Techniques for Analyzing High Throughput Molecular Biology Data

by

Linghong Lu

M.Sc., University of Victoria, 2008

M.Sc, Yunnan University, 2005

B.Sc, Yunnan University, 2002

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Techniques for Analyzing High Throughput Molecular Biology Data

by

Linghong Lu

M.Sc., University of Victoria, 2008

M.Sc, Yunnan University, 2005

B.Sc, Yunnan University, 2002

Supervisory Committee

Dr. Mary Lesperance, Supervisor

(Department of Mathematics and Statistics)

Dr. Julie Zhou, Departmental Member

(Department of Mathematics and Statistics)

Supervisory Committee

Dr. Mary Lesperance, Supervisor

(Department of Mathematics and Statistics)

Dr. Julie Zhou, Departmental Member

(Department of Mathematics and Statistics)

ABSTRACT

The application of ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) technology to identify and quantify metabolomics data is relatively new. An important feature of the FTICR-MS metabolomics data is the high percentage of missing values. In this thesis, missing value analysis showed that the missing value percentages were up to 50% and the control treatment, NaOH.ww, had the highest missing value percentage among the treatments in the aqueous FTICR-MS sets. A simulation study was done for the FTICR-MS data to compare selection methods, the Kruskal-Wallis test and the MTP and Limma functions in Bioconductor, an open source project to facilitate the analysis of high-throughput data. The study showed that MTP was sensitive to variations among treatments, while the Kruskal-Wallis test was relatively conservative in detecting variations. As a result, MTP had a much higher false positive rate than Kruskal-Wallis test. The performance of Limma for sensitivity and false positive rate was between the Kruskal-Wallis test and MTP. Data sets with missing values were also simulated to assess the performance

of imputation methods. Study showed that variances among treatments diminished or disappeared after imputations, but no new differentially expressed masses were created. This gave us confidence in using imputation methods. Summary of analysis results of some of the frogSCOPE data sets was given in the last chapter as an illustration.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

When the first draft of the thesis was finally written, the past two months passed in quick procession before my eyes. I just couldn't believe that I pulled it off. Here I would like to thank those who made it happen:

I would like to thank my parents and friends for listening to me talking about nothing but thesis in the past two months. Especially, I would like to thank Amy and Danny, for stuffing my fridge and taking me out for fresh air. They certainly made the whole thesis writing experience less painful.

Also, I would like to thank faculties and staffs at our department for their kindness and support. My sincerest thanks go to Dr. Julie Zhou for serving on my supervisory committee.

This thesis is inspired by the frogSCOPE project with Dr. Caren Helbing. Her enthusiasm for research was always motivating and inspiring.

Last but not least, I would like to thank my supervisor Dr. Mary Lesperance. All my friends know that I have a brilliant supervisor because I just couldn't stop talking about her. I would like to take this opportunity to thank her for her endless support. She was always there to show me the right track when I needed help. It was her valuable suggestions, guidance and encouragement that made the thesis possible.

Chapter 1

Introduction

The existence of microarray and novel mass spectrometry technologies brings tremendous changes to the data structure. Biological samples are being analyzed in even greater detail, and there are usually hundreds and thousands of measurements per sample. Scientists in genomics and proteomics are more interested in measurements that differ among treatments because those measures could serve as valuable biomarkers. The final goal is to take full advantage of the advanced data identification and quantification technologies, pick out genes/proteins/masses with differential expression among treatments, and unveil something spectacular.

However, dealing with high-throughput data with parallel measurements of large numbers of molecular species means considering thousands of statistical inferences simultaneously. The critical problem hiding behind this is not very clear at first glance. Numbers in the following example will help us to visualize the problem. We know that for a hypothesis test conducted at significance level 0.05, there is only a 5% chance of incorrectly rejecting the null hypothesis if the null hypothesis is true. But when we consider a family of 100 comparisons together, where all null hypotheses are true, the probability of at least one incorrect rejection is $1 - 0.95^{100} = 99.4\%$ if

the tests are independent. Hence, when considering multiple comparisons as a whole, hypothesis tests that incorrectly reject the null hypothesis are more likely to occur. In statistics, this is called the multiple comparisons (or "multiple testing") problem.

