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

Author: Moaraj Hasan
Direct Supervisor: Dr. Evan William Greal,
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

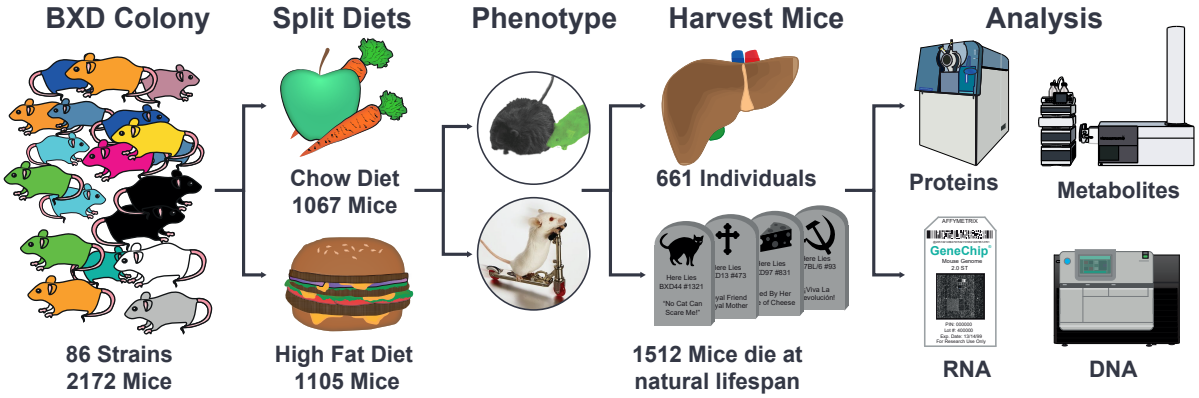
Introduction

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 Time-line

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 differnet quetsion P53 kn jownout effect too large, reverse genetics

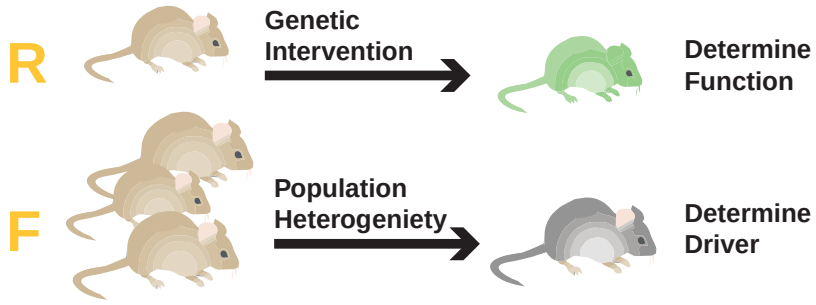


Figure 1.1: Forward and Reverse Genetic Paradigms

1.3 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

1.4 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*

Stocahstic manifestation of environmental factor , especially long term disease for chronic disease

1.5 Why Model Organism Population facilitate Complex Trait Analysis

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

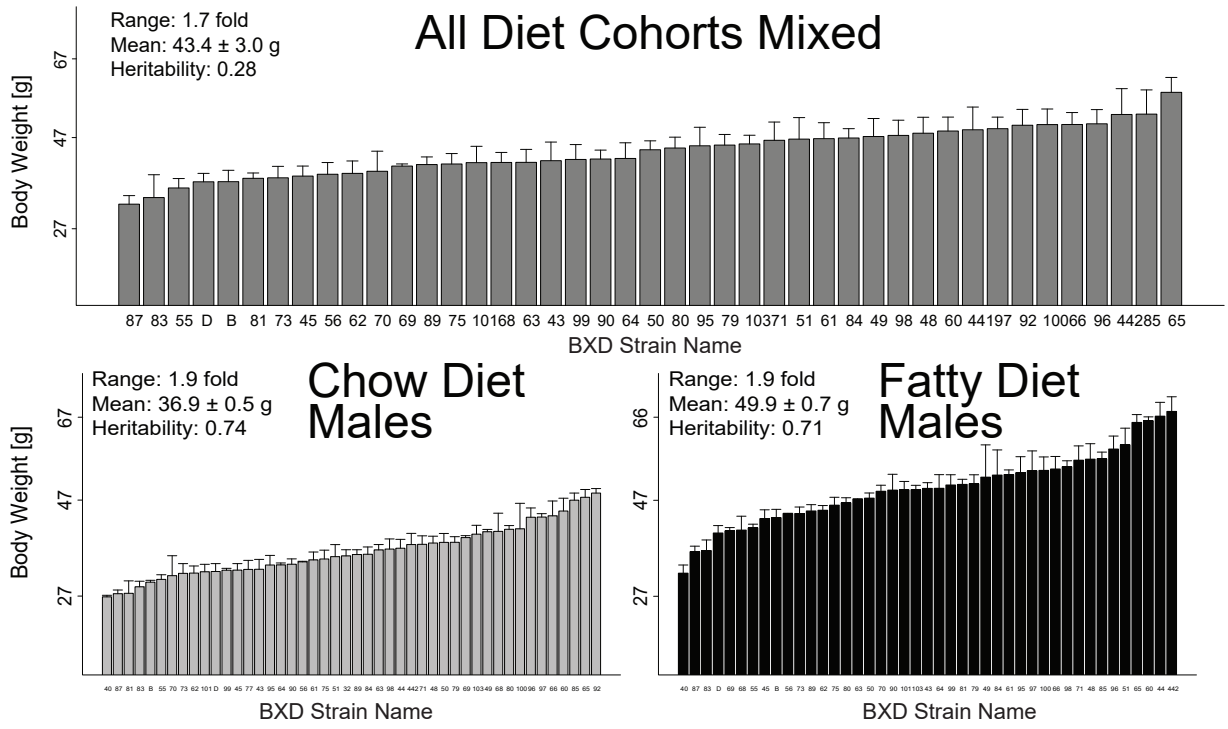


Figure 1.2: Heritability of Body Weight and fixed and mixed Diet population

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.

Heritablle trait is heritable, if its core physiology and behavioural trait distinction

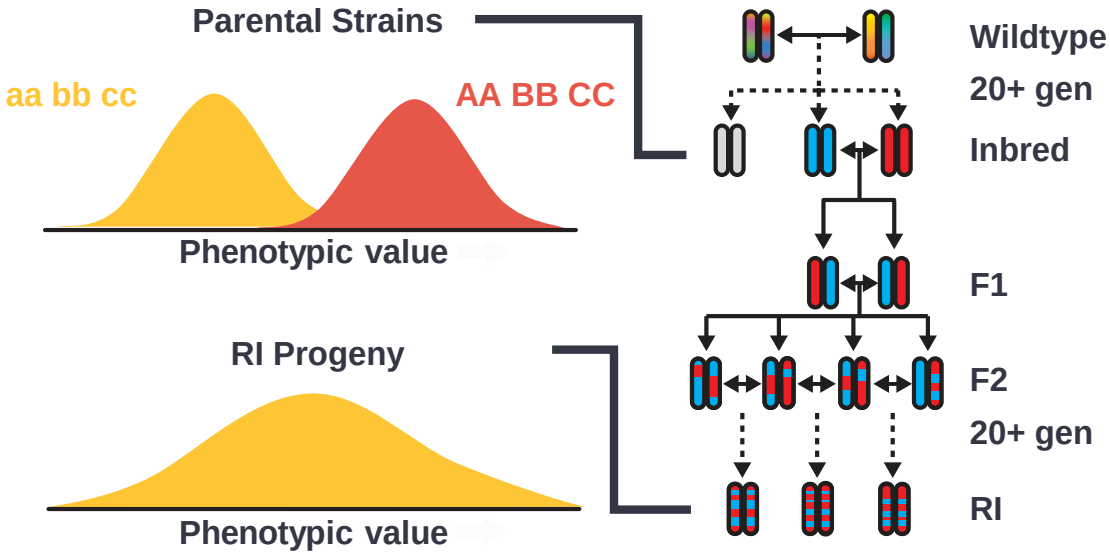
Applications of heritability estimation are broad and cross a range of disciplines, from evolutionary biology to agriculture to human medicine. In humans, estimation of heritability has been applied to diseases and behavioral phenotypes (e.g., IQ), and it has helped establish that a substantial proportion 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

1.6 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.



1.7 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 (CITE EVANs Papers). 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 (CITE EVANS TEXTBOOK), 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 (CITE EVANS TEXTBOOK). 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

1.8 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 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

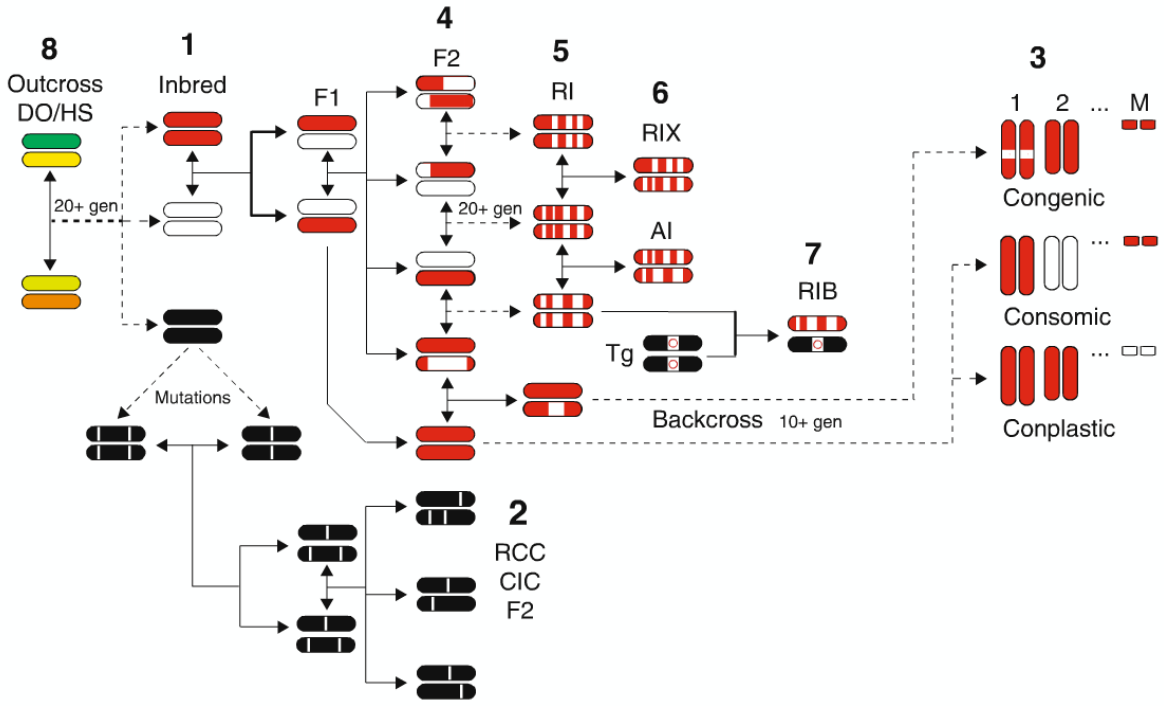


Figure 1.3: Produciton of RI stock and validation of QTL results in congenic lines

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.].

1.9 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.

1.10 Study Design

1.10.1 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.

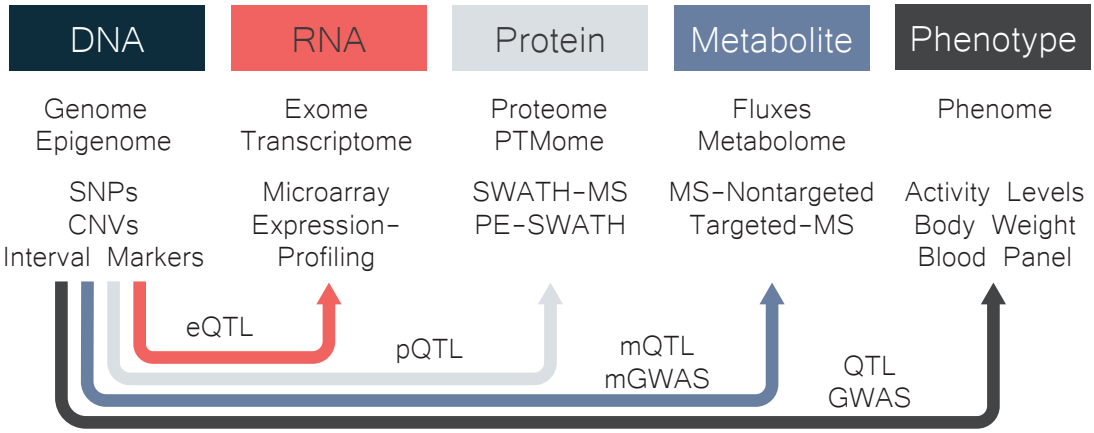
1.10.2 Study Design: Mouse Sex

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

1.11 Why Multiple Omics

Chapter 2

Metabolomics

2.1 Introduction to Non-Targeted Metabolomics

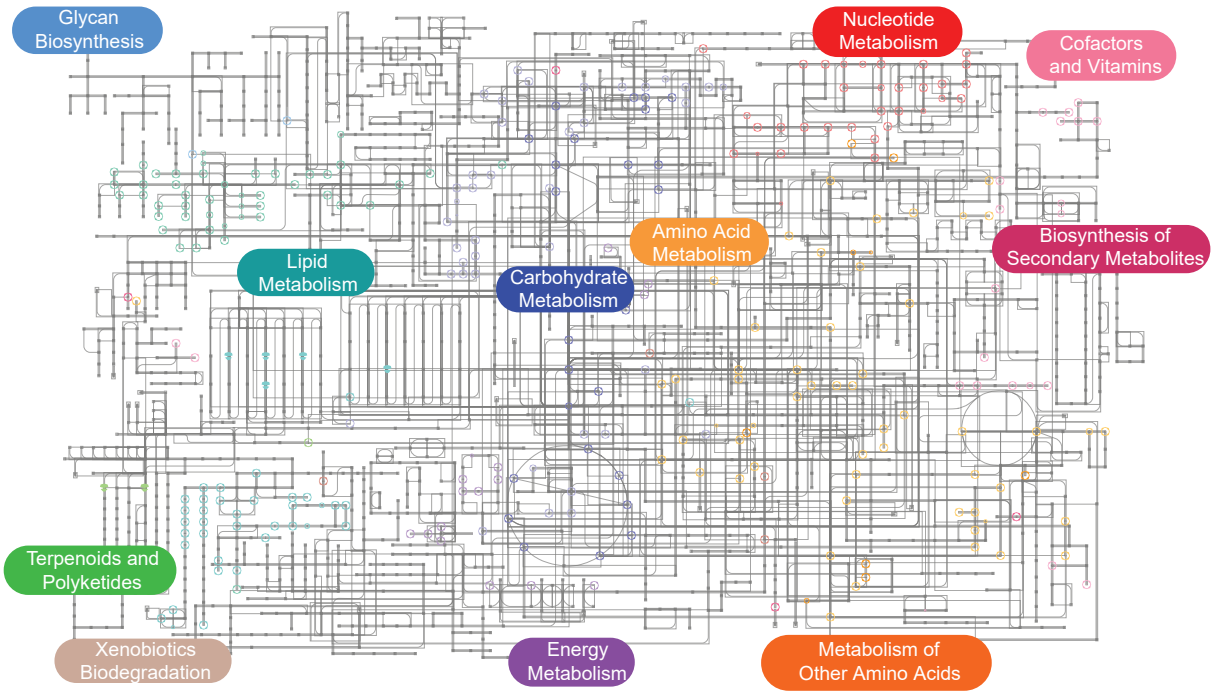
The metabolism is the sum of all the biochemical reaction that occur in an organism. Metabolites are shuttled through multiple enzymatically catalyzed reactions in the fundamental processes of energy production, growth and integrates multiple levels of information about the environment and cellular state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

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. Non-targeted Metobolomics is a mass spectrometer used to determine the steady-state concentration of major metabolic compounds in a sample. The advantage of metabolomics is its simplicity and speed in comparison to other omics techniques. Small molecules can be extracted from biological samples using simple and inexpensive procedures that require little processing before analyzing with a mass spectrometer.

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.[citation]

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.

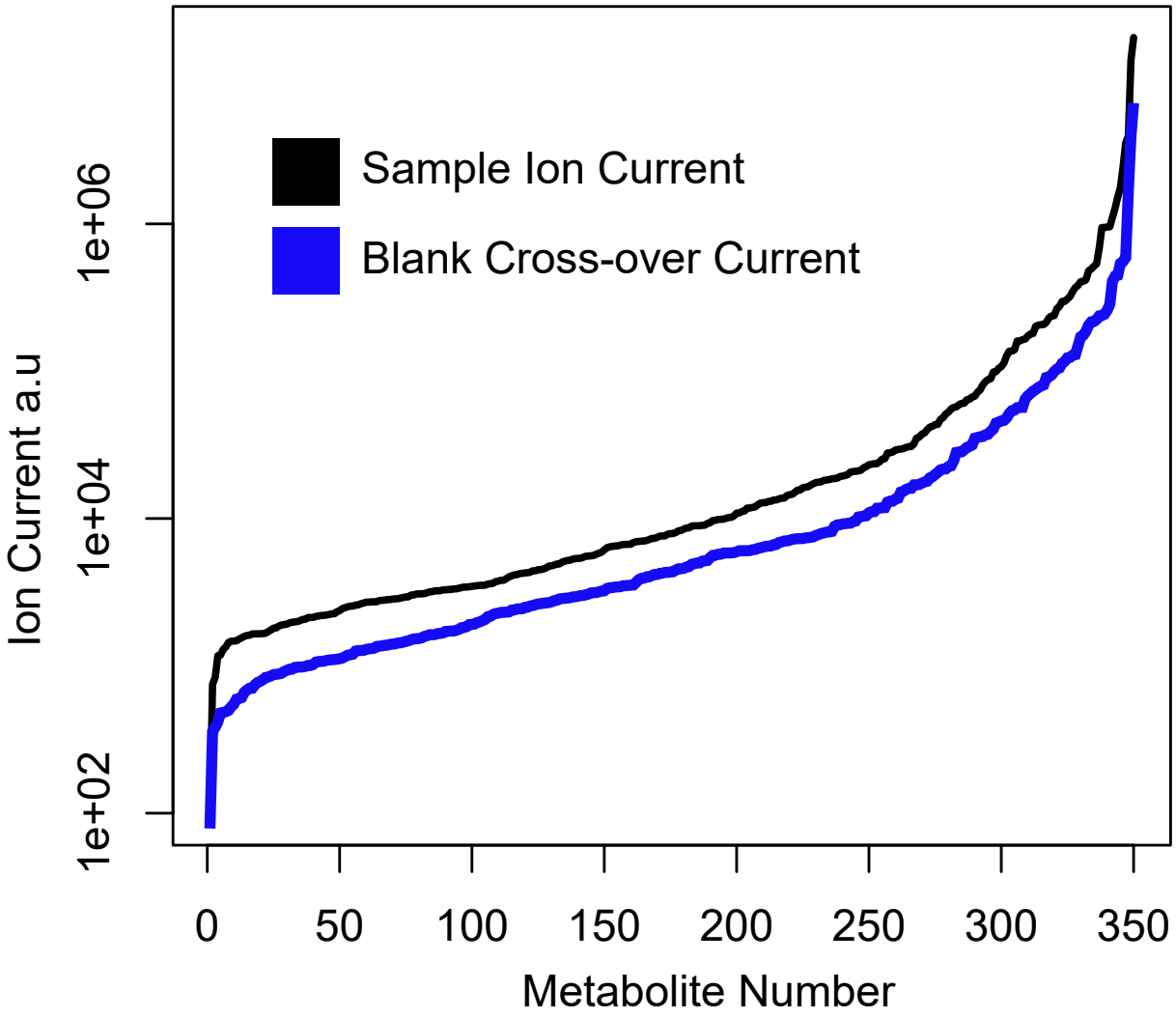
Metabolomics Optimization

2.2 Raw Data Processing

2.3 Data Conditioning

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.

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


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

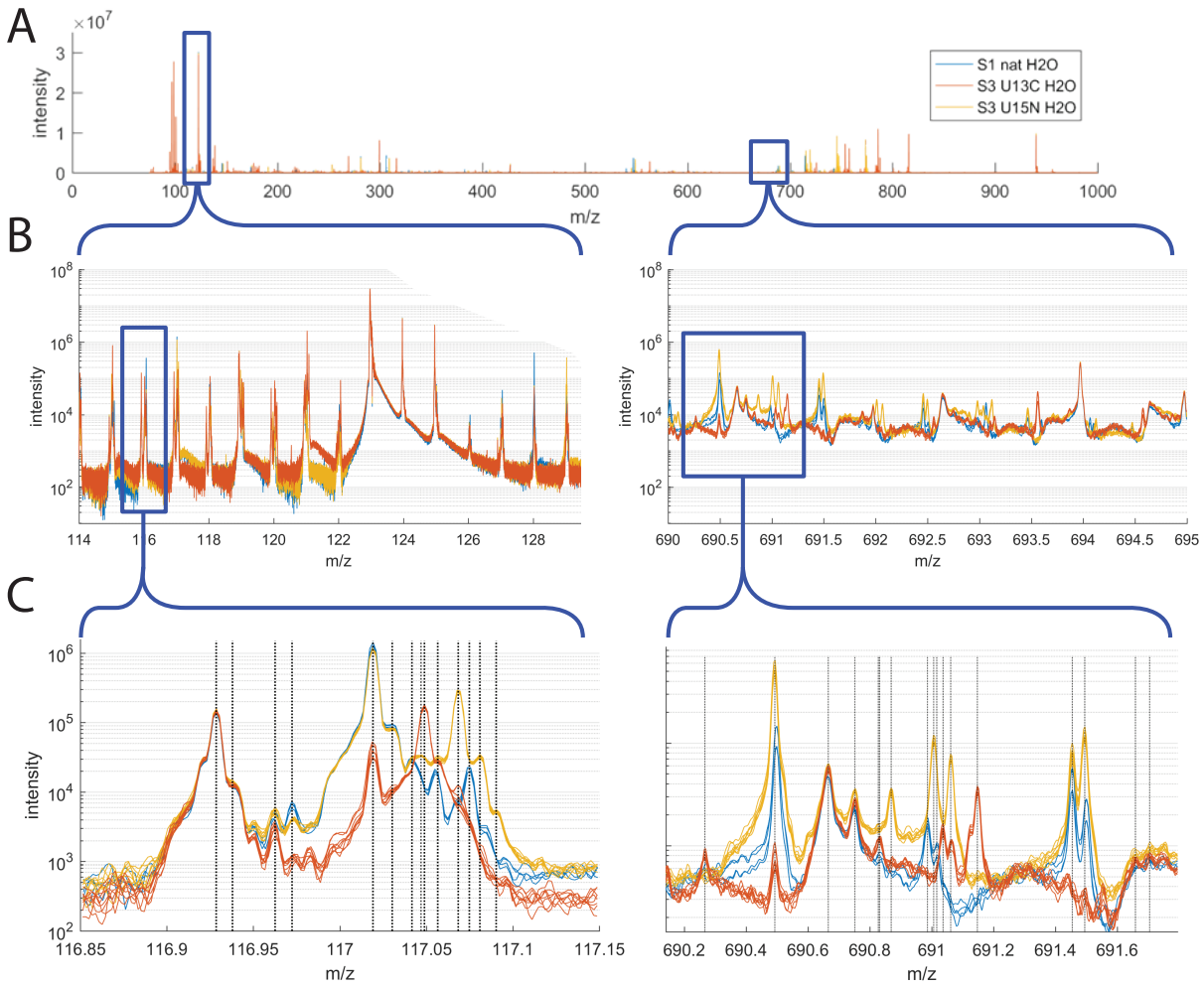


Figure 2.1: A. B. C.

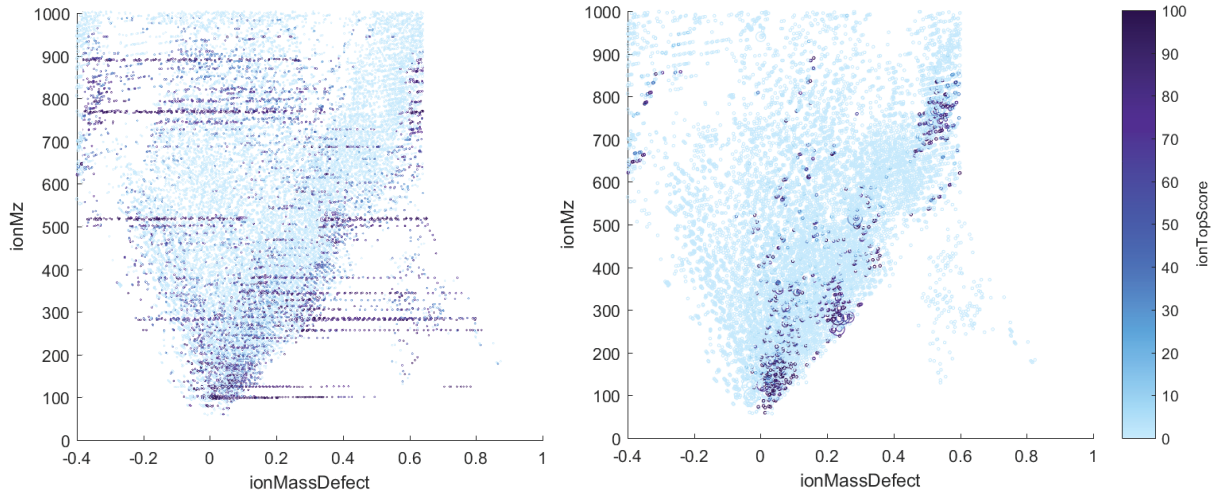
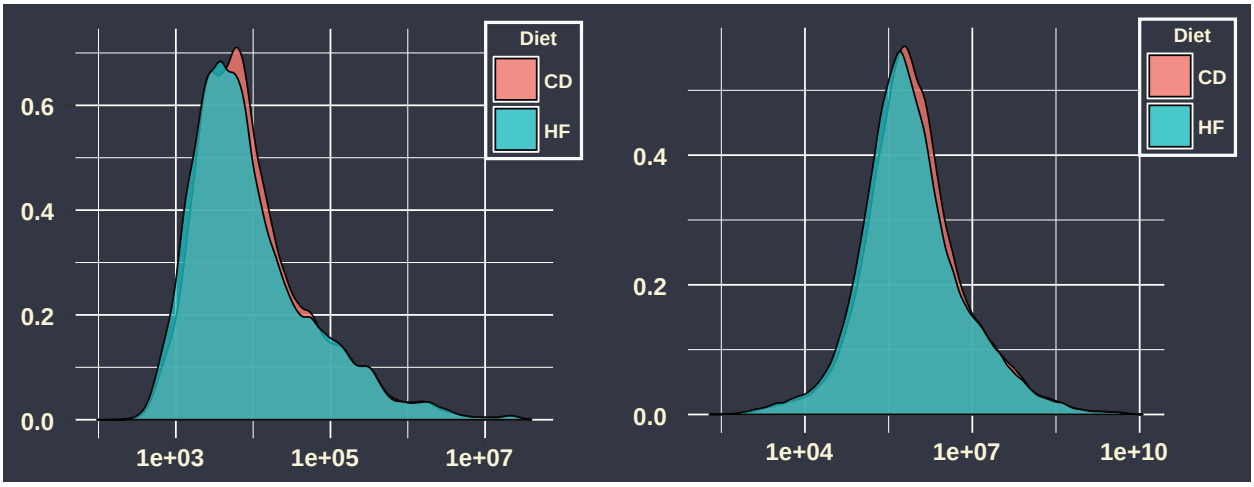
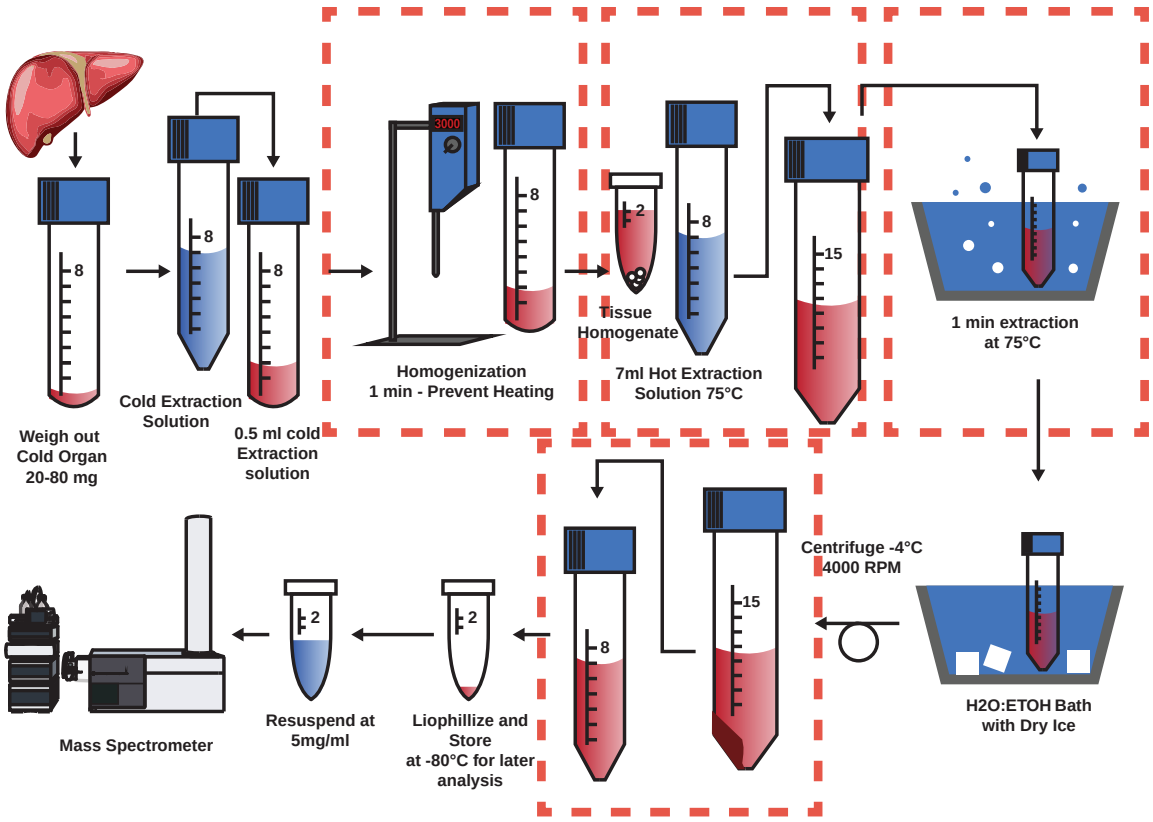


Figure 2.2: Left: All annotated peak annotated automatically
Right: Ringing Peaks and impossible mass combinations filtered



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.

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.



2.4 Extraction Conditions

Four extraction conditions were tested to determine the optimal extraction efficiency and robust metabolic converge. In a hot extraction, timing must be kept meticulously, in order to minimize degradation and restarting metabolic reactions. A homogenization step is included in the protocol to lyse cells as the buffer sufficiently micelles disrupting to dissolve cellular membrane. After homogenization, the hot bath is meant fo denature enzymes that increase the extraction rate. Like wise in the cold extraction [40:40:20] MeOH, ACN and H2O extraction mixture is meant to solubilize cells and poison enzymes. Two cold extraction times were used , 1hours and 24 hours at -20C to determine the extraction kinetics and optimal extraction times. We also attempted a 24 hour cold extraction without the homogenization step as it can introduce impurities and cross samples containments if the homogenization head is not properly cleaned. Moreover, the homogenization step adds a minute to each protocol and adds timing complication into the protocol, warranting us to determine if it can be circumvented.

2.4.1 Warm Extraction

In the warm samples extraction protocol, samples are kept at -20°, weighed in cell culture tubes to allow for the homogenization head to read the bottom of the plate. They are homogenized to lyse cells, at which point viscous heating from the homogenizer brings the sample temperature up quite quickly. It is for this reason they must rapidly be transferred into a 75° bath to prevent

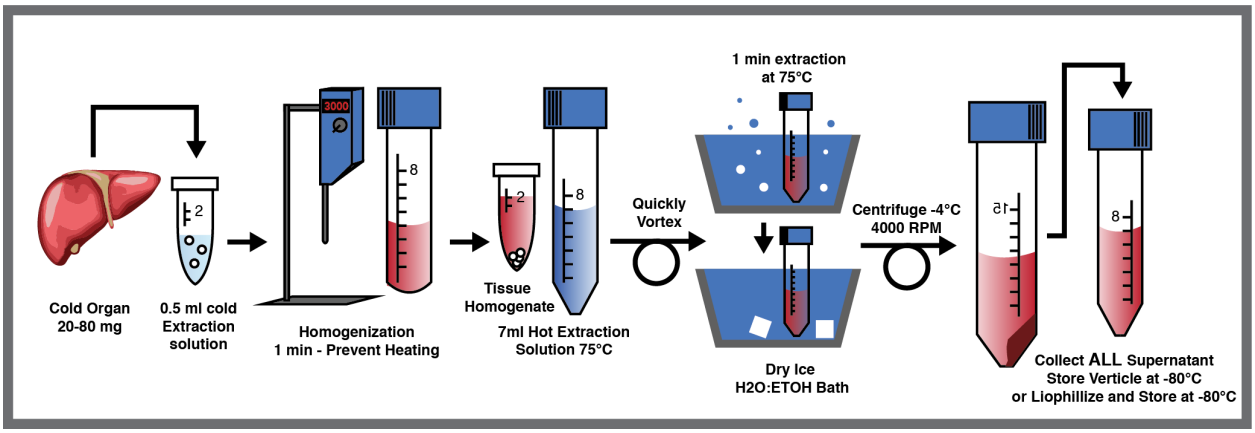


Figure 2.3: Hot Polar Metabolite Extraction Protocol

2.4.2 Cold Extraction

The cold extraction protocol uses a MeOH, ACN and H_2O mixture to quickly neutralize enzymes that may be active in the solution and thus allow for easier handling. Samples are kept at -20° while preparing and during extraction however, they are still heated during the homogenization step.

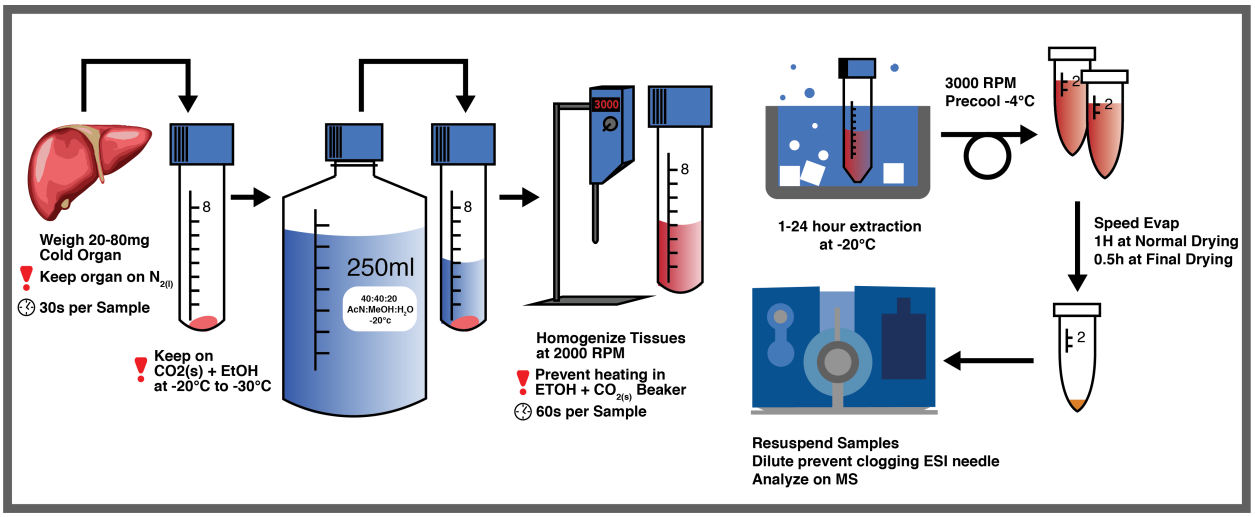


Figure 2.4: Cold Polar Metabolite Extraction Protocol

2.5 Extraction Results

where are the adjusted p values and p values generated

2.5.1 Hot and Cold Extraction Performance

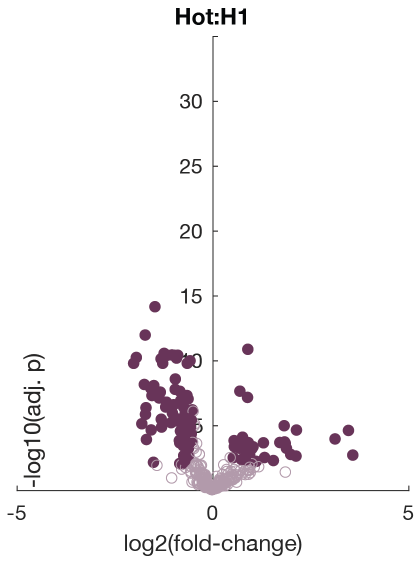


Figure 2.5: Volcano Plot of P-Values and Log2 Fold Changes seen between Hot Extraction protocol and Standard Cold Extraction Protocol

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

2.5.2 Extraction Time Performance

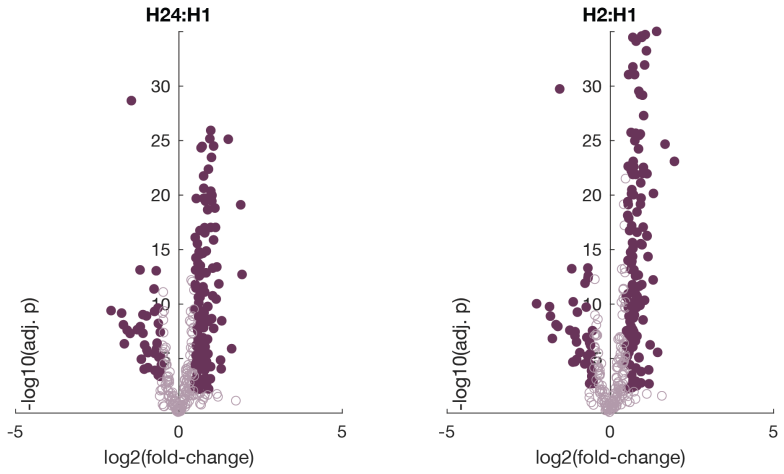


Figure 2.6: Volcano Plot of P-Values and Log2 Fold Changes seen between Standard Cold Extraction Protocols with an additional 1hour and 24 hour extraction period as compared to the standard cold extraction protocol

2.5.3 Effect of Homogenization on Performance

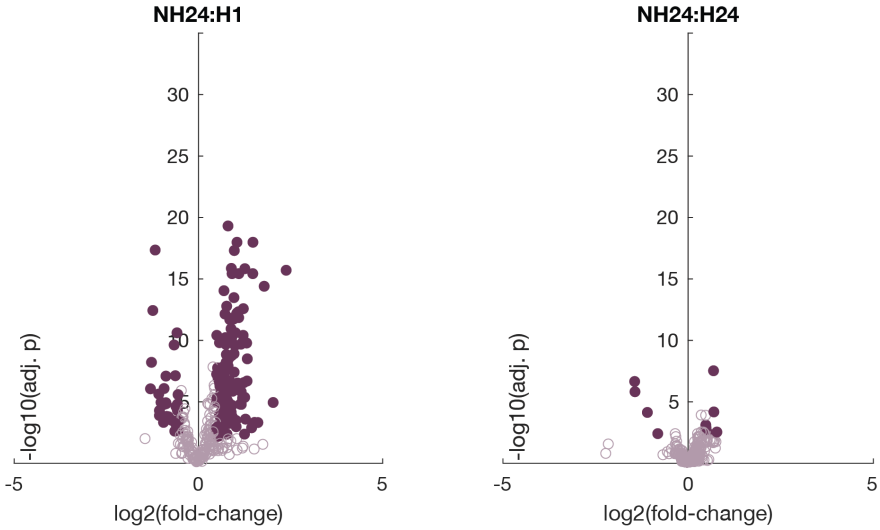
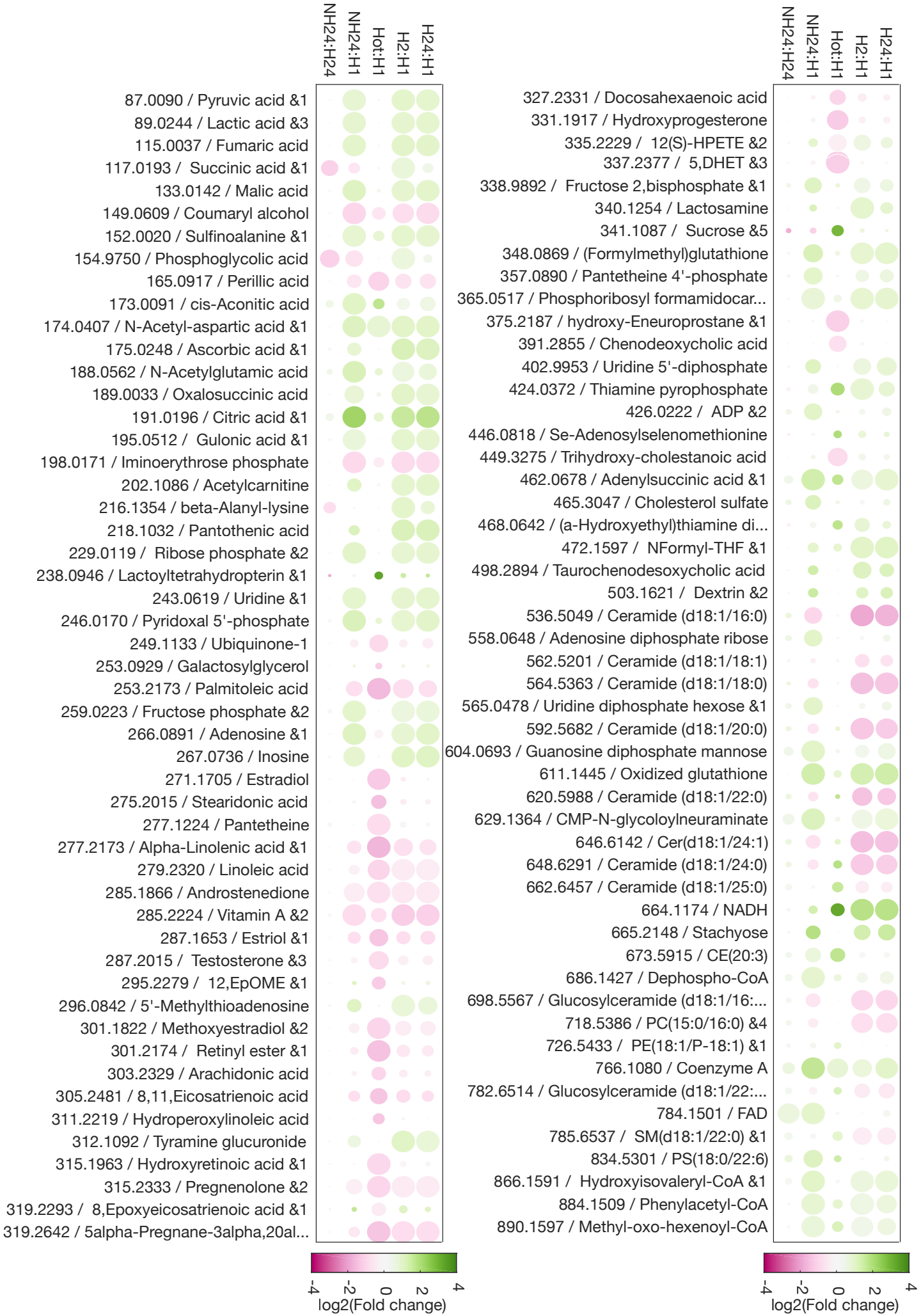
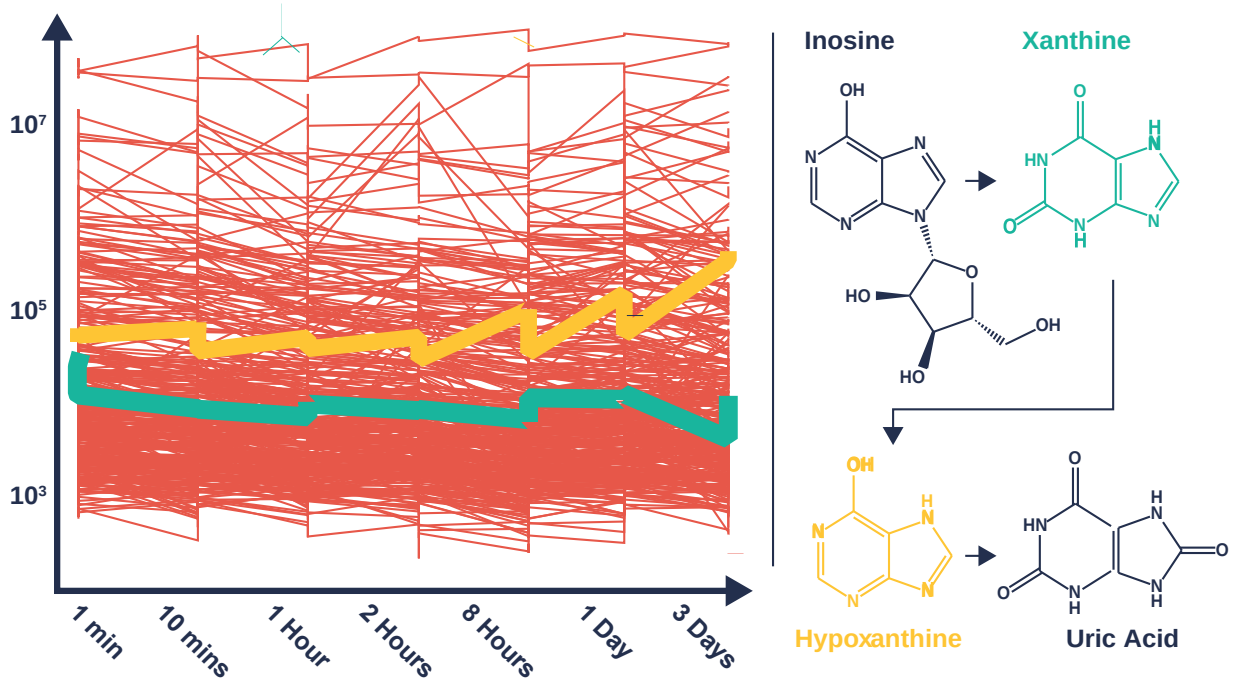


Figure 2.7: Volcano Plot of P-Values and Log2 Fold Changes seen between Standard Cold Extraction Protocols with an additional 1hour and 24 hour extraction perdion as compared to the standard cold extraction protocol


NH24:H24
NH24:H1
Hot:H1
H2:H1
H24:H1
87.0090 / Pyruvic acid &1
89.0244 / Lactic acid &3
115.0037 / Fumaric acid
117.0193 / Succinic acid &1
133.0142 / Malic acid
149.0609 / Coumaryl alcohol
152.0020 / Sulfinoalanine &1
154.9750 / Phosphoglycolic acid
165.0917 / Perillic acid
173.0091 / cis-Aconitic acid
174.0407 / N-Acetyl-aspartic acid &1
175.0248 / Ascorbic acid &1
188.0562 / N-Acetylglutamic acid
189.0033 / Oxalosuccinic acid
191.0196 / Citric acid &1
195.0512 / Gulonic acid &1
198.0171 / Iminoerythrose phosphate
202.1086 / Acetylcarnitine
216.1354 / beta-Alanyl-lysine
218.1032 / Pantothenic acid
229.0119 / Ribose phosphate &2
238.0946 / Lactoyltetrahydropterin &1
243.0619 / Uridine &1
246.0170 / Pyridoxal 5'-phosphate
249.1133 / Ubiquinone-1
253.0929 / Galactosylglycerol
253.2173 / Palmitoleic acid
259.0223 / Fructose phosphate &2
266.0891 / Adenosine &1
267.0736 / Inosine
271.1705 / Estradiol
275.2015 / Stearidonic acid
277.1224 / Pantetheine
277.2173 / Alpha-Linolenic acid &1
279.2320 / Linoleic acid
285.1866 / Androstenedione
285.2224 / Vitamin A &2
287.1653 / Estriol &1
287.2015 / Testosterone &3
295.2279 / 12,EpOME &1
296.0842 / 5'-Methylthioadenosine
301.1822 / Methoxyestradiol &2
301.2174 / Retinyl ester &1
303.2329 / Arachidonic acid
305.2481 / 8,11,Eicosatrienoic acid
311.2219 / Hydroperoxylinoleic acid
312.1092 / Tyramine glucuronide
315.1963 / Hydroxyretinoic acid &1
315.2333 / Pregnenolone &2
319.2293 / 8,Epoxyeicosatrienoic acid &1
319.2642 / 5alpha-Pregnane-3alpha,20al...
-4
-2
0
2
4
log2(Fold change)
NH24:H24
NH24:H1
Hot:H1
H2:H1
H24:H1
327.2331 / Docosahexaenoic acid
331.1917 / Hydroxyprogesterone
335.2229 / 12(S)-HPETE &2
337.2377 / 5,DHET &3
338.9892 / Fructose 2,bisphosphate &1
340.1254 / Lactosamine
341.1087 / Sucrose &5
348.0869 / (Formylmethyl)glutathione
357.0890 / Pantetheine 4'-phosphate
365.0517 / Phosphoribosyl formamidocar...
375.2187 / hydroxy-Eneuroprostane &1
391.2855 / Chenodeoxycholic acid
402.9953 / Uridine 5'-diphosphate
424.0372 / Thiamine pyrophosphate
426.0222 / ADP &2
446.0818 / Se-Adenosylselenomethionine
449.3275 / Trihydroxy-cholestanoic acid
462.0678 / Adenylsuccinic acid &1
465.3047 / Cholesterol sulfate
468.0642 / (a-Hydroxyethyl)thiamine di...
472.1597 / NFormyl-THF &1
498.2894 / Taurochenodesoxycholic acid
503.1621 / Dextrin &2
536.5049 / Ceramide (d18:1/16:0)
558.0648 / Adenosine diphosphate ribose
562.5201 / Ceramide (d18:1/18:1)
564.5363 / Ceramide (d18:1/18:0)
565.0478 / Uridine diphosphate hexose &1
592.5682 / Ceramide (d18:1/20:0)
604.0693 / Guanosine diphosphate mannose
611.1445 / Oxidized glutathione
620.5988 / Ceramide (d18:1/22:0)
629.1364 / CMP-N-glycoloylneuraminate
646.6142 / Cer(d18:1/24:1)
648.6291 / Ceramide (d18:1/24:0)
662.6457 / Ceramide (d18:1/25:0)
664.1174 / NADH
665.2148 / Stachyose
673.5915 / CE(20:3)
686.1427 / Dephospho-CoA
698.5567 / Glucosylceramide (d18:1/16:...
718.5386 / PC(15:0/16:0) &4
726.5433 / PE(18:1/P-18:1) &1
766.1080 / Coenzyme A
782.6514 / Glucosylceramide (d18:1/22:...
784.1501 / FAD
785.6537 / SM(d18:1/22:0) &1
834.5301 / PS(18:0/22:6)
866.1591 / Hydroxyisovaleryl-CoA &1
884.1509 / Phenylacetyl-CoA
890.1597 / Methyl-oxo-hexenoyl-CoA
-4
-2
0
2
4
log2(Fold change)



2.5.4 Extraction Times

2.5.5 Effect of Freeze Thaw Cycles

2.5.6 Pilot Study Results

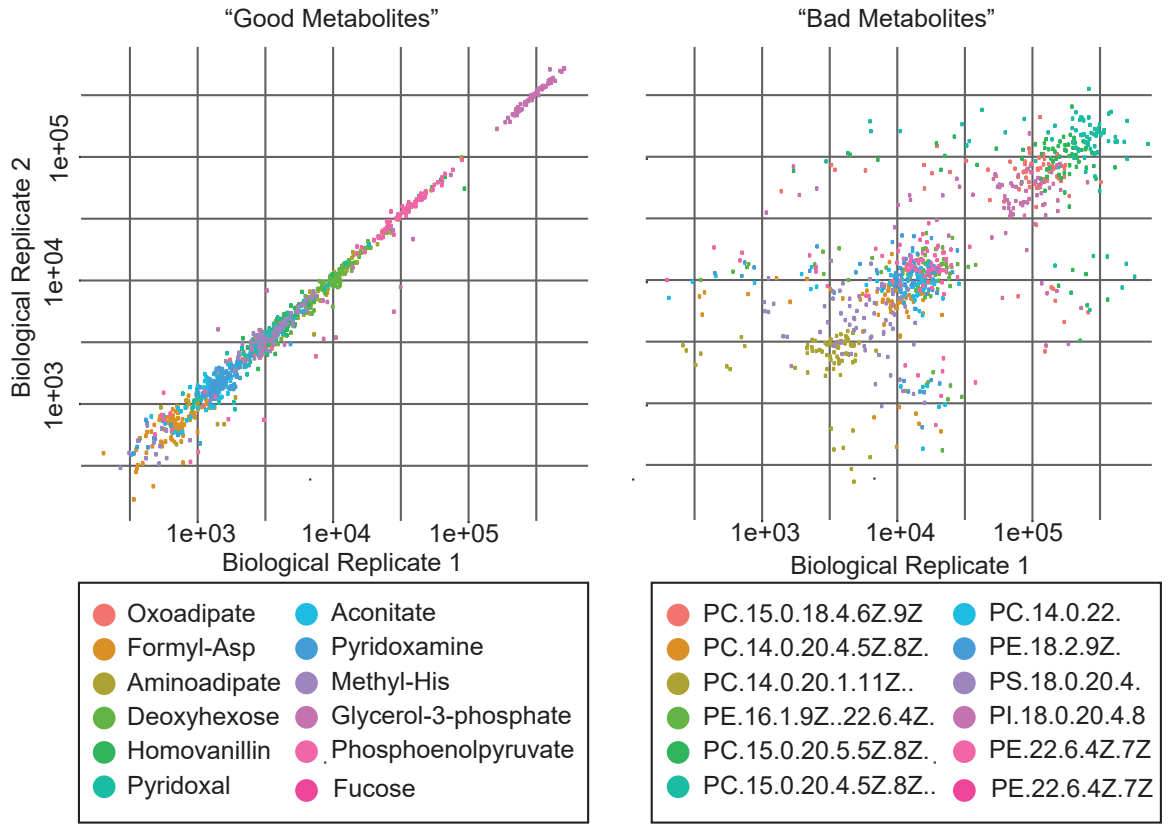
2.6 Differential Metabolites

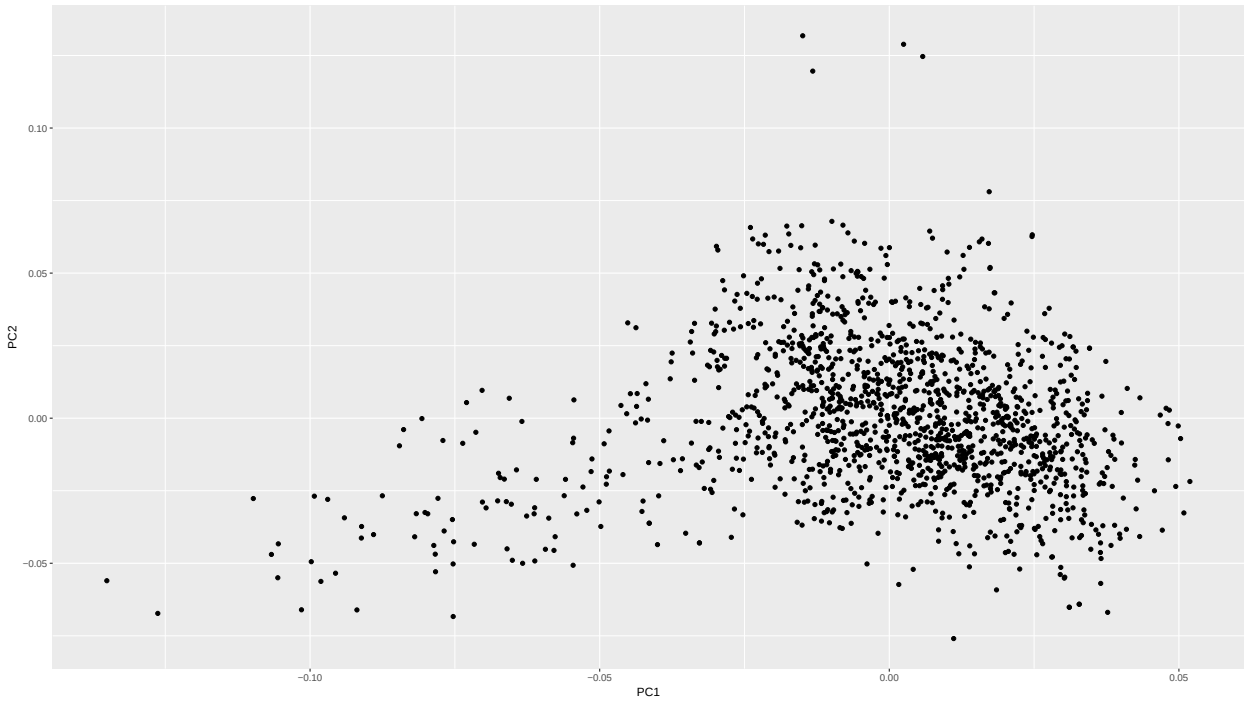
2.7 Data Conditioning

2.8 Analysis of Metabolic Data

2.9 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


“Good Metabolites”
“Bad Metabolites”
Biological Replicate 2
1e+03
1e+04
1e+05
Biological Replicate 1
Oxoadipate
Formyl-Asp
Aminoadipate
Deoxyhexose
Homovanillin
Pyridoxal
Aconitate
Pyridoxamine
Methyl-His
Glycerol-3-phosphate
Phosphoenolpyruvate
Fucose
PC.15.0.18.4.6Z.9Z
PC.14.0.20.4.5Z.8Z.
PC.14.0.20.1.11Z..
PE.16.1.9Z..22.6.4Z.
PC.15.0.20.5.5Z.8Z.
PC.15.0.20.4.5Z.8Z..
PC.14.0.22.
PE.18.2.9Z.
PS.18.0.20.4.
PI.18.0.20.4.8
PE.22.6.4Z.7Z
PE.22.6.4Z.7Z


PC2
0.10
0.05
0.00
−0.05
−0.10
−0.05
0.00
0.05
PC1

the two approaches identical.

2.10 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

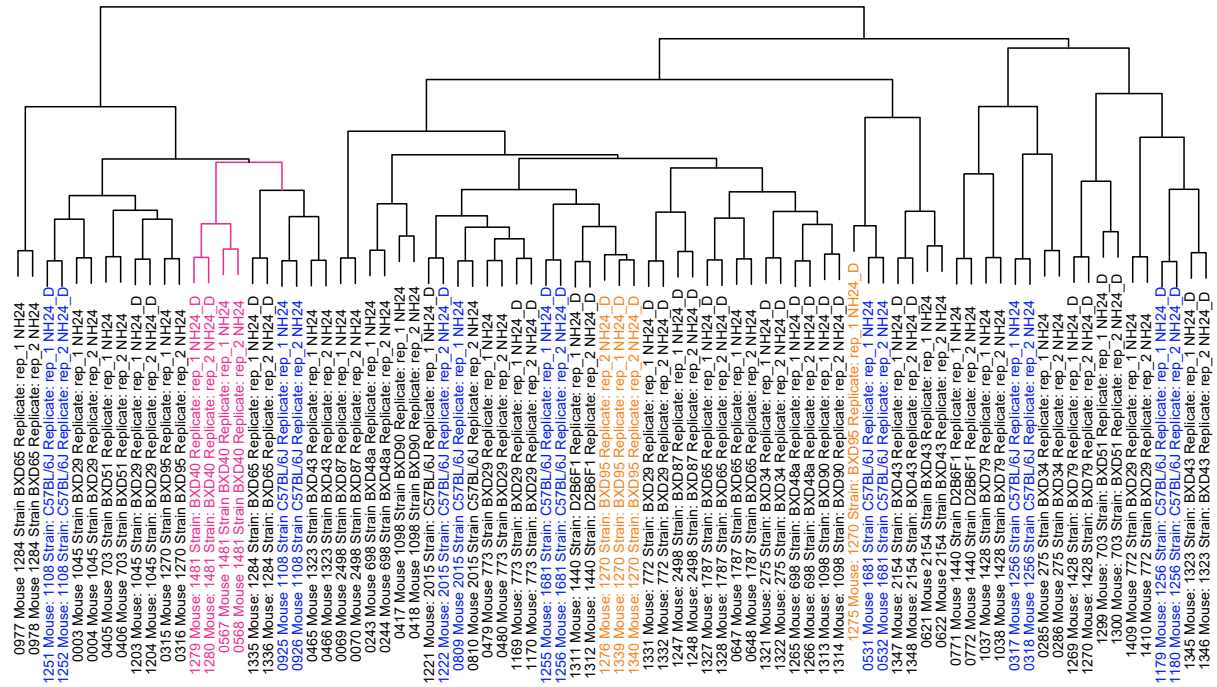
https://math.stackexchange.com/questions/128255/what-is-the-correct-definition-of-minkowski-distance

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.

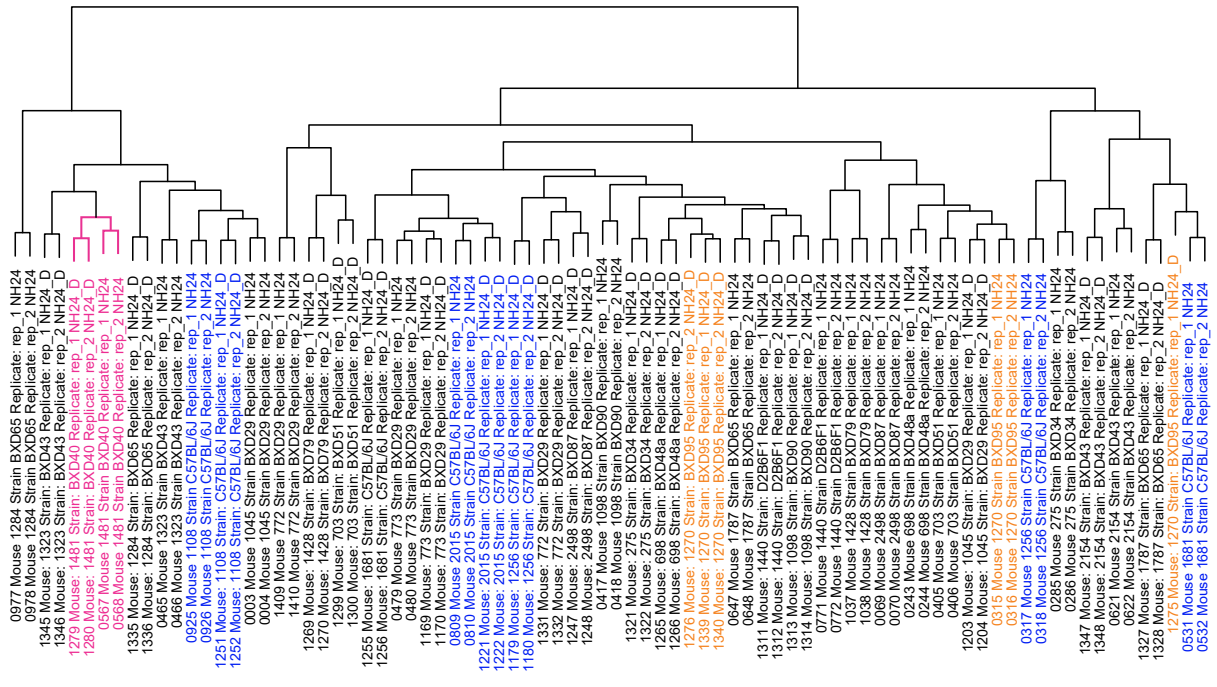
Agglomerative hierarchical clustering, instead, builds clusters incrementally, producing a dendogram. As the picture below shows, the algorithm begins by assigning each sample to its own cluster (top level). At each step, the two clusters that are the most similar are merged; the algorithm continues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

Manhattan Clustering Dendrogram



Minkowski Cluster Dendrogram



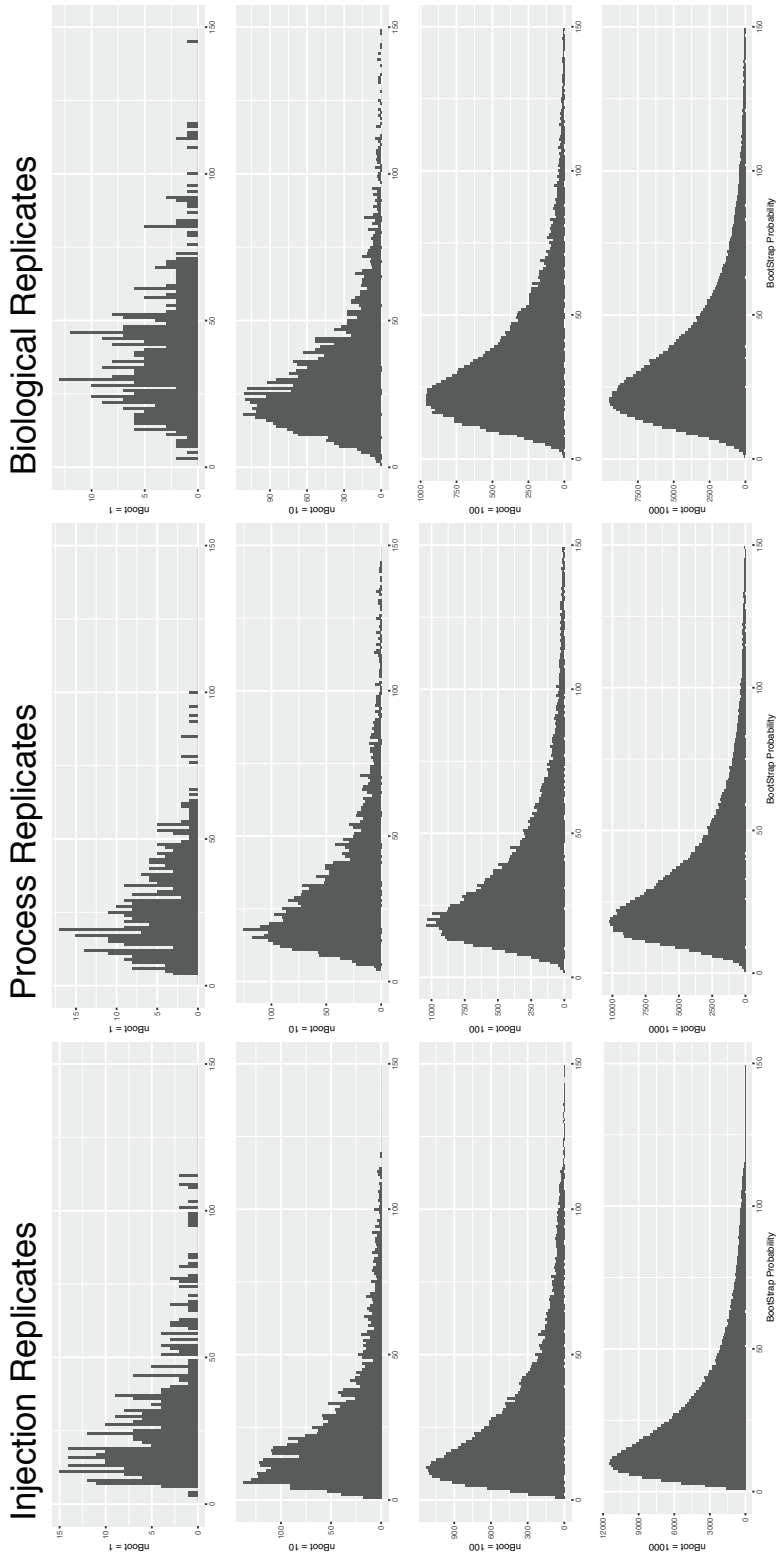


Figure 2.8: Boot Strap distribution of the CV between injections, Process Replicates and Biological Replicates

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

Chapter 3

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.

3.1 Introduction to SWATH MS

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

To date, few proteomics studies have investigated aging in mammalian tissues. Effects of senescence on the left rat heart ventricle was addressed using two-dimensional gel electrophoresis or iTRAQ labeling and matrix-assisted laser desorption/ionization (MALDI)-based quantitative mass spectrometry in which differential expression of metabolic enzymes, structural and antioxidant proteins were reported (2426). Very recently, Mao et al. published a two-dimensional gel-based time course analysis of aging mouse brain. The authors suggest that aging is associated with a reduction in abundance of proteasomal subunits and an accumulation of non-functional proteins (27). In general, the depth and reliability of quantification of the above proteome studies was low because of technical limitations of the methods used.

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 are fragmented. 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.

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

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

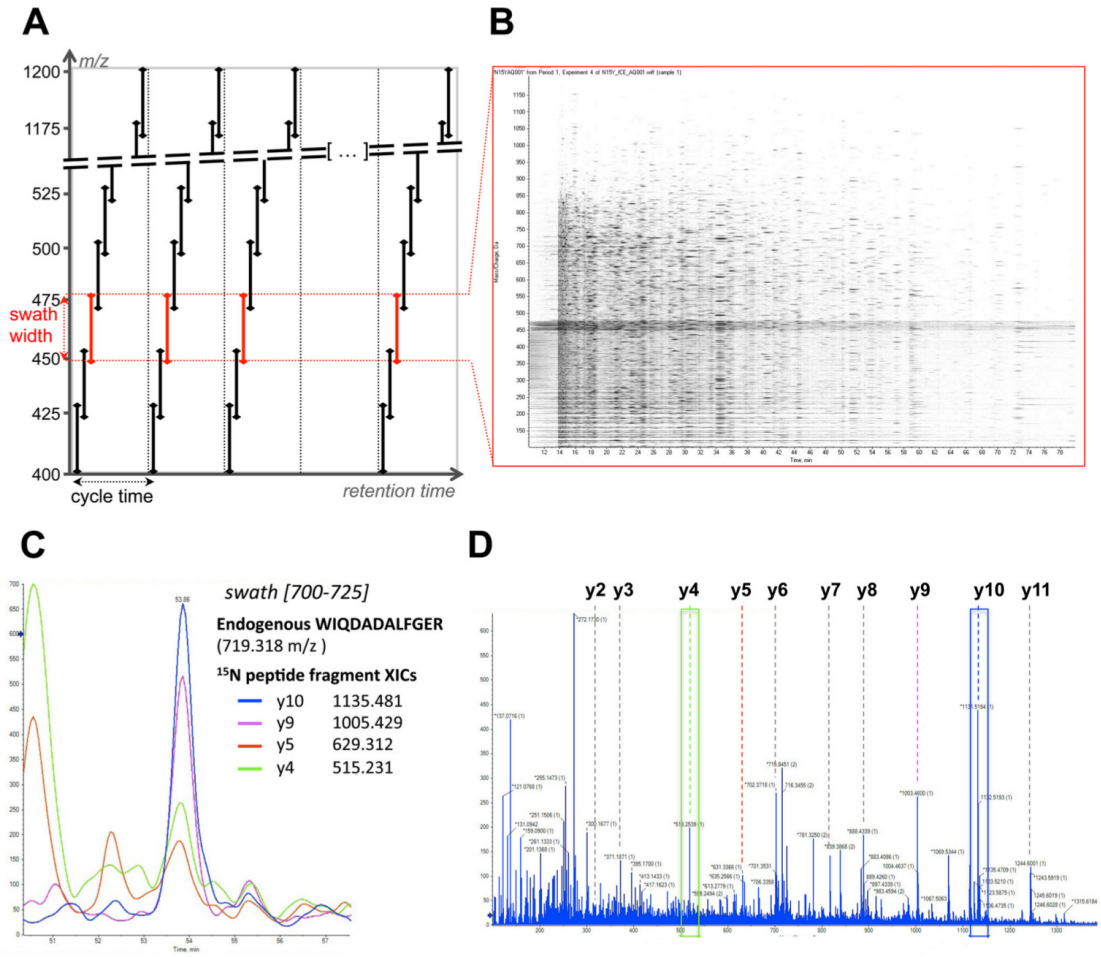


Figure 3.1: Number of RSUs that each vehicle has encountered

3.1.1 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.

3.1.2 Reagents & Materials

- Water (HPLC-grade)

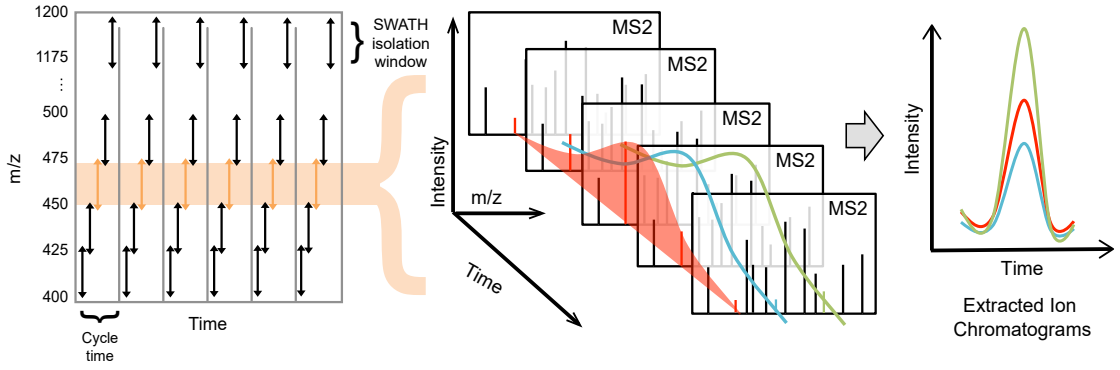


Figure 3.2: An Illustration of the SWATH-MS Duty cycle. (1) As peptides elute off an orthogonal chromatography column into the mass spectrometer, a window of mass ranges OR SWATHS are isolated and fragmented. (2) The ions that results from the fragmented peptides is recorded as a convoluted spectrum 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.

- 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)

3.1.3 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.

3.2 Results

3.3 Quality Control

3.3

3.4 Biomarker Analysis

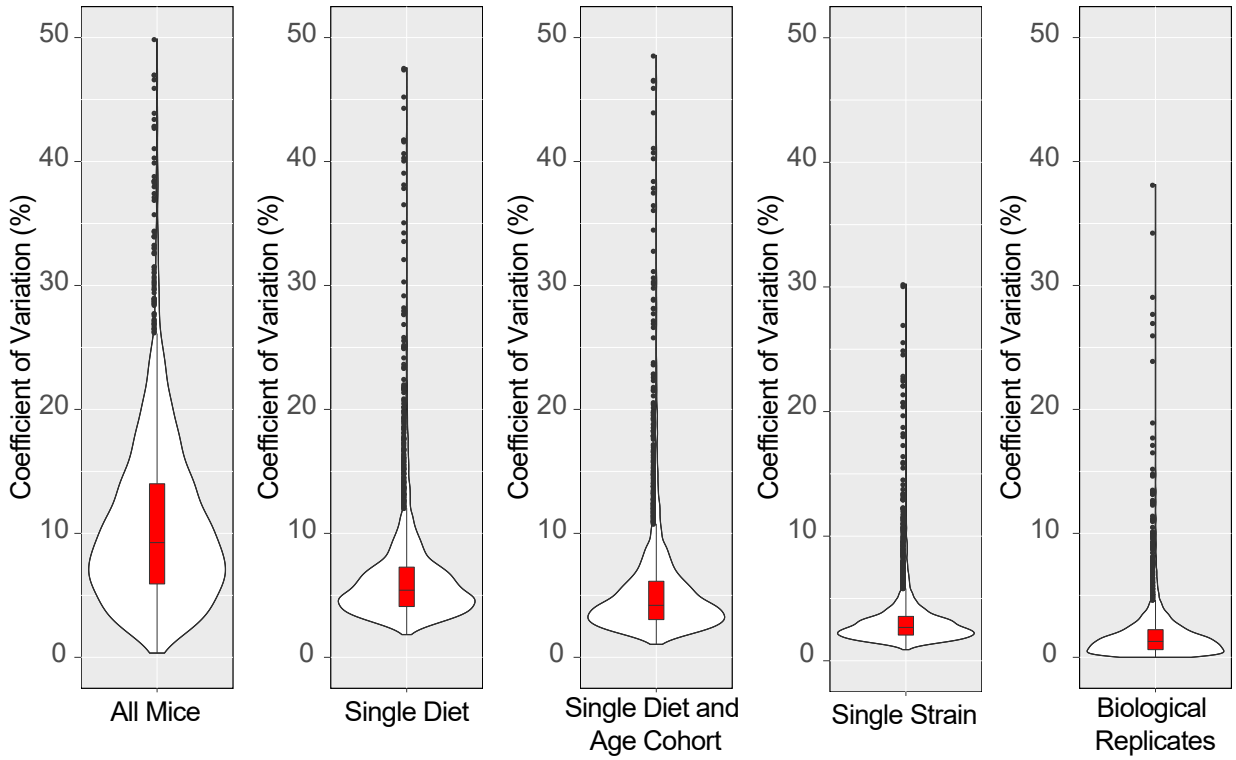
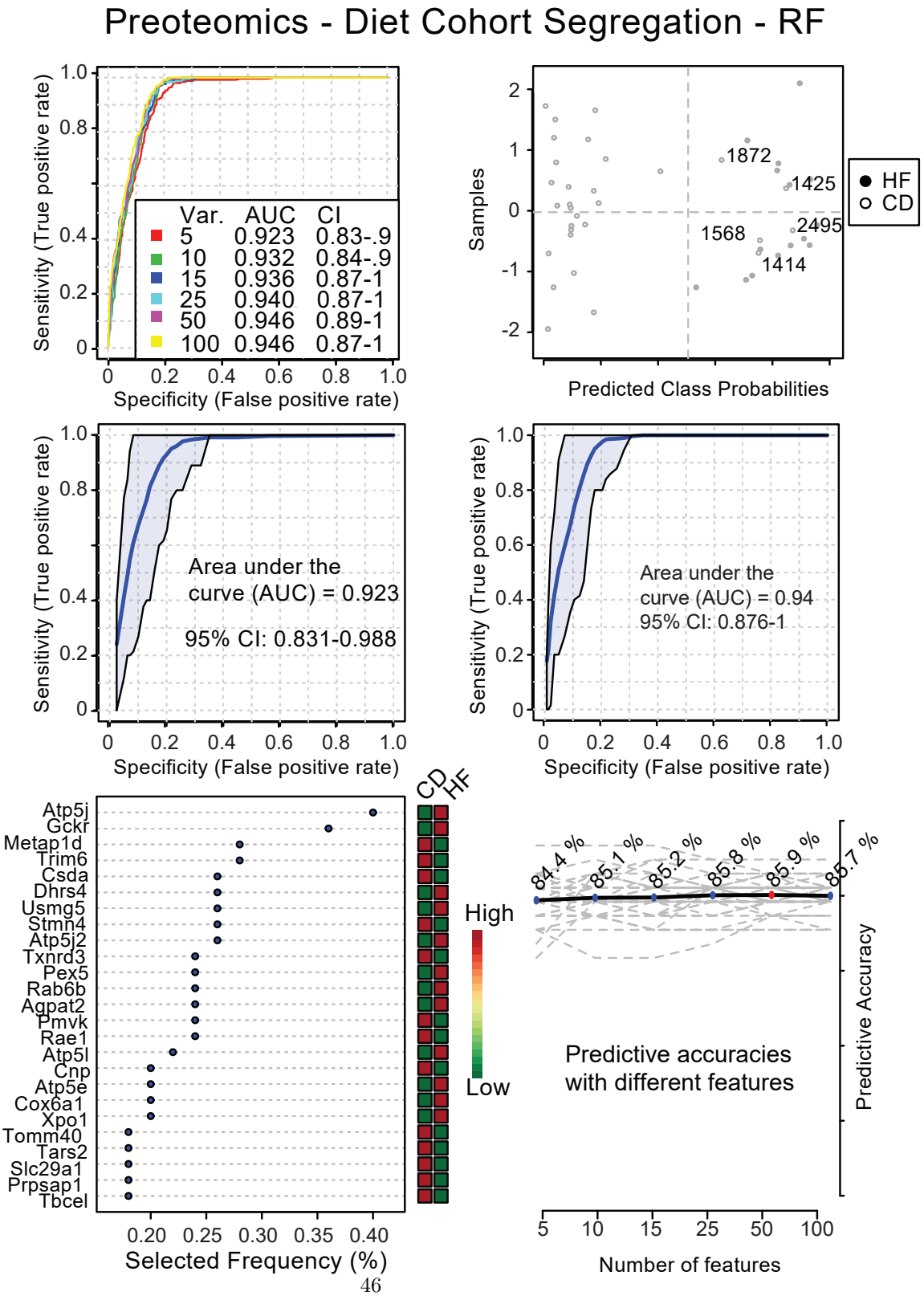
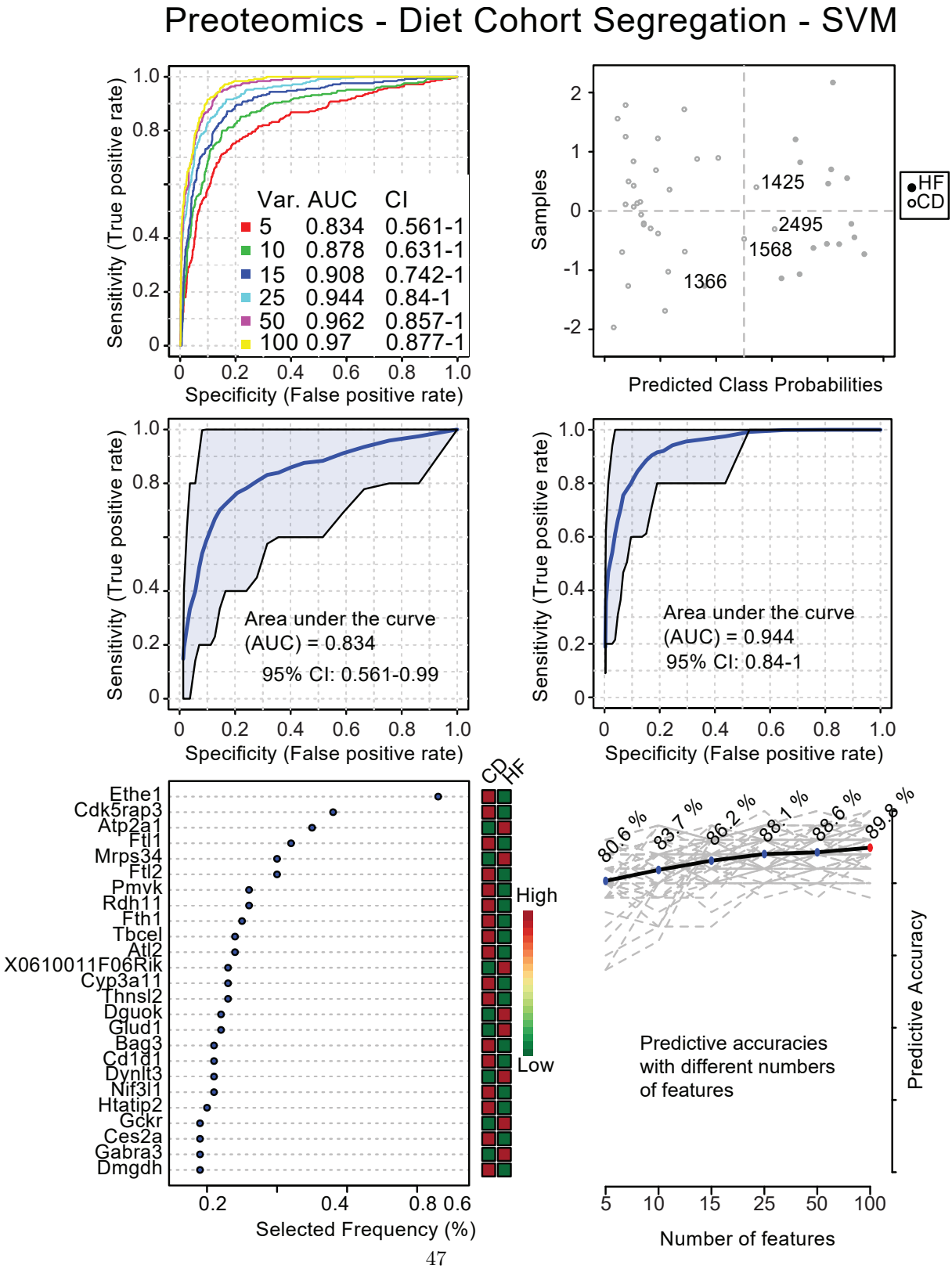


Figure 3.3: Coefficient of Variation of SWATH-MS Run between Mice measured on two days


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
Predicted Class Probabilities
1872
1425
2495
1568
1414
HF
CD
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
1425
2495
1568
1366
HF
CD
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
Selected Frequency (%)
High
Low
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 4

Transcriptomics

4.1 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

4.1.1 Microarray Experimental Methods

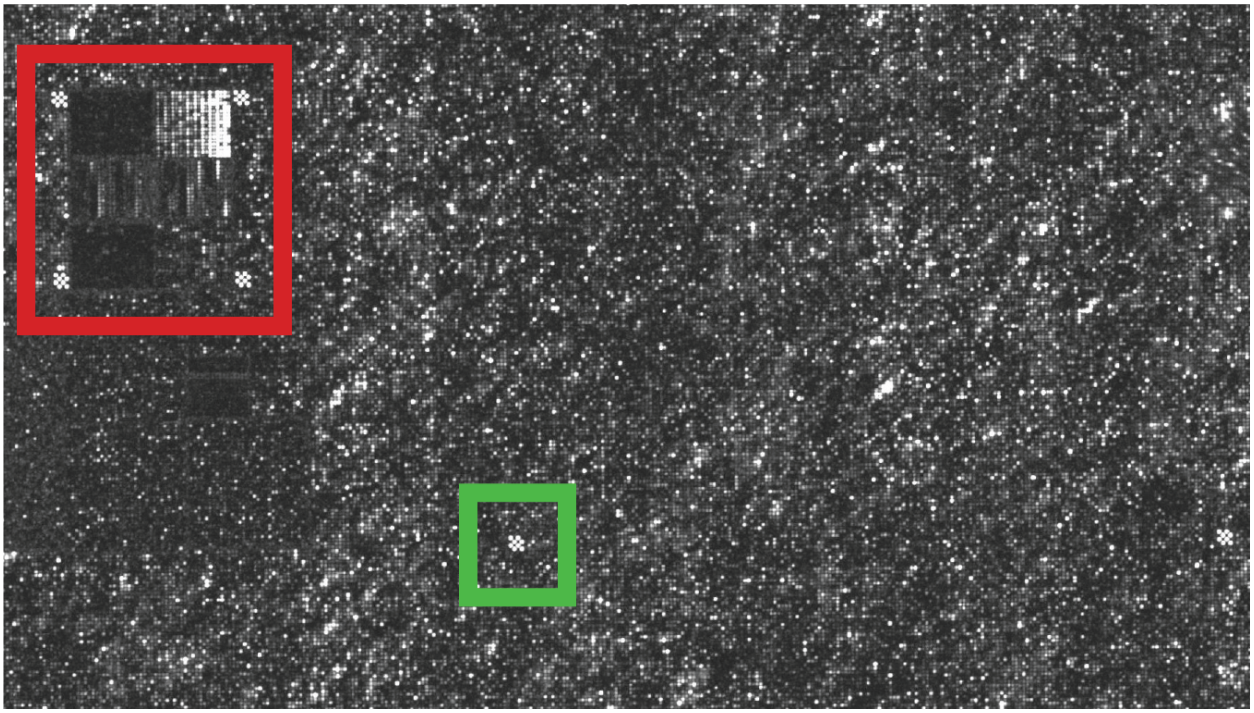
Liver tissues frozen in $N_(2l)$ were pulverized to a powder in a Motor and Pestle and shipped to the to the United states for precessing

4.1.2 Transcriptomics Data Processing

4.1.3 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

4.1.4 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.

4.1.5 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 $b_i\eta_{ki}$.

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

this yields the final model given below:

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

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

This equation allows us to model all 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.

4.1.6 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

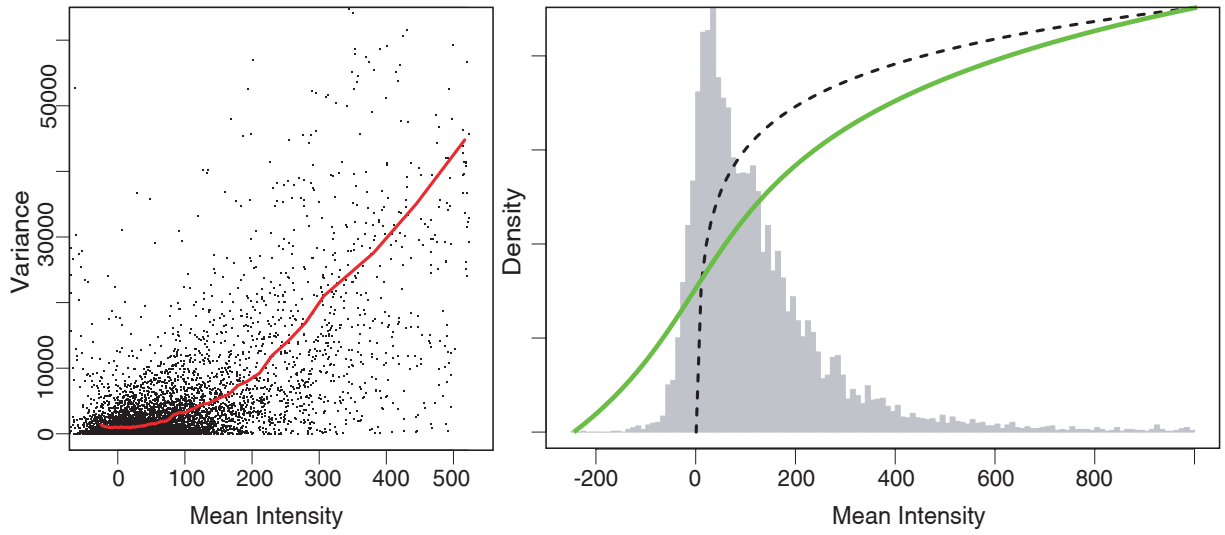


Figure 4.1: (Left) The variance and mean relationship found in experimentally produced microarrray data. The red line shows a plot of the v(u) described in at the end of section 2.3.6(right) The variance stabilization transformation performed on the data. The green line represents a log transformation, 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

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

4.1.7 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.

$$b_i\beta_k = b\beta_k$$

Figure 4.2: CV Analysis of the fist Micro array data

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

4.1.8 Results

4.1.9 Quality Control

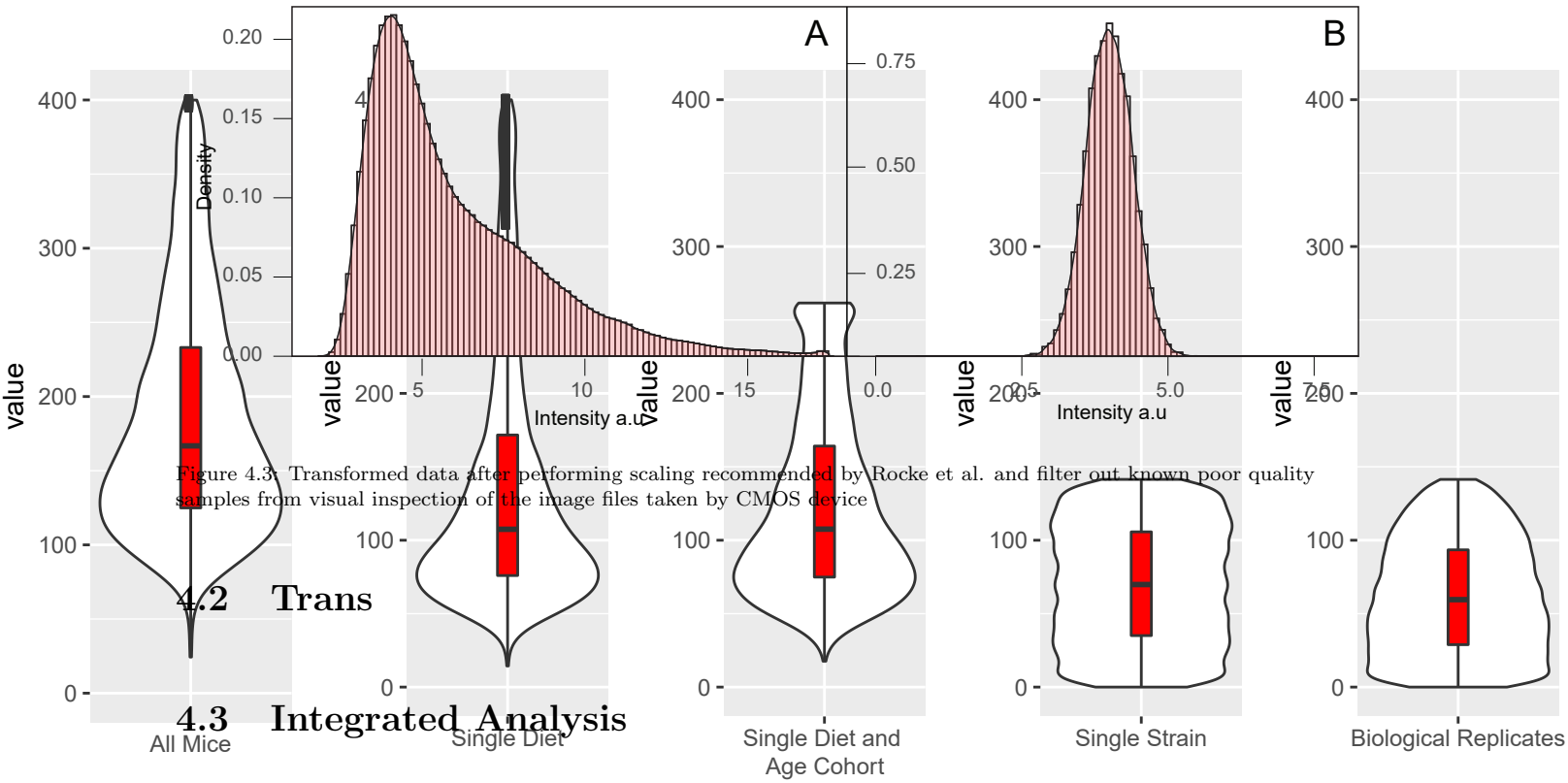


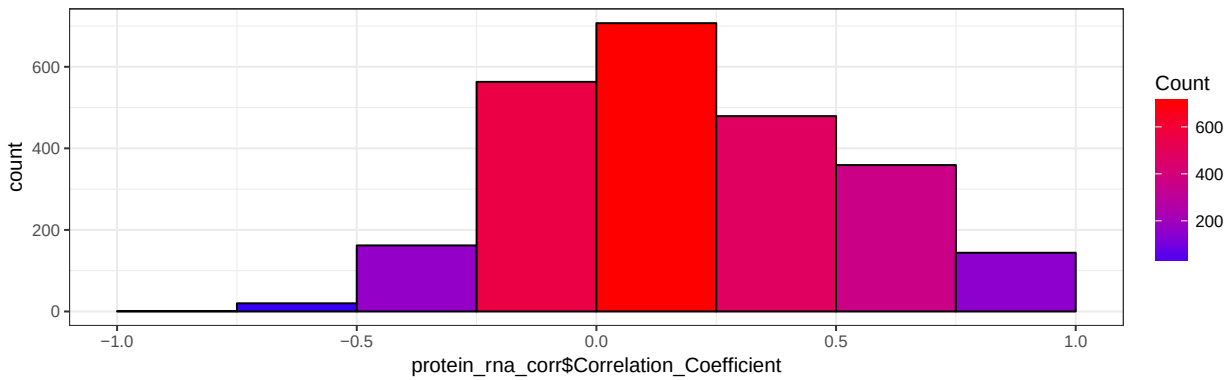
Figure 4.3: Transformed data after performing scaling recommended by Rocke et al. and filter out known poor quality samples from visual inspection of the image files taken by CMOS device

4.2 Trans

4.3 Integrated Analysis

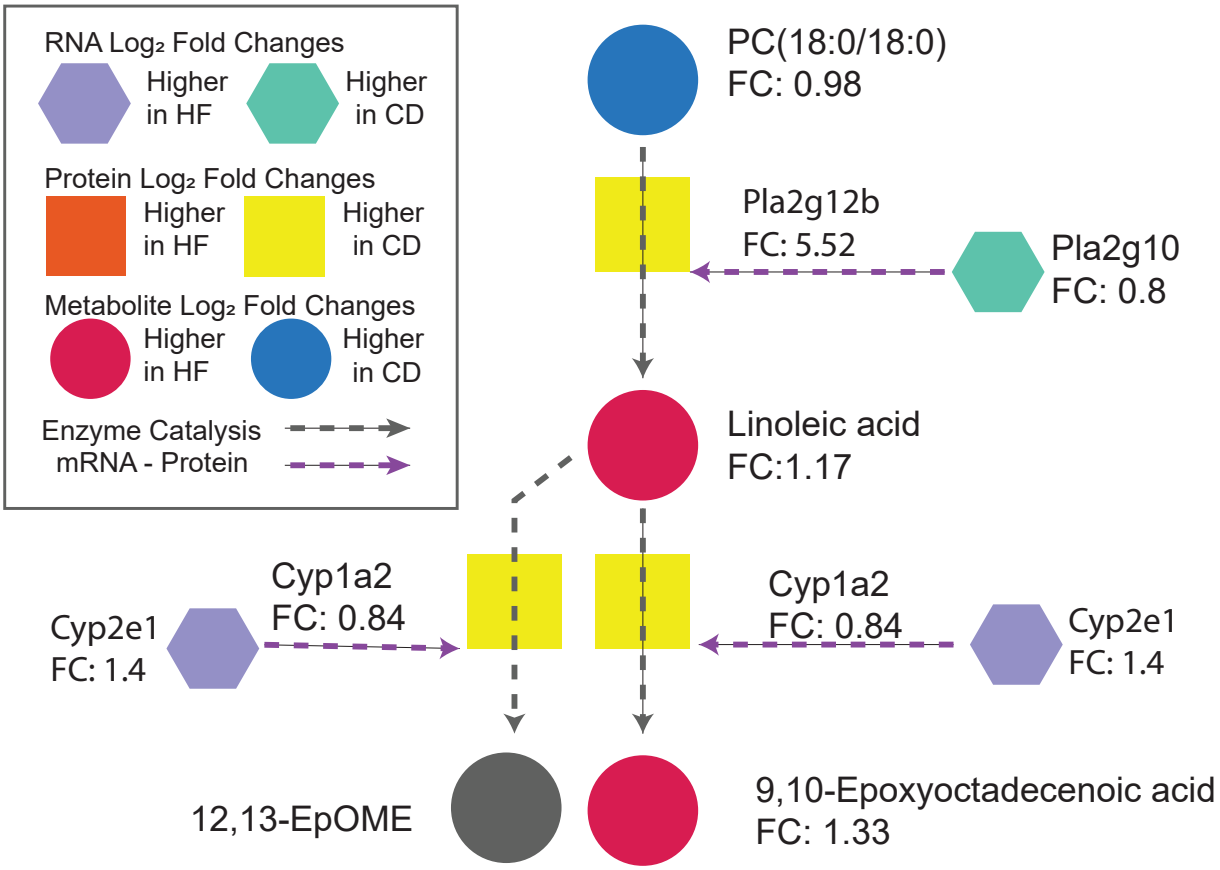
Figure 4.4: A subfigure

Figure 4.5: A subfigure



RNA/Protein/Metab Network Analysis of Linoleic Acid Metabolism

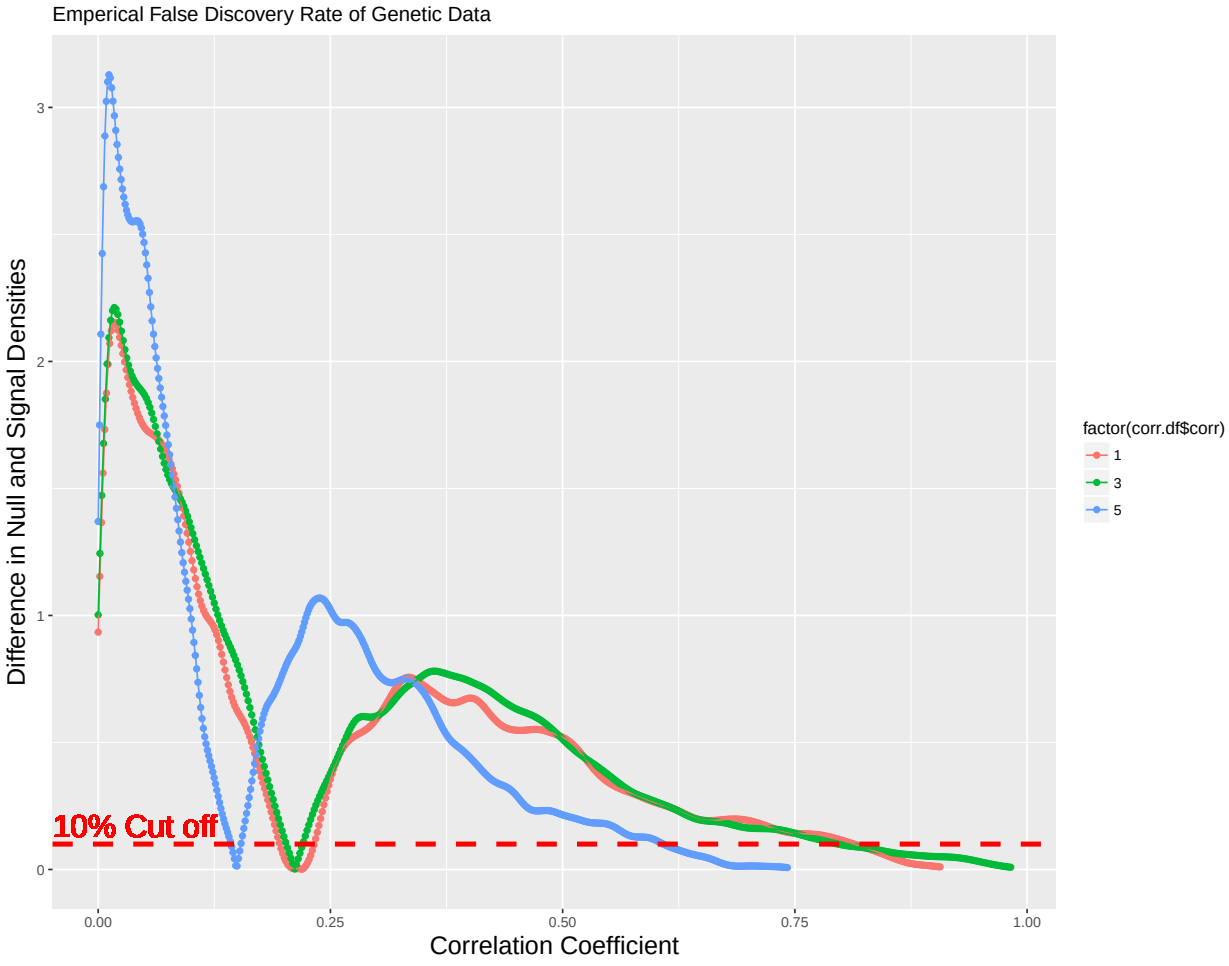
All Fold Changes are the Log_2 ratio of HF/CD



Chapter 5

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 6

Metabolite Set Analysis

6.1 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.

6.2 limitations

6.3 Diet Related Metabolite Set Enrichment Analysis

6.4 Age Related Metabolite Set Enrichment Analysis

Chapter 7

Biomarker Analysis

The ideas is to find in the very large set of data a smaller regularized set that may yield

7.1 Introduction

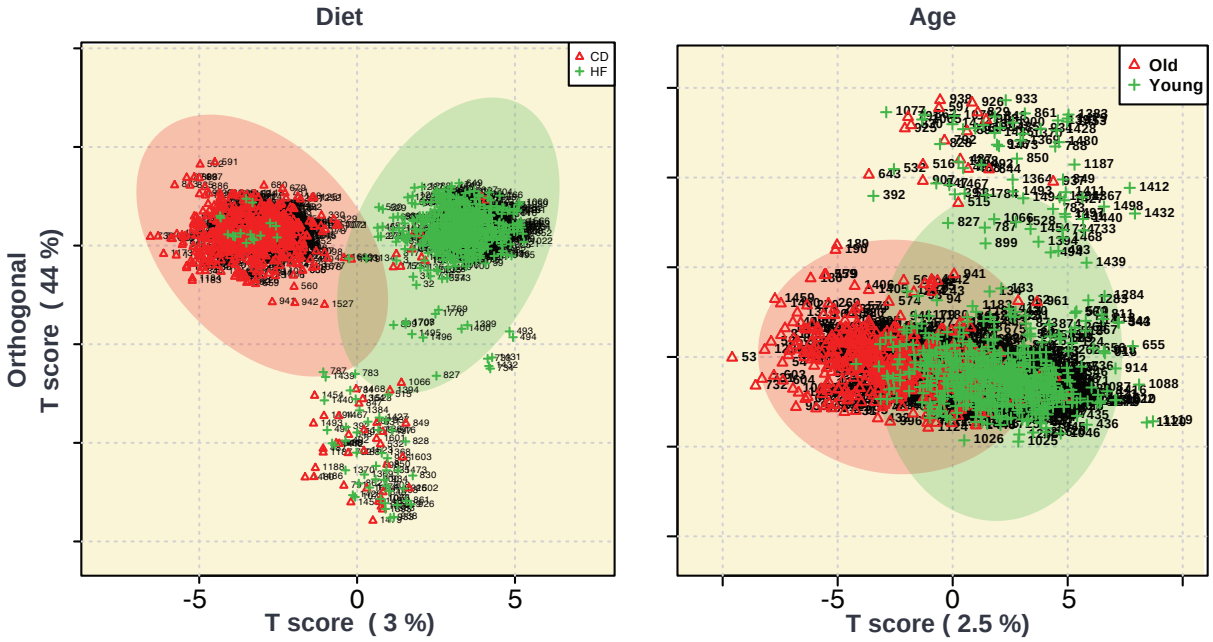
7.2 ROC Curves

7.3 PLS-Da

7.4 Tree and Random Forest

The Decision tree classifiers uses greedy approach hence an attribute chooses at first step cant be used anymore which can give better classification if used in later steps. Also it overfit the training data which can give poor results for unseen data. So, to overcome this limitation ensemble model is used. In ensemble model results from different models are combined. The result obtained from an ensemble model is usually better than the result from any one of individual models.

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.

7.5 Support Vector Machine

7.6 Neural Networks

Chapter 8

QTL and Genetic Analysis

8.1 QTL Mapping

8.1.1 Good QTL

8.1.2 Bad QTL

8.1.3 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.

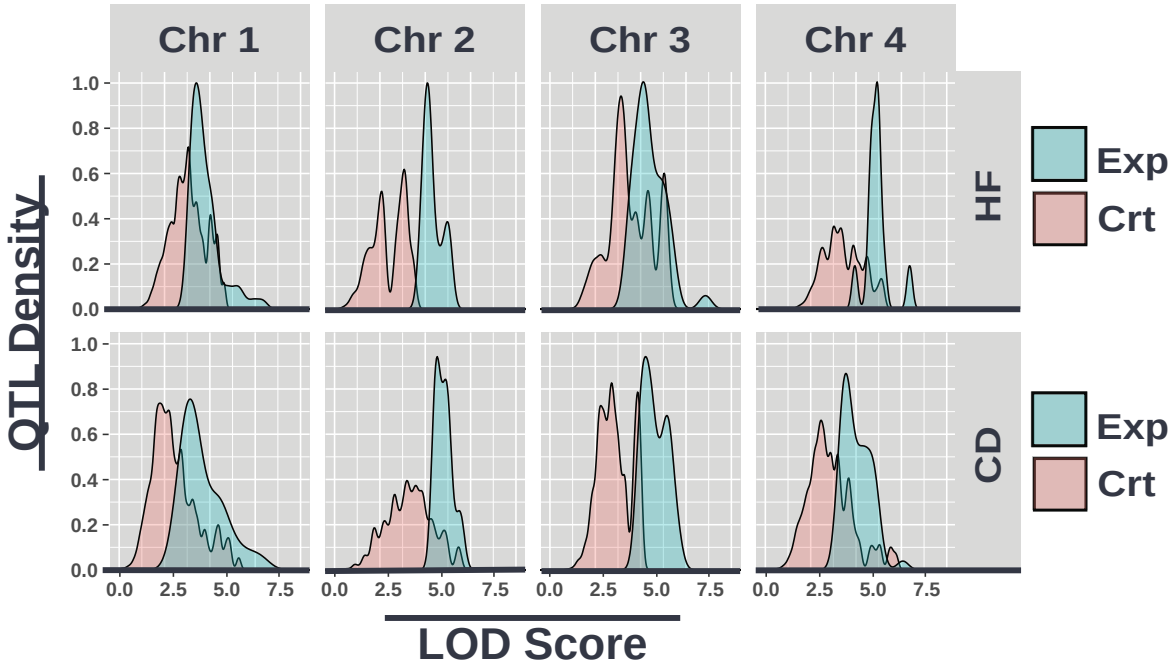
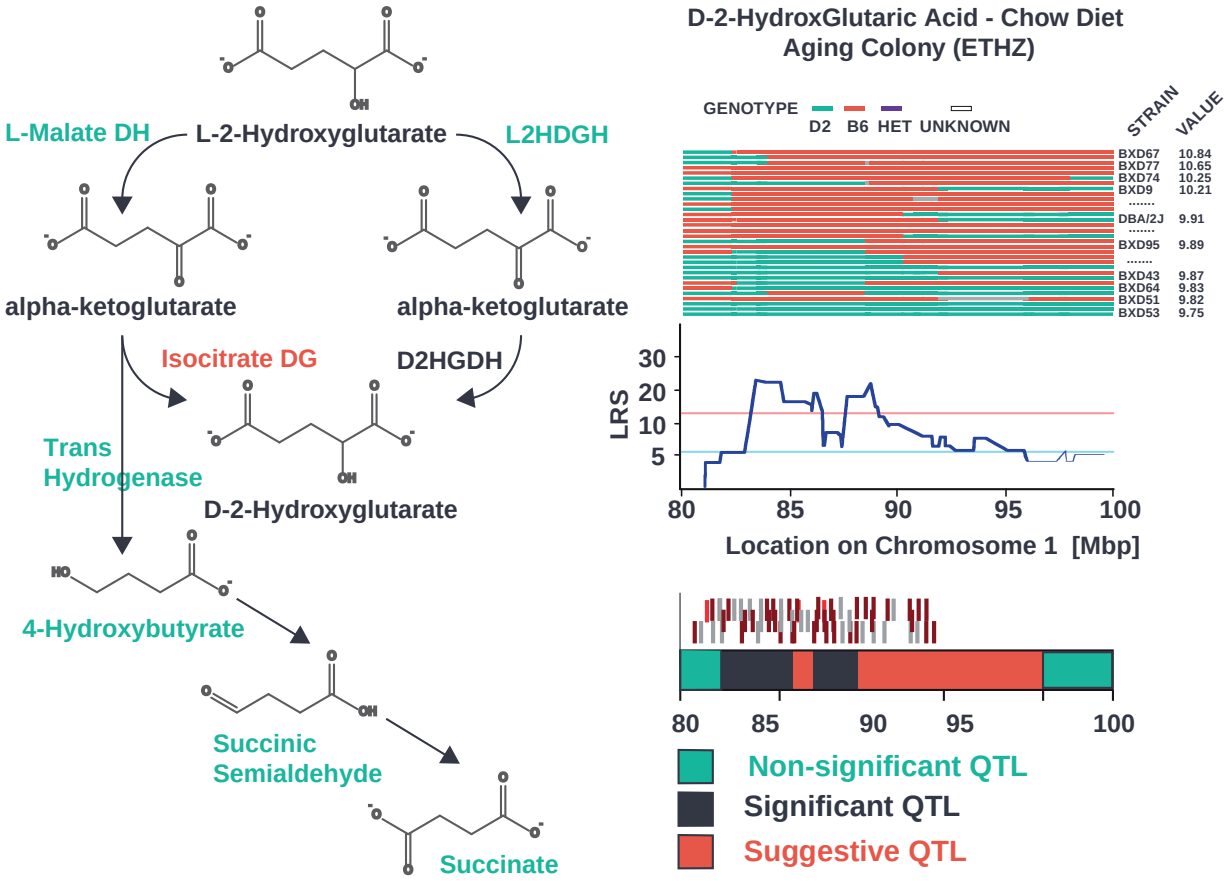
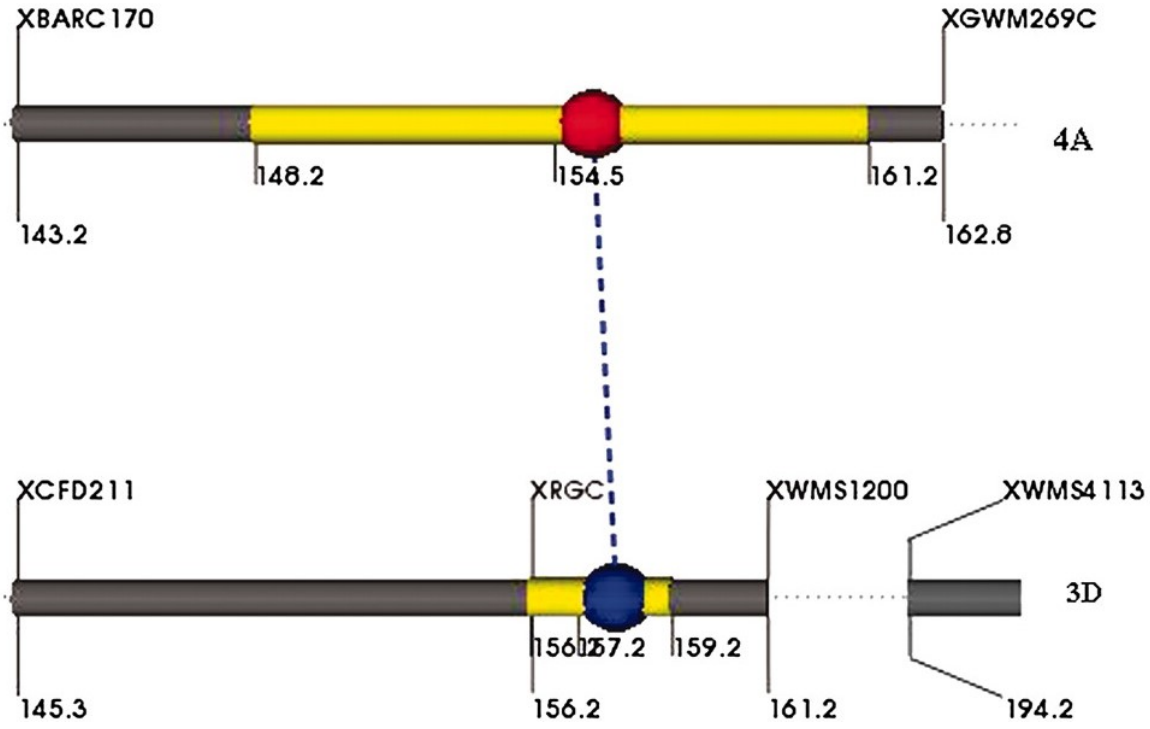

Chr 1
Chr 2
Chr 3
Chr 4
HF
CD
Exp
Crt
Exp
Crt
QTL Density
LOD Score
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5

Figure 8.1


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


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

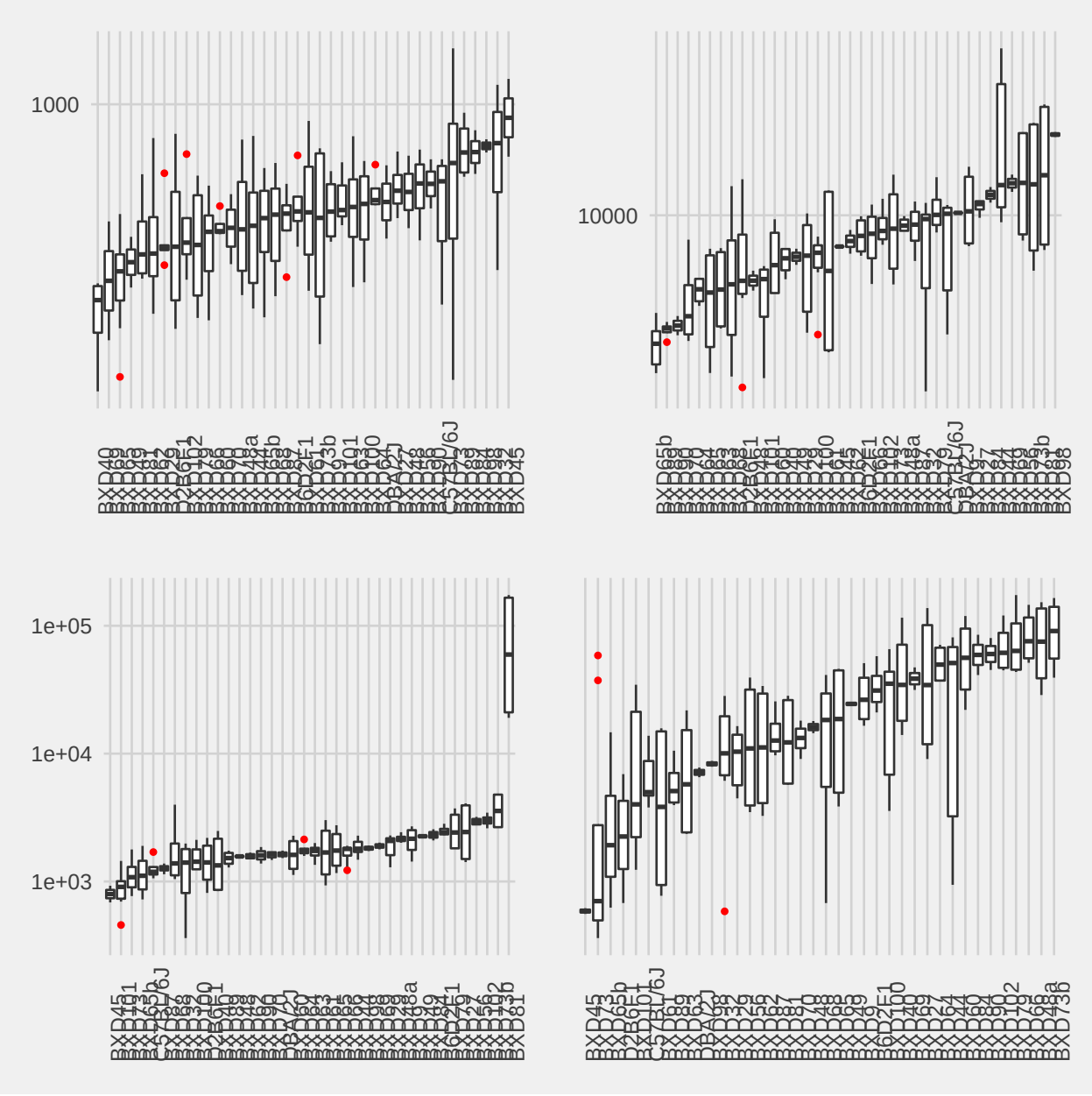
Results

Chapter 10

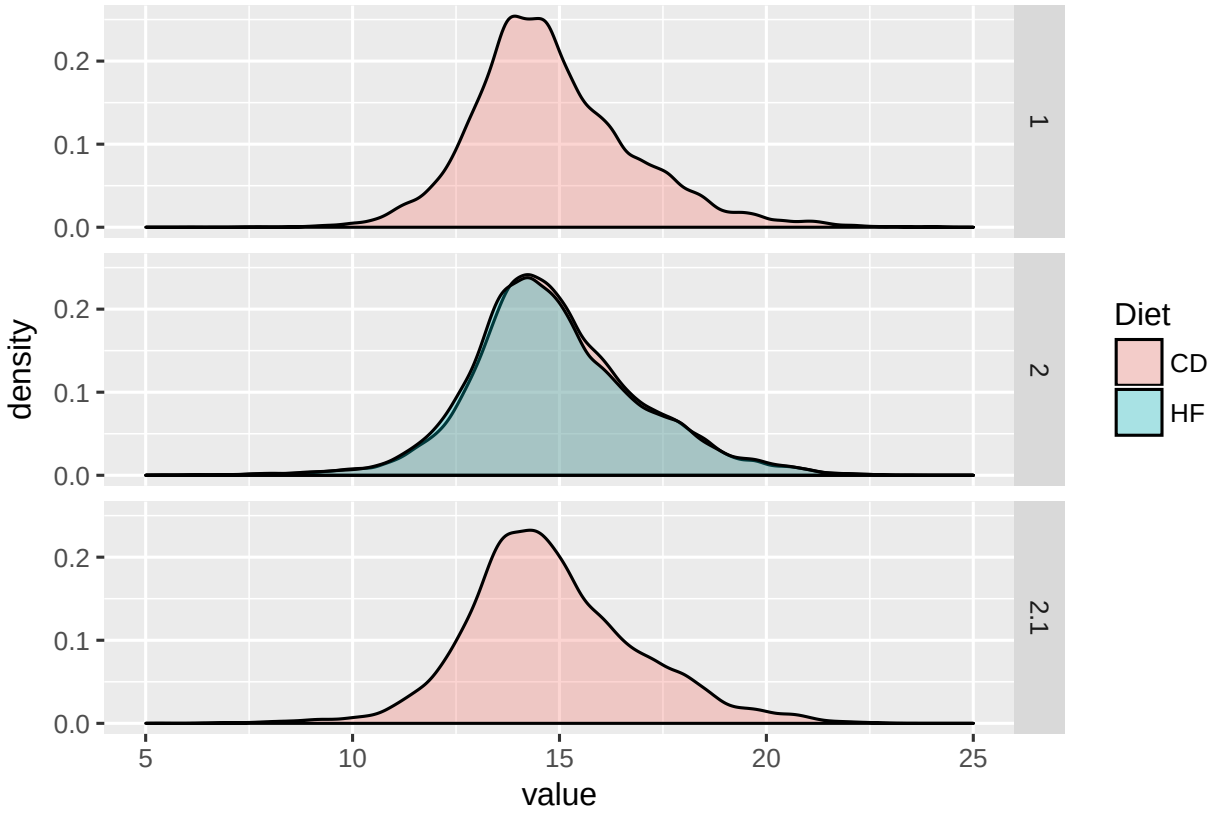
References

Full Table

Pathway Name	Match Status	p	-log(p)	Holm p	FDR	Impact
Valine, leucine and isoleucine degradation	15/38	1.9582E-5	10.841	7.0497E-4	3.6717E-5	0.27135


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