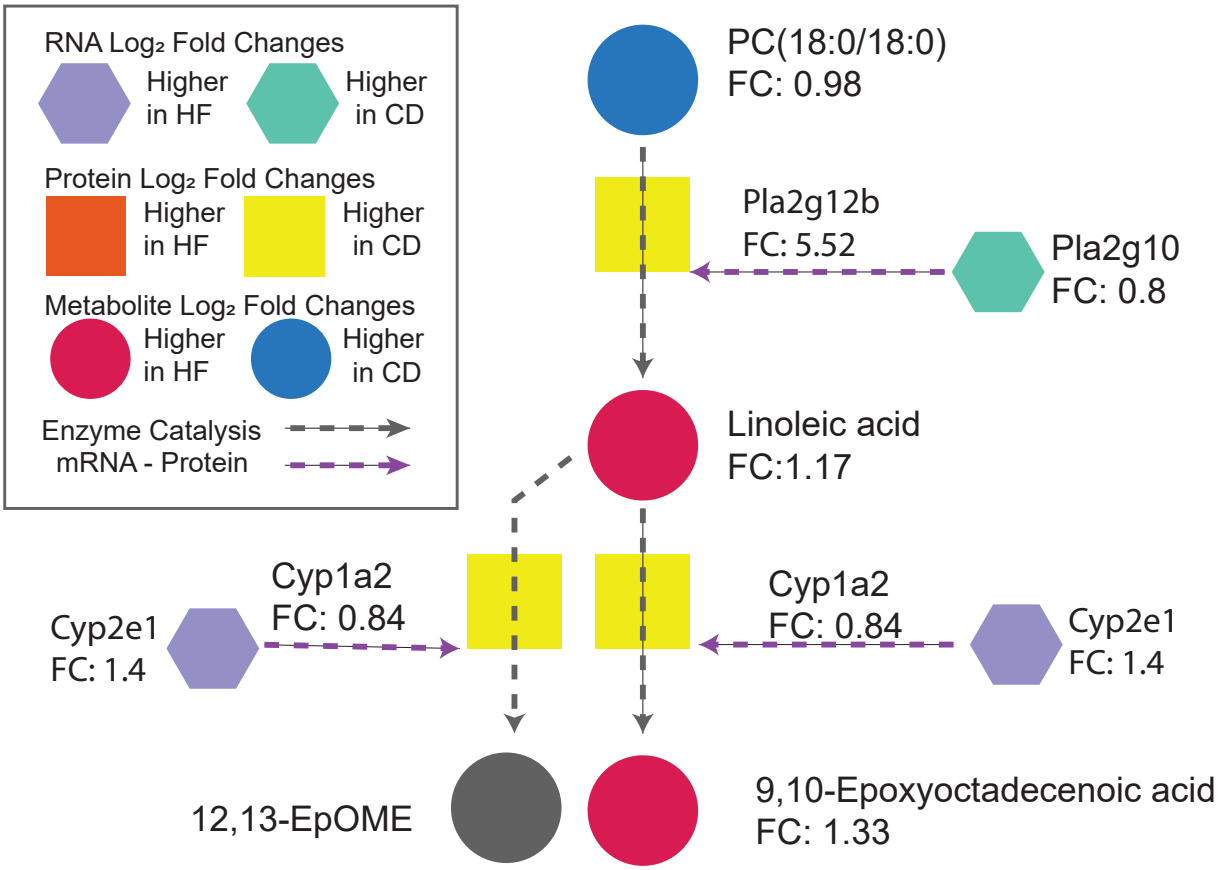
Figure 5.4: A subfigure

Figure 5.5: A subfigure



RNA/Protein/Metab Network Analysis of Linoleic Acid Metabolism

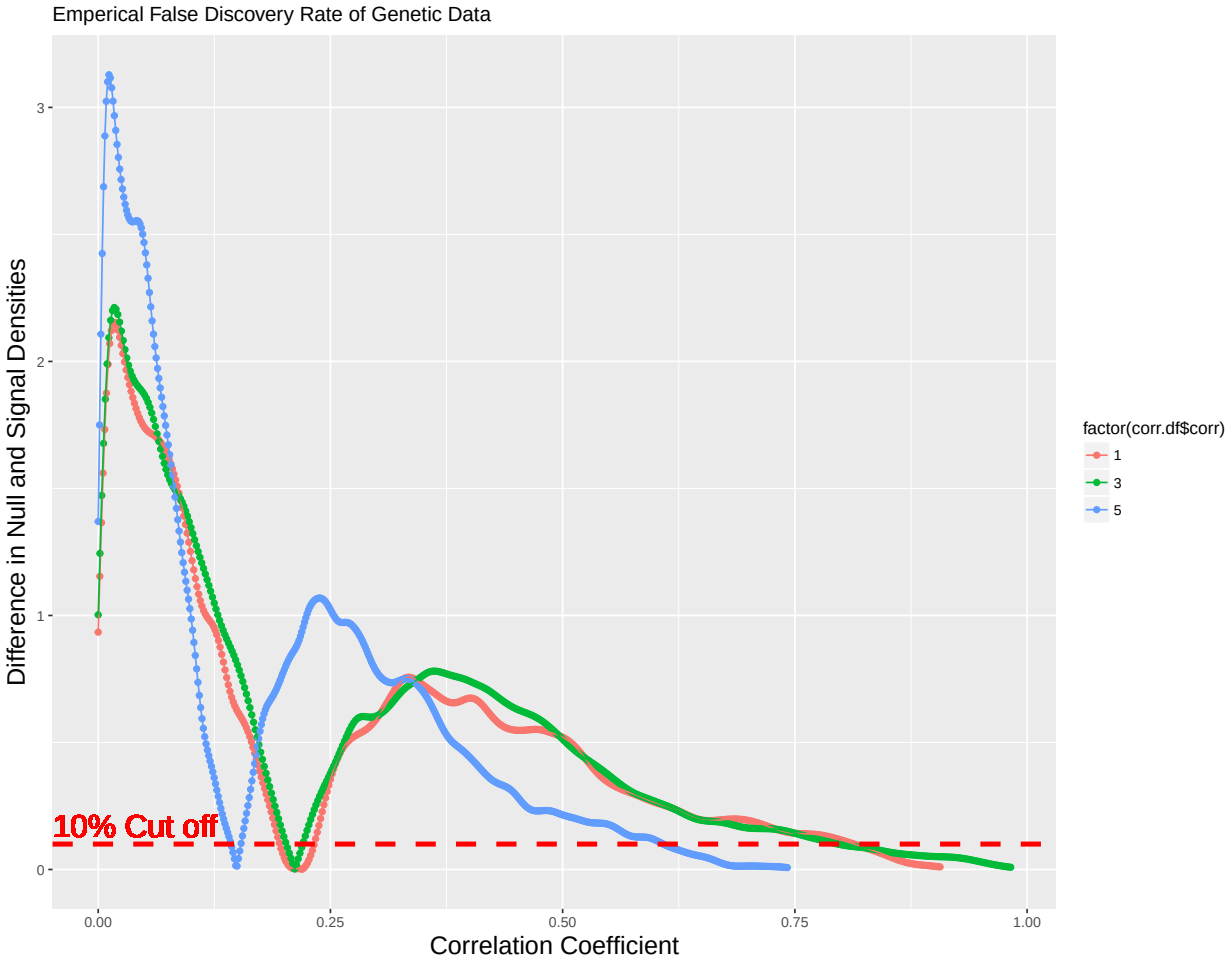
All Fold Changes are the Log_2 ratio of HF/CD



Chapter 6

Correlation Network

Determine which metabolites correlate best within their functional flux networks


Emperical False Discovery Rate of Genetic Data
Difference in Null and Signal Densities
Correlation Coefficient
10% Cut off
0
1
2
3
0.00
0.25
0.50
0.75
1.00
factor(corr.df$corr)
1
3
5

Chapter 7

Biomarker Analysis

7.1 Introduction to Multi-Omics Biomarkers

One of the goals of this project is to find biomarkers and biological networks that are indicators of aging and metabolic disease. Biomarkers refer to a broad category of physiological factors that provides an objective indication of the physiology of the animal or patient which can be measured externally. These factors have to be measured accurately and reproducibly and ideally with little invasiveness in order to be widely used as a proxy for physiology that could otherwise be difficult to measure. In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention. (Baker and Sprott, 1988) Examples of biomarkers include everything from pulse, body weight and blood test panels to complex genomic, transcriptomic, proteomic and metabolic signatures.

high-throughput technologies, such as genomic microarrays, proteomic and metabolomic mass spectrometry, have been used to generate large amount of data from single experiments that allow for global comparison of changes in molecular profiles that underlie particular cellular phenotypes. As a result, current biomarkers are mostly molecular makers, such as genes, proteins, metabolites, glycans, and other molecules, that can be used for disease diagnosis, prognosis, prediction of therapeutic responses, as well as therapeutic development (C. H. Johnson, Ivanisevic, and Siuzdak, 2016). The key issue however is determining the mechanistic relationship between any given measurable biomarker and biological endpoints and validating whether there is a robust correlation between the biomarker and the trait being indirectly measured. With enough tests one can find many biological correlations that do not actually correspond to an animals clinical state, or whose variations are undetectable and without a large effect on the physiology of the observed animal. Moreover,

even within inbred strain with only two founder strain, the variance of certain biological characteristics is so great as to render them all but useless as reliable predictors of the biological state of the animal.

7.1.1 Biomarkers for Aging and metabolic disease

In human graying hair and wrinkling in the skin are externally observable features that increase with age but cannot be called biomarkers of aging because they are not indicator of biological functionality and future function. A biomarker of aging must be a measurable biological feature of an organism that predicts functional capacity at some later age better than chronological age (Baker and Sprott, 1988). The identification of such biomarkers is an important goal for aging research, with thus far limited success. A major limiting factor in determining age related biomarkers is the requirement of a large set of individuals with feature specific measurement throughout their life spans. This is compounded by the logistical difficulty of maintain similar environment states of the individuals in order to remove environmentally determined factors from more intrinsic genetically driven factors. Moreover, the measurements in animal studies are destructive (the animal must be euthanized in order to allow for the extraction of the organs) or interventional (taking a sufficient amount of blood may markedly influence the physiology of an animal), and therefore it is not always possible to determine the normal lifespan of the animal from which the measurements were taken. This is why a large mouse cohort is used in this study, in order to provide us sufficient coverage of biologically identical mice at many ages at the three measured omics levels. 631 of the 2000 BXD mice could be used in the study in the end. The other mice having died of natural causes were not harvested on a specific time schedule and thus may have been in the cages for many hours before being found by the facility staff. The rigorous schedule of the animals sacrifice allows to preclude other environmental factor and age-related correlations with more power (Moeller et al., 2014).

7.2 ROC Curves

ROC(receiver operating characteristic curve) curves originally developed in world war 2 for radar signal detection is a graphical method method of evaluating the performance of binary classifier systems. ROC curve, which is defined as a plot of test sensitivity (the number of true positive decisions/the number of actually positive cases) as the y coordinate versus its specificity(the number of true negative decisions/the number of actually negative cases) or false positive rate (FPR) as the x coordinate. In the context of this project, ROC curves are used to determine which classifier and ensemble of factors are able to differentiate mice on different diets and in different age cohorts.

The result of each classification falls in to one of two obviously defined categories such as a mouse being

classified as one that has been eating a high fat or low fat diet. If only a single test is performed then the classifier has only one pair of sensitivity and specificity values. However, in many diagnostic situations, making a decision in a binary mode is both difficult and impractical. There is be a considerable variation in the diagnostic confidence levels between the radiologists who interpret the findings. As a result, a single pair of sensitivity and specificity values is insufficient to describe the full range of diagnostic performance of a test.

When given a large set of metabolites to classify mice in different diet categories, a method may use only metabolites that are significantly enriched in mice with a high fat diet in order to reduce the chance of incorrectly classifying a mouse as being a part of the chow diet cohort but then reduces its sensitivity to mice that how marginal differences in these metabolites and vise versa.

7.3 Cross-Validation

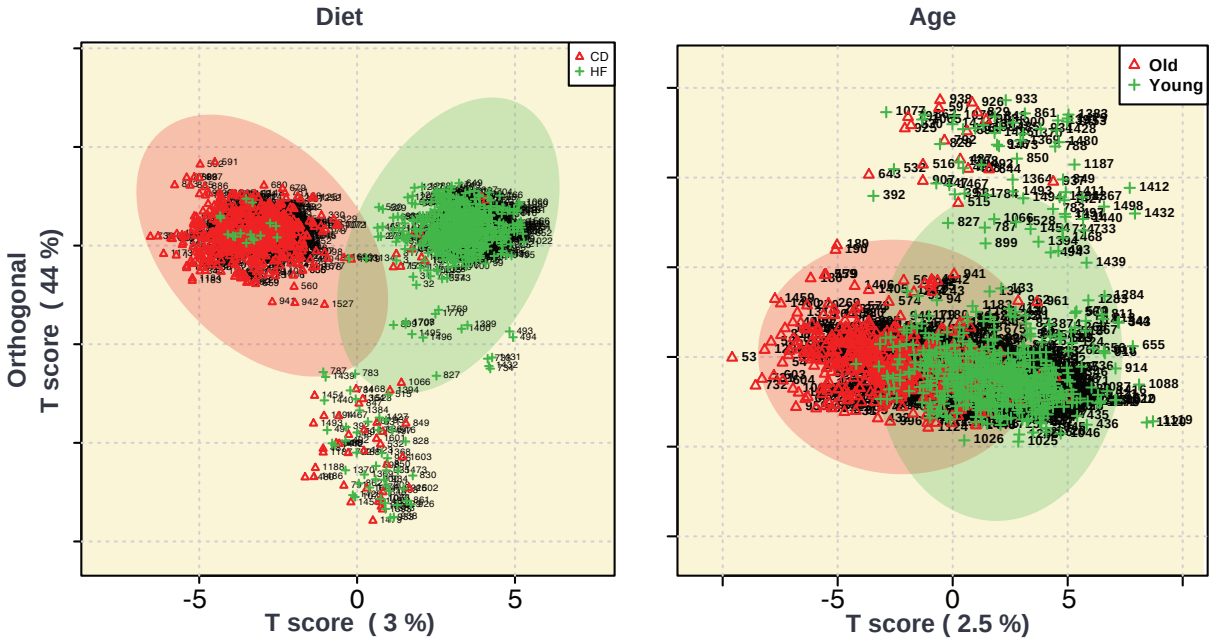
Cross-validation is a technique that can be applied to different types of models, including classification models. Leave-one-out cross-validation and k-fold cross-validation are popular choices. As an example, assume the sample size is n. For leave-one-out cross-validation, every time n1 samples are used as a training set to fit a classification model, and the remaining sample that is left out is used for testing. This process is repeated n times, and every sample serves as a test data once and only once. A model that is built on n1 samples is nearly as accurate as the model built on all n samples. The classification error rate is estimated as the proportion of misclassified test data points.

As leave-one-out cross-validation fits the classification model n times, it is computationally demanding. K-fold cross-validation simplifies this process. The whole dataset is divided into k equal size subsets (e.g., k=5 or k = 10). For each iteration, k1 subsets are combined and serve as a training set, and the one remaining subset serves as the test set. Again, every sample serves as a test data point once and only once. Both the leave-one-out and k-fold cross-validation error rate estimates are unbiased.

7.4 PLS-Da

7.5 Tree and Random Forest

One of the very nice properties of trees is their interpretation. Displaying information in terms of a tree structure is extremely useful because metabolites that partition the Diet and Age classes can be readily



read from the final figure. From the view of trees it also becomes clear that trees with depth ‘ may include interaction terms of degree ‘ between different predictor variables.

Regarding performance, classification trees often yield quite good prediction results with the Diet classifications because there are a few very strongly represented molecules can vegetables that are one found in the High Fat diet. The Age related classification is now as good as fats are often selected for partitioning the groups. Fats are not well extracted using the polar small molecule extraction. It may be worthwhile to point out that tree models/algorithms are doing variable selection automatically and preclude the introduction of bias from *a priori* information.

This method is not used in the final analysis because of the a few disadvantages of tree methods. The regression function estimate, or probability estimate in classification, is piecewise constant: this is not usually the form one thinks of an underlying true function. This also implies that the prediction accuracy for regression curve estimation (or probability estimation in classification) is often not among the best. Additionally, the greedy tree-type algorithm produces fairly unstable splits: in particular, if one of the first splits is wrong, everything below this split (in terms of the tree) will be wrong. Thus, despite the simplicity of interpreting the tree, it may be incorrect to do so with great certainty.

To overcome the instability of the tree methods a random forest type algorithm is employed.

Random Forests is an ensemble classifier which uses many decision tree models to predict the result. A different subset of training data is selected, with replacement to train each tree. We can get an idea of the mechanism from the name itself-”Random Forests”. A collection of trees is a forest, and the trees are

being trained on subsets which are being selected at random, hence random forests. This can be used for classification and regression problems. Class assignment is made by the number of votes from all the trees and for regression the average of the results is used.

7.6 Support Vector Machine

7.7 Neural Networks

Chapter 8

QTL and Genetic Analysis

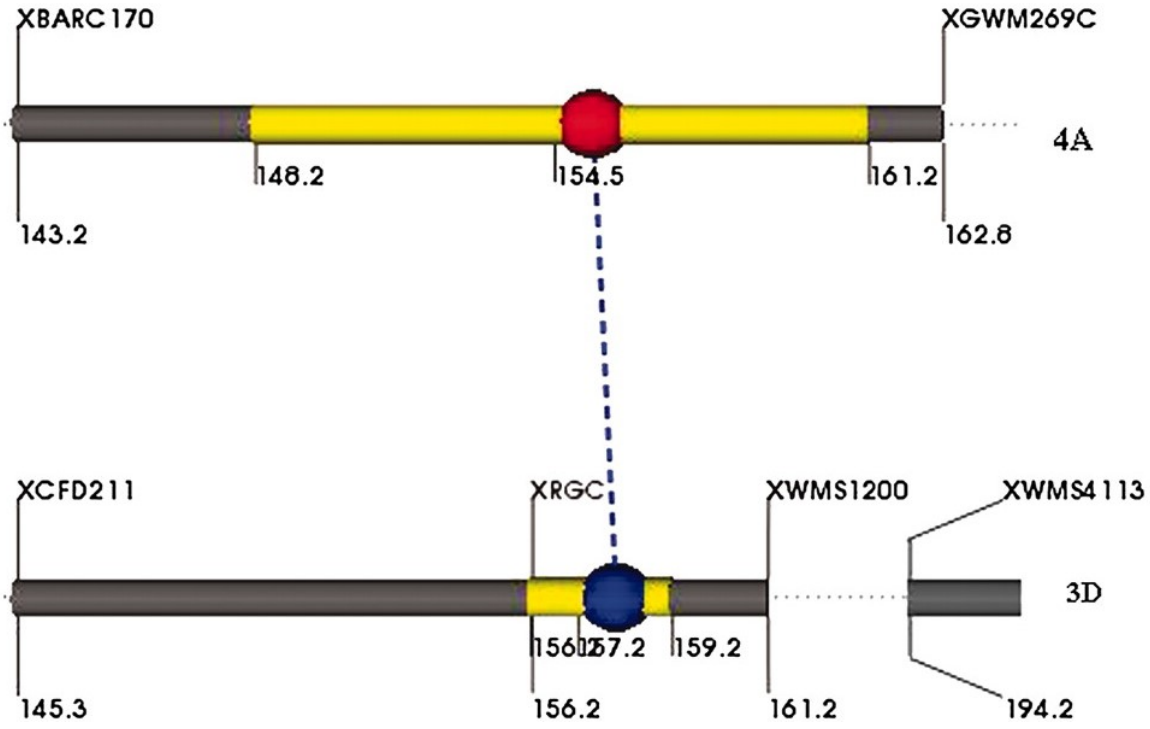
8.1 QTL Mapping

glutathione paper http://www.sciencedirect.com/science/article/pii/S1568163705000103?via

8.1.1 Bad QTL

8.1.2 Epitatsis

None of these tools can simultaneously investigate epistasis and QTLenvironment (QE) interactions. The use of QTL mapping software called QTLNetwork and 'CAPE' may allow us to dissect the genetic architecture of complex traits into single-locus effects (additive and/or dominance), epistatic effects (additive by additive, additive by dominance, dominance by additive and dominance by dominance) and their QE interaction effects, and also to visualize the analysis results by a series of graphs.


XBARC170
XGWM269C
4A
148.2
154.5
161.2
143.2
162.8
XCFD211
XRGC
XWMS1200
XWMS4113
3D
156.2
157.2
159.2
145.3
156.2
161.2
194.2

Chapter 9

Conclusions and Follow up Experiments

askldfjasklfjasdkl (Moeller et al., 2014)

References

Bibliography

Aksenov, Alexander A et al. (2017). "Global chemical analysis of biology by mass spectrometry". In: 1, p. 54. URL: http://dx.doi.org/10.1038/s41570-017-0054%2010.1038/s41570-017-0054%20https://www.nature.com/articles/s41570-017-0054#supplementary-information.

Baker, G T and R L Sprott (1988). "Biomarkers of aging." In: *Experimental gerontology* 23.4-5, pp. 223–39. ISSN: 0531-5565. URL: http://www.ncbi.nlm.nih.gov/pubmed/3058488.

Haynes, Christopher A. et al. (2009). "Sphingolipidomics: Methods for the comprehensive analysis of sphingolipids". In: *Journal of Chromatography B* 877.26, pp. 2696–2708. ISSN: 15700232. DOI: 10.1016/j.jchromb.2008.12.057. URL: http://linkinghub.elsevier.com/retrieve/pii/S1570023208009537.

Hu, Chunxiu et al. (2009). "Analytical strategies in lipidomics and applications in disease biomarker discovery". In: *Journal of Chromatography B* 877.26, pp. 2836–2846. ISSN: 15700232. DOI: 10.1016/j.jchromb.2009.01.038. URL: http://linkinghub.elsevier.com/retrieve/pii/S1570023209000737.

Johnson, Andrew D and Christopher J O'Donnell (2009). "An open access database of genome-wide association results." In: *BMC medical genetics* 10, p. 6. ISSN: 1471-2350. DOI: 10.1186/1471-2350-10-6. URL: http://www.ncbi.nlm.nih.gov/pubmed/19161620%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2639349.

Johnson, Caroline H., Julijana Ivanisevic, and Gary Siuzdak (2016). "Metabolomics: beyond biomarkers and towards mechanisms". In: *Nature Reviews Molecular Cell Biology* 17.7, pp. 451–459. ISSN: 1471-0072. DOI: 10.1038/nrm.2016.25. URL: http://www.nature.com/doifinder/10.1038/nrm.2016.25.

Moeller, Mark et al. (2014). "Inbred mouse strains reveal biomarkers that are pro-longevity, antilongevity or role switching". In: *Aging Cell* 13.4, pp. 729–738. ISSN: 14749718. DOI: 10.1111/acel.12226. URL: http://doi.wiley.com/10.1111/acel.12226.

Mushtaq, Mian Yahya et al. (2014). "Extraction for Metabolomics: Access to The Metabolome". In: *Phytochemical Analysis* 25.4, pp. 291–306. ISSN: 09580344. DOI: 10.1002/pca.2505. URL: http://doi.wiley.com/10.1002/pca.2505.

Venable, John D et al. (2004). "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra". In: *Nature Methods* 1.1, pp. 39–45. ISSN: 1548-7091. DOI: 10.1038/nmeth705. URL: http://www.nature.com/doifinder/10.1038/nmeth705.

Williams, Evan G. et al. (2016). "Systems proteomics of liver mitochondria function". In: *Science* 352.6291. URL: http://science.sciencemag.org/content/352/6291/aad0189.

Williams, Robert W. and Evan G. Williams (2017). "Resources for Systems Genetics". In: Humana Press, New York, NY, pp. 3–29. DOI: 10.1007/978-1-4939-6427-7{_}1. URL: http://link.springer.com/10.1007/978-1-4939-6427-7_1.

Wu, Yibo et al. (2014). "Multilayered Genetic and Omics Dissection of Mitochondrial Activity in a Mouse Reference Population". In: *Cell* 158.6, pp. 1415–1430. ISSN: 00928674. DOI: 10.1016/j.cell.2014.07.039. URL: http://www.sciencedirect.com/science/article/pii/S0092867414009891.

Chapter 10

Appendix

10.1 mQTL Results

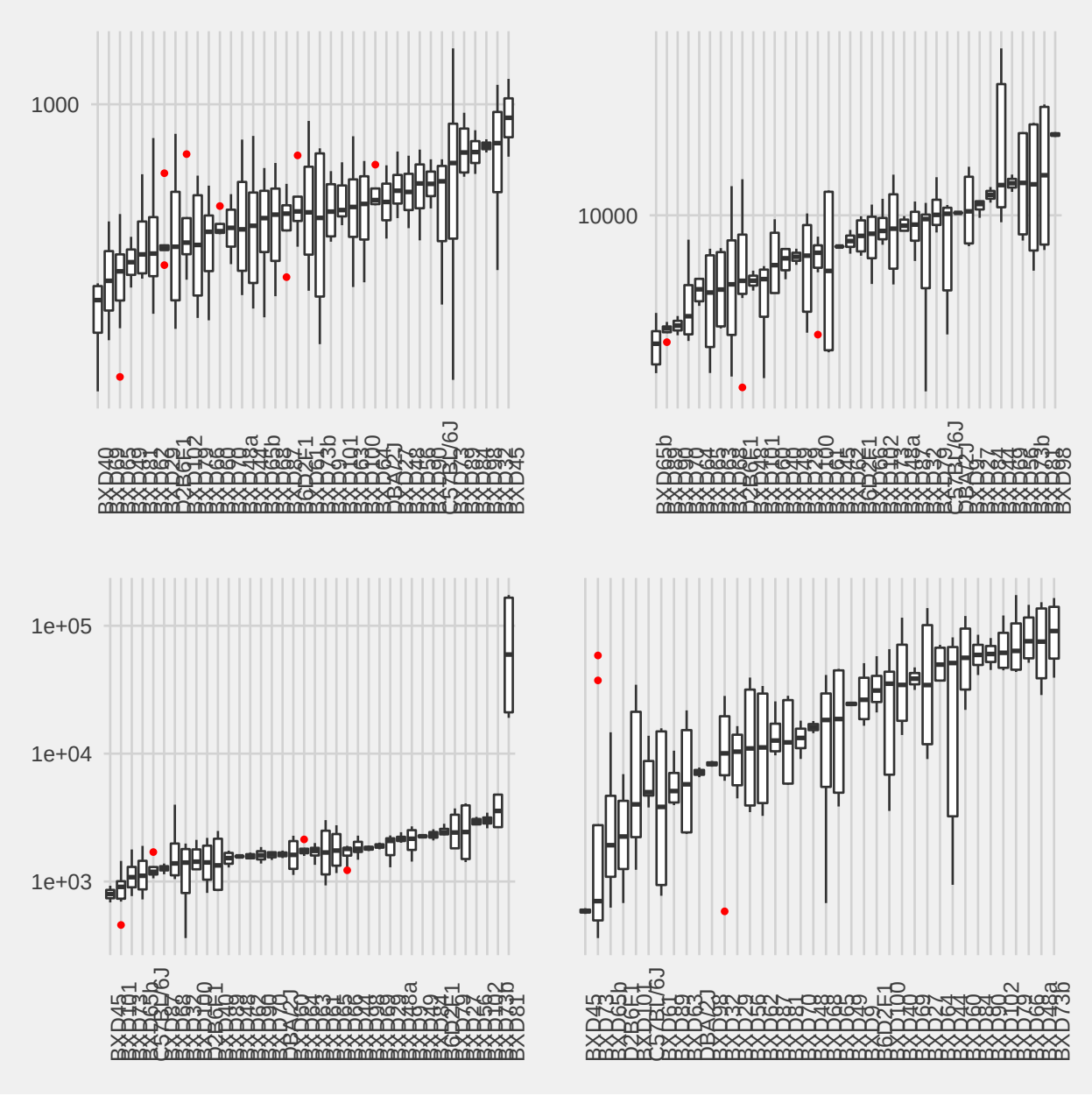
10.2 Metabolomics Protocol Optimization

10.3 protein

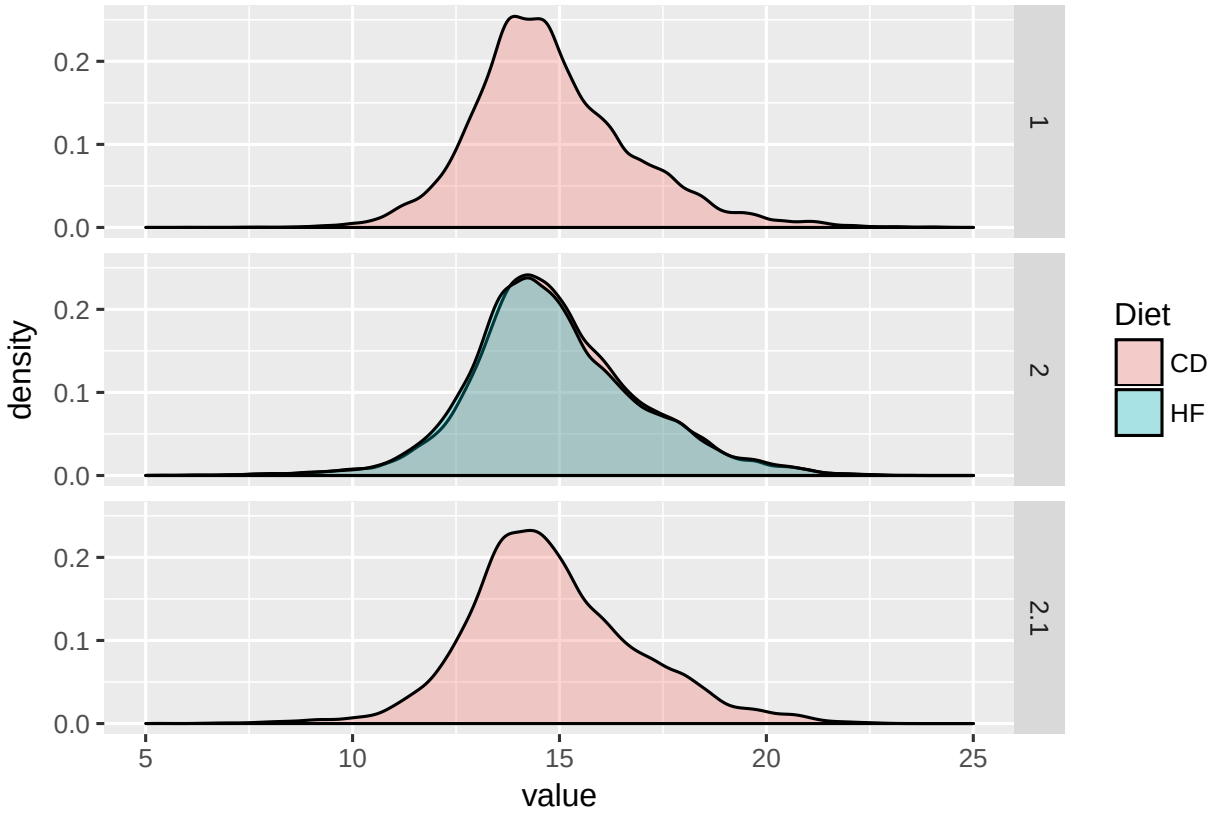
[h]

10.4 MSEA Table

[h]


1000
10000
1e+05
1e+04
1e+03


B6D2F1
BXD100
BXD101
BXD102
BXD11
BXD13
BXD24
BXD27
BXD29
BXD32
BXD34
BXD39
BXD40
BXD43
BXD44
BXD45
BXD48
BXD48a
BXD49
BXD50
BXD51
BXD53
BXD55
BXD56
BXD60
BXD61
BXD62
BXD63
BXD64
BXD65
BXD65b
BXD66
BXD67
BXD68
BXD69
BXD70
BXD73
BXD73b
BXD74
BXD75
BXD77
BXD79
BXD81
BXD83
BXD84
BXD86
BXD87
BXD88
BXD89
BXD9
BXD90
BXD91
BXD95
BXD98
BXD99
C57BL/6J
D2B6F1
DBA/2J
25
20
15
10
CD
HF
log(value, 2)
Diet
Meta
CD Filtered
CD Original
HF Filtered
HF Original


density
0.2
0.1
0.0
1
2
2.1
5
10
15
20
25
value
Diet
CD
HF

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

Appendix: Summary Statistics

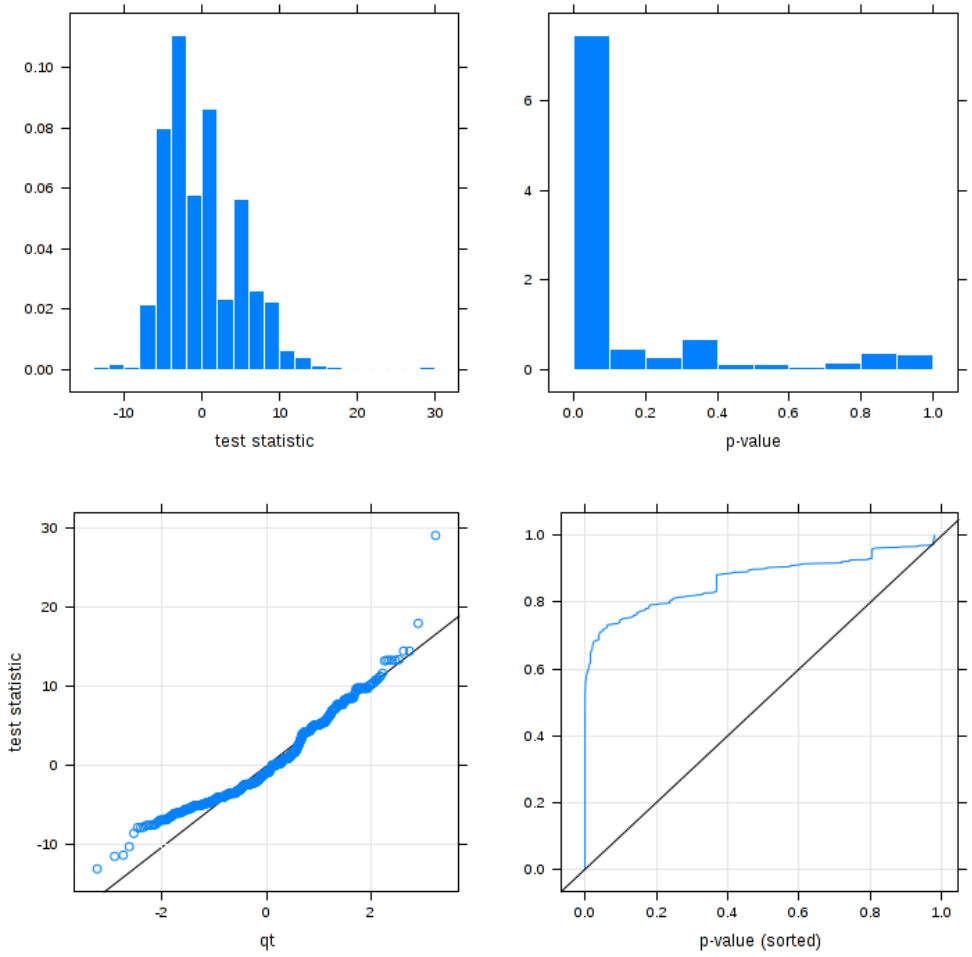


Figure 10.1: Summary Statistic for the metabolite data

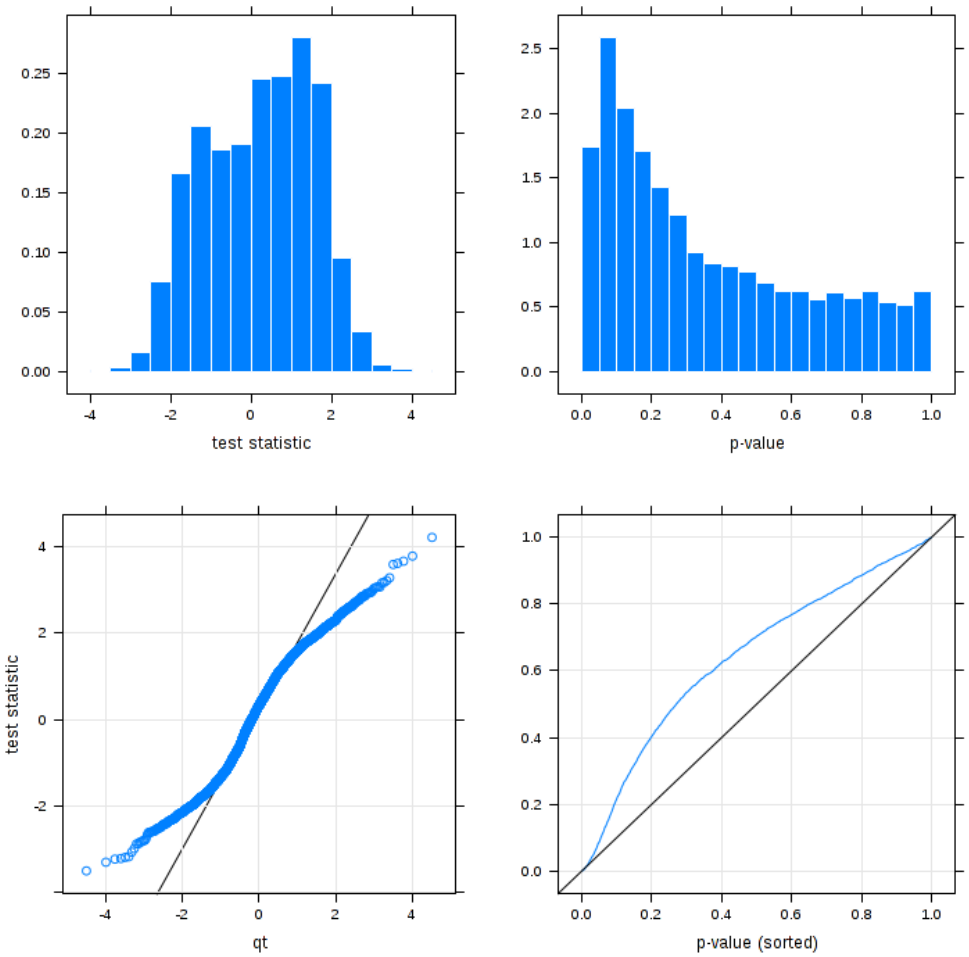


Figure 10.2: summary statics for the proteomics data

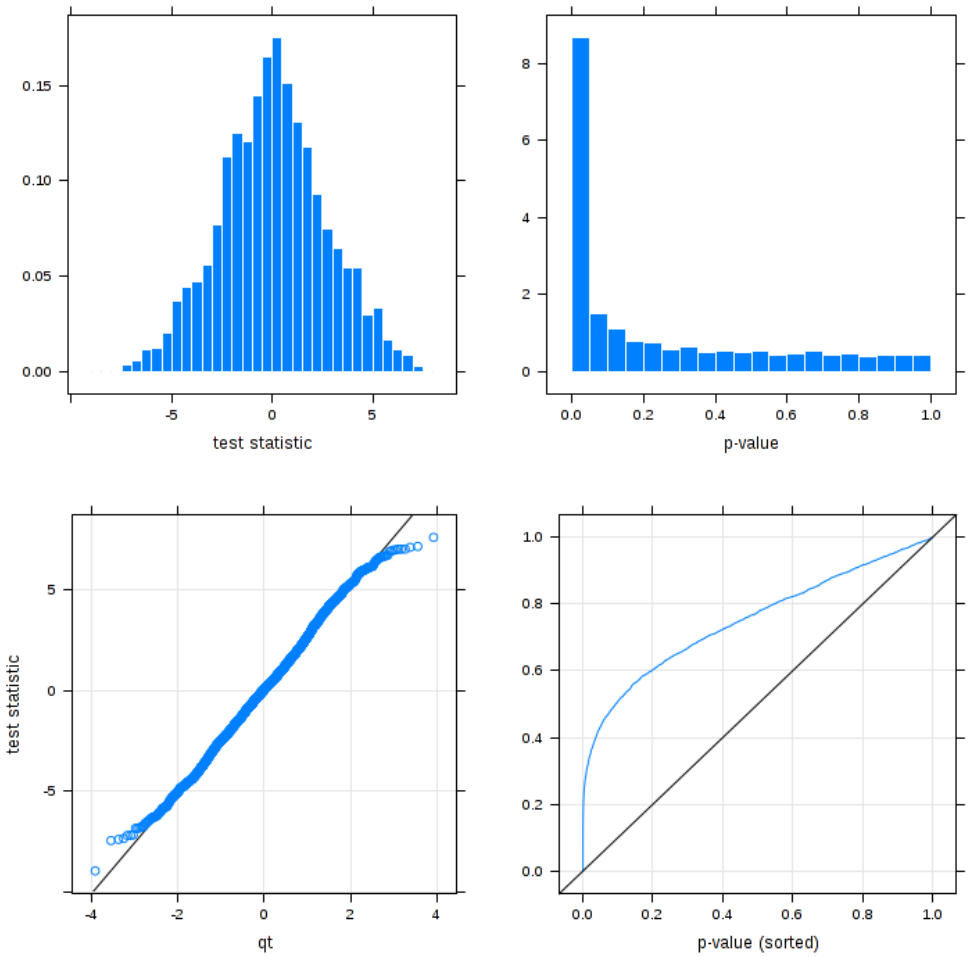


Figure 10.3: summary statistics for the transcriptomics data