Methods have been developed to control the false positive error rate associated with performing multiple statistical tests. These techniques generally adjust the significance level for individual comparison and require a stronger level of evidence to be observed in order for an individual comparison to be considered significant. The adjustment methods include the Bonferroni correction (Abdi, 2007) and some other less conservative correction methods, such as the Holm's multiple procedure (Holm, 1979), the Hochberg's procedure (Hochberg, 1988) and the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

The multiple comparisons problem is very common in genomic data sets. In a software project called Bioconductor (Gentleman et al, 2004), which is aimed at the analysis and comprehension of genomic data from molecular biology, the adjustment methods are implemented into some of the selection methods to facilitate the analysis. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases every year. However, most of the methods often disagree in their analyses and their results are hard to interpret.

Data sets in this study were from frogSCOPE project (Helbing et al, 2010). Tissue samples were prepared at Dr. Caren Helbing's lab. Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) and ultra performance liquid chromatography mass spectrometry (UPLC-MS) technologies were applied to identify and quantify metabolomics data, and iTraq technology was used for protein identification, characterization, and quantization. Metabolomics and iTraq were analyzed at the UVic Genome BC Proteomics Centre, while microarrays were processed and analyzed in the Helbing lab. More information on experimental design and data structure are

given in Chapter 3.

An important feature of the FTICR-MS data sets is the large amount of missing values. Simulation is an important method to compare the feasibility of imputation methods. There are many research papers on the microarray data sets, including simulation and methods discussion (Hoyle et al, 2002). But FTICR-MS metabolomics data are relatively less studied since the application of FTICR-MS technology to identify and quantify metabolites is relatively new (Han et al, 2008). In this study, features of FTICR-MS data were examined, including missing value patterns and data distributions. FTICR-MS data were simulated based on studies of real data sets. Further studies were carried out to compare different imputation methods and newly developed selection methods from the Bioconductor project.

Research topics in this thesis about the study of FTICR-MS metabolites data, including data simulation, selection method comparison and imputation method assessment, were inspired by the features of frogSCOPE FTICR-MS metabolites data, and were vital to statistical analysis for potential biomarker discovery afterward.

The organization of this thesis is as follows:

- Chapter 2

  A brief review of methods we applied or tested for the analysis of frogSCOPE data is given in this chapter. Bonferroni correction was used for adjusting p-values to control type I errors when needed; logistic regression models were applied for missing values pattern detection; Levene's test was used as a equal variance test without the requirement of normality; Kruskal-Wallis rank sum test was one of the selection methods we tested, and was carried out in data analysis to pick potential biomarkers; both Correspondence Analysis (CA) and Principal Component Analysis (PCA) were conducted to display the relative relationship between different measurements, treatments, and samples in two

dimensions; methods that were tested but not included in analysis are also listed for later reference.

- Chapter 3

  The frogSCOPE project was described in this chapter, including the experimental design and data types. There were three types of data sets in this project, transcriptomic data, proteomic data and metabolomics data. These data sets fully described the sample tissues from gene structure to protein components to metabolite elements.

- Chapter 4

  FTICR-MS metabolomics data sets displayed large missing value percentages. The application of FTICR-MS technology to metabolomics data is relatively new, therefore missing value analysis before any filtering, was necessary to detect any possible relationship between missing value patterns and treatment. The analysis in this chapter could help technicians to tune parameters during the process of gene/protein/mass identification, characterization, and quantification.

- Chapter 5

  In this chapter, we simulated FTICR-MS metabolomics data based on real data from frogSCOPE project. Three selection methods: Kruskal-Wallis test, multiple testing procedures (MTP) and linear models for microarray data (Limma) were compared. Performance of three imputation methods: Bayesian Principle Component Analysis (BPCA), Singular Value Decomposition (SVD) and Local Least Squares (LLS) were also evaluated.

- Chapter 6

  A summary of analysis results of some of the frogSCOPE data are included in

this Chapter. PCA plots and CA plots are included.

- Chapter 7

  Conclusions are stated in this chapter.

- Appendix A

  A workflow for FTICR-MS data from UVic proteomics center is attached. More tables from missing values analysis and sensitivity tests of the FTICR-MS data are included here. Boxplots and interaction plots from the data analysis of the aqu.pos set are also included.

- Appendix B

  Boxplots and interaction plots from the data analysis of the frogSCOPE Microarray Sets are included here.

- Appendix C

  Boxplots and interaction plots from the data analysis of the frogSCOPE iTraq Sets are included here.

Chapter 2

Literature Review

The frogSCOPE experiment was a two-factor factorial design. One factor, Hormone, had two levels: NaOH and T3. Another factor, Silver, had three levels: ionic, nano and well water (ww). Each combination had replicates. Multiple sample tissues were examined by microarray, QPCR, iTRAQ, FTICR-MS and UPLC-MS techniques. For the sake of unity, we refer to the genes, proteins and masses found in the tissues by the powerful analytical techniques as components. Components that differ among treatments were considered to be ecological sensors of chemicals in the environment.

Most statistical analyses, for example, the analysis of variance and the Kruskal -Wallis test (Corder and Foreman, 2009; Myles and Douglas, 1973), assume that variances are equal across groups of samples. The analyses are unreliable when the equal variances assumptions are not satisfied. We started our analysis by testing the 6 treatments having equal variance for each component. Bartlett's test (Snedecor and Cochran, 1989) was used to serve this purpose at the beginning of the frogSCOPE project. Bartlett's test, however, is sensitive to departures from normality. That is, if the samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. The Levene's test (Levene, 1960) is an alternative to

the Bartlett's test that is less sensitive to departures from normality. Therefore, the Levene's test was used to replace Bartlett's test to conduct the equal variances tests to separate the components into two groups, equal variances group and unequal variances group, at the first step of the statistical analysis.

Bonferroni correction (Abdi, 2007) was applied for the FTICR-MS data sets to obtain the test level for each comparison for a desired family-wise error rate (FWER). Missing value analysis was carried out using logistic regression models (Hilbe, 2009; Hosmer and Lemeshow, 2000), and the Kruskal -Wallis test (Corder and Foreman, 2009; Myles and Douglas, 1973) was performed on the equal variances group to detect components that differ among treatments. Correspondence Analysis (Greenacre, 1984, 2007; Nenadi and Greenacre, 2007) was conducted to identify the associations between components and the treatment combinations, while Principal Component Analysis (Jolliffe, 2002) was performed to highlight similarities and differences in the samples.

For each of the components picked by the Kruskal-Wallis test, a box plot was provided to indicate possible outliers if there were any. An interaction plot of medians with 95% confidence interval for each treatment was also provided to show the interactions between the treatments. The approximate 95% confidence interval formula used was (Brown, 1985)

$$\text{median} \pm 1.96 * 1.253 * \frac{\sigma}{\sqrt{n}},$$

where $\sigma = \frac{\text{median absolute deviation}}{0.6745}$ (Maronna et al, 2006).

Linear models using generalized least squares (Carroll and Ruppert, 1988), permutation tests (Good, 2000), linear mixed-effects models (Davidian and Giltinan, 1995) and generalized estimating equations (Liang and Zeger, 1986) were the methods we tried in the case of unequal variances and repeated measurements.

2.1 Bonferroni Correction

For hypothesis testing, the problem of multiple comparisons (also known as the multiple testing problem) results from the increase in type I error that occurs when statistical tests are used repeatedly. Bonferroni correction (Abdi, 2007) is a method used to address the problem of multiple comparisons. Let M be the number of comparisons performed and α be the desired FWER, the Bonferroni correction sets each of the individual tests at a significance level of α/M.

$\alpha = 0.05$ was applied throughout the analysis. For data sets from iTRAQ and microarray techniques, however, no components were found to be statistically significant when applying Bonferroni correction to the hypothesis tests. We used 0.05 as the significance level of each test to identify components with differential expressions. To verify the significance of those components, box plots and interaction plots were used as further selection tools.

2.2 Logistic Regression Models

FTICR-MS data and iTraq data have large numbers of missing values in general. It is hard to say whether they are technically missing values or biologically missing values. On one hand, if they are missing because of a failure to detect them, it is reasonable to treat them as NAs (not available). In this case, components with too many missing values within any of the treatment combinations need to be deleted in order to conduct statistical analysis. On the other hand, if the missing values are missing because they just do not exist in some of the treatment combinations, those components are actually the ones scientists are interested in. Deleting components with no or less observed values will filter out many important and interesting components.

With missing status as a binary indicator, 0 for missing and 1 of not missing,

logistic regression models (Hilbe, 2009; Hosmer and Lemeshow, 2000) were used to examine the missing value structures of the two-factor factorial design.

2.3 Levene's Test

Levene's test (Levene, 1960) is an inferential statistic used to assess the equality of variances in different groups. It tests the null hypothesis that the population variances are equal (called homogeneity of variance). One advantage of Levene's test is that it does not require normality of the underlying data.

The test statistic W defined as follows is tested against $F(\alpha, k-1, N-k)$.

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i(Z_{i.} - Z_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - Z_{i.})^2}, \tag{2.1}$$

where

1. k is the number of different groups to which the observations belong;
2. N is the total number of observations;
3. N_i is the number of observations in the ith group;
4. Y_{ij} is the value of the jth observations from the ith group;
5. $Z_{ij} = |Y_{ij} - \bar{Y_{i.}}|$, $\bar{Y_{i.}}$ is a mean or median of the ith group;
6. $Z_{..}$ is the mean of all Z_{ij};
7. $Z_{i.}$ is the mean of the Z_{ij} for group i.

2.4 Kruskal-Wallis Rank Sum Test

The Kruskal-Wallis rank sum test is a non-parametric method for testing equality of population medians among groups when the ANOVA normality assumptions may not apply (Corder and Foreman, 2009). Let $n_i, i = 1, 2, \ldots, k$ represent the sample sizes for each of the k groups in the data and R_i be the sum of ranks for group i. Then the Kruskal-Wallis test statistic is:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \tag{2.2}$$

This statistic has approximately a chi-square distribution with k-1 degrees of freedom if the null hypothesis of equal populations is true, $k > 3$ and $n_i \geq 5$ (Devore, 1995). That is, each of the n_i should be at least 5 for the approximation to be valid.

In this study, for the FTICR-MS data set, components with less than 10 observations in any of the 6 treatments were deleted for higher reliability of the test. For iTraq data, since the components were not very significantly different among treatments, components with less than 5 observations in any of the 6 subgroups were deleted in order to keep most of the components in the data set while the tests were still valid.

The Kruskal-Wallis test is a non-parametric method, i.e., it does not assume a normal population. However, it assumes that the shapes of the distributions in different groups are the same. In the frogSCOPE project, the n_is are small. There is not enough information to determine shape. Kruskal-Wallis tests were performed on the components that passed the equal variances tests, which ensured that at least "part" of the distribution was the same.

2.5 Correspondence Analysis

Correspondence analysis (CA) is based on the singular value decomposition (SVD) in matrix theory. To summarize the theory, let N denote the $I \times J$ data matrix with positive entries and n be the grand total of the data. Define a new matrix $P = N/n$ called correspondence matrix for notational simplicity, and let r and c be the row and column marginal totals of P respectively, $D_r = diag(r)$ and $D_c = diag(c)$. The standardized matrix is then $D_r^{-\frac{1}{2}} \times (P - rc^T) D_c^{-\frac{1}{2}}$ denoted by S, and the SVD of this matrix is $S = UD_aV^T$, where $U^TU = V^TV = I$ and D_a is the diagonal matrix of (positive) singular values in descending order $a_1 \geq a_2 \geq \cdots$. a_k^2 is the principal inertia on the kth dimension, $k = 1, 2, \ldots, K$ where $K = min\{I-1, J-1\}$. Then the principal coordinates of rows are $F = D_r^{-\frac{1}{2}} UD_a$ the principal coordinates of columns are $G = D_c^{-\frac{1}{2}} VD_a$, the standard coordinates of rows are $X = D_r^{-\frac{1}{2}} U$ and the standard coordinates of columns are $Y = D_c^{-\frac{1}{2}} V$. The row and column principal coordinates are scaled in such a way that $FD_rF^T = GD_cG^T = D_a^2$, whereas the standard coordinates have weighted sum-of-squares equal to I: $XD_rX^T = YD_cY^T = I$ (Nenadić and Greenacre, 2007).

There were two types of data for CA. Using treatment NaOH.ww as control, correspondence analysis was performed on the fold changed medians. An asymmetric map, called CA plot hereafter, was produced to represent the association between the components and the 6 treatments in a two-dimensional graph, where the components were scaled in principal coordinates and treatments were in standard coordinates (Greenacre, 2007).

On the CA plot of the fold changed medians, the further away a given point representing a component is relative to the origin, the more the observed response deviates from the control response. If the point lies close to or along the line representing a treatment condition, then the more strongly related the component response is to

that treatment (Greenacre, 2007). In order to improve the visibility, numbers were used to label the components on the CA plot.

To see the effect of different silver levels on the components clearly, CA was also done on the NaOH group and the T3 group separately. In those CA plots, we didn't use NaOH.ww as base, instead, the data was standardized by variance within treatments by dividing each observation by the square root of the pooled variance s_P^2 defined as:

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{n_1 + n_2 + \cdots + n_k - k}, \tag{2.3}$$

where n_i is the sample size of the ith sample, s_i^2 is the variance of the ith sample, and k is the number of samples being combined.

2.6 Principal Components Analysis (PCA)

The idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe, 2002).

Patterns in data can be hard to visualize in high dimensions. PCA allows us to restrict the data to a subspace of lower dimensionality, while filtering noise out and extracting and highlighting similarities and differences in the data. It is a powerful tool for analyzing data.

There are different ways to perform PCA in R. A function called pca in library pcaMethods was used to do the analysis, and the imputation method used in the analysis was Nonlinear Iterative Partial Least Squares algorithm (nipals). The PCA

R code for the jth component was

$$pca(data[,j], nPcs = 2, method = "nipals"). \tag{2.4}$$

2.7 Linear Model Using Generalized Least Squares, Permutation Test

For the unequal variance group, we did a lot of research trying to find an applicable method. There were methods for analysis of unequal variances data. For example, we could fit a linear model using generalized least squares (gls). The errors were allowed to be correlated and/or had unequal variances. But the choice of suitable weights to describe the within-group heteroscedasticity structure was a big obstacle to this test. There were a few standard classes of variance function structures available in the R library called nlme. One standard class called varIdent, which allowed different variances according to the levels of a classification factor, was tested on the components with unequal variances. For the jth component, the R function used to fit the model was gls given in library nlme,

$$\begin{aligned} gls(data[,j] \ &\sim \ Hormone.Silver, weights = varIdent(form =\sim 1|Hormone.Silver), \\ &na.action = na.omit), \end{aligned} \tag{2.5}$$

where Hormone.Silver was a categorical factor with 6 levels. The normality tests of the residuals resulted in very small p-values, which indicated that the residuals were not normally distributed, and that made the results from gls unreliable.

Another method we tried was permutation tests for linear models using an R function called lmp given in library lmPerm (Wheeler, 2010). This test is distribution

free and it was valid even when the data was drawn from non-normal populations, but an important assumption behind this test is that the observations are exchangeable under the null hypothesis which resulted in the requirement of equal variances.

No practical test was found to carry out analysis for the components in the unequal variances group. The components that didn't pass the equal variances tests were just a small portion of the data. Boxplots were used to help us visualize the data and decide if further analysis was necessary.

2.8 Linear Mixed-Effects Models and Generalized Estimating Equations

Generalized Estimating Equations (GEE) are a general method for analyzing data collected in clusters where observations within a cluster may be correlated (Liang and Zeger, 1986), while Mixed-Effects Models are often appropriate for representing dependent data (Davidian and Giltinan, 1995). The GEE approach is implemented in the R package geepack. A call to function lme in R package nlme fits a linear mixed-effects model allowing for nested random effects, and the within-group errors are allowed to be correlated.

Both methods were tested in the analysis of the morphology tissue data on the changes in stages, weights, tail lengths, etc collected over time. Examples of calling geeglm to fit GEE and lme to fit a linear mixed-effects model were

$$geeglm(response \sim time, id = id, family = poisson(\text{“}log\text{”}), data, corstr = \text{“}ar1\text{”})$$

and

$$lme(response \sim 1, random = \sim time|id, data).$$

The features of the morphology tissue data are different from other data set from frogSCOPE project, and this data set is not included in this thesis.

Chapter 3

the frogSCOPE project

A growing number of substances released into the environment have been identified as disruptors of hormone-dependent mechanisms in humans and animals. Thyroid hormones play crucial roles in regulation of growth, development and metabolism in vertebrates and are targets for endocrine disruptive agents.

The frogSCOPE project focused on the effects of one of the most extensively used nano-particles, nano silver, whose environmental impact was unknown. Many home appliances such as refrigerator or air conditioner have silver nano coating to their surfaces for an overall anti-bacterial and anti-fungal effect. Experiments were done on *Rana catesbeiana* (American bullfrog) tadpoles. Amphibian wildlife are sentinels of our environment and are used as indicators of ecosystem health. Due to their distinctive life history, they transition from aquatic to terrestrial environments and are particularly vulnerable to pollutants.

The objective of this project was to develop sensitive indicators for the determination of disruptions of tadpoles based upon transcriptomic, proteomic and metabolomic analyses of target tissues. These studies set the foundation for the long term objective of developing better predictive and diagnostic tools for aquatic environmental

protection.

3.1 Experimental Design

Three experiments were conducted using different silver (nano-silver and ionic-silver) concentrations: 6.0 μg/L, 0.6 μg/L and 0.06 μg/L. Ionic-silver was used here as a control of nano-silver. Each experiment was a two-factor factorial design. One factor, Hormone, had two levels: sodium hydroxide (NaOH) and $3, 5, 3'$-Triiodothyronine (T3). Another factor, Silver, had three levels: ionic, nano and well water. There were 6 treatment combinations in total, and each combination had replicates.

The whole experiment took 28 days. Tadpoles were kept in tanks with different type and concentration of silvers. All tadpoles were injected with either T3 or NaOH on day 4, and tissue samples were extracted on day 6 and day 28.

3.2 frogSCOPE Data

In this experiment, tissues of interest were liver, brain, tail fin and serum. Three types of data sets were generated for this project using Microarray, iTRAQ and FTICR-MS/UPLC-MS technologies. Tissues and the technologies applied are shown in the table below.

Tissue Collected	Technology	Data Type (components identified and quantified)
Serum	FTICR-MS/UPLC-MS	Metabolomics data (Masses)
Liver, Brain	Microarray	Transcriptomic data (Gene transcripts)
Liver	iTRAQ	Proteomic data (Proteins)

Table 3.1: Tissues collected and the technologies applied.

3.2.1 Metabolomics Data

There were four sets of metabolomics data from the analysis of Bullfrog tadpole (*Rana catesbeiana*) serum samples. They were labeled as hiAgd6, hiAgd28, loAgd6, and loAgd28, where 'hi' and 'lo' refer to two concentrations of the silver (nano-silver and ionic-silver), 6.0 μg/L and 0.06 μg/L respectively, and 6 and 28 were the days that the premetamorphic *Rana catesbeiana* (American bullfrog) tadpoles were exposed to the nano-silver or ionic silver.

The techniques applied were FTICR-MS and UPLC-MS. UPLC-MS is a common and standard technique, while the FTICR-MS technique is a more sensitive new technique for high-throughput metabolomic analysis. FTICR-MS was used to analyze the hiAgd6 samples as a pilot experiment. However, the overall number of metabolites obtained was relatively small and we then used the UPLC-MS technique for the remaining three treatment sets. The hiAgd6 data sets obtained from FTICR-MS were kept for statistical analysis since the use of two techniques would give us independent confirmation of metabolite masses that are measured in both.

1. The hiAgd6 sets

The hiAgd6 sets were generated first at the beginning of the project using direct infusion-FTMS (FTICR-MS) analysis method at the UVic Genome BC Proteomics Centre. There were two extraction methods Aqueous and Organic, and each extraction method had positive and negative ion modes. So there were four sets for hiAgd6: Aqu.pos, Aqu.neg, Org.pos and Org.neg. The normalized peak intensities of neutral (monoisotopic) masses were recorded for each sample in the hiAgd6 sets. A brief workflow used in the data extraction and alignment at the Proteomics Centre is listed below. A more detailed workflow for the FTICR-MS method provided by the

Proteomics Centre is given in Appendix A.1.

1. Monoisotopic peak picking using MIPP v2.0, S/N=3 cut-off
2. Peak alignment with 1D PAMD v1.1, a custom software written using the NI LabView suite (Han et al, 2008)
   (a) Adduct ions, Pos, Na; Neg, Cl
   (b) Normalization to total ion intensity, then ×10000
   (c) Alignment within 2.5 or 3 ppm & 24% (8/36, at least 8/12)
3. Output format: neutral (monoisotopic) mass vs. normalized peak intensity

Among all the four hiAgd6 sets, the positive organic phase FTMS data were the first dataset acquired at the Proteomics Centre as a test run. Every sample was acquired twice from 250 scans each time. Each acquisition took about 6 min, and it took about 10-12 hours for 36 samples plus time for instrument tuning and blank sample analysis. During the acquisition of the org.pos set, the technician at the Proteomics Centre found the overall MS signals were a little weak due to low sample concentration. Therefore, for the remaining three hiAgd6 datasets, each sample was only acquired once with the scan numbers doubled to have stronger MS signals and better data quality. One sample in the negative organic data set also had a technical replicate due to poor data quality in the first acquisition.

The original data sets obtained from the Proteomics Centre had the NaOH samples and the T3 samples recorded in separate data sets for each of the four hiAgd6 sets. In order to compare the metabolite features in the NaOH group with the T3 group, the two separate data sets were combined together by the Proteomics Centre according to our request. When combining the data sets together, criteria were needed to establish that a particular mass in the NaOH set corresponds to another slightly different mass

in the T3 set. The same metabolite feature was assigned to masses within a mass error 3 ppm across all samples in each of the four hiAgd6 sets. The raw data sets, NaOH set and T3 set, were then normalized together for each of the four analysis modes according to this criterion, and each mass in the normalized data set can be regarded as a unique metabolic feature. According to the Proteomics Centre, when combining the NaOH and T3 groups, the data sets were processed as follows:

1. For data alignment, the percentage of occurrence of each mass across all samples is 16%. i.e., all masses have to show up at least in 12 of 72 data (samples) in order to keep it.

2. 3 ppm mass accuracy is used to align and combine each unique mass across all samples.

3. All kept masses are < 1000 Da.

Before starting the statistical analysis of the normalized data sets, the technical replicates needed to be merged because the observations were from the same sample and they were highly correlated. There were many missing values in the data set, and how missing values were treated would change the data sets and finally affect the analysis results. A discussion of the missing value analysis is provided in Section 4.2. After lengthy discussions, we came to an agreement to censor the blanks (missing observations) in the data set since we could not determine whether the blanks were due to true zero values or due to suppression effects. For all the technical replicates, if there was a blank in one of the replicates, the non-missing value was taken. If there were two values, the average of those was taken.

After merging the technical replicates, the following eight samples were removed from the hiAgd6 sets due to poor quality. The data for those samples were largely NA and where there were values, they were lower than the values of the other treatment

groups.

1. Aqu.Neg - sample 73 and sample 65;

2. Aqu.Pos - sample 73;

3. Org.Neg - sample 73 and sample 27;

4. Org.Pos - sample 49, sample 53 and sample 75.

The final sample allocation of the FTICR-MS hiAgd6 sets is in the first section of Table 3.2.

2. The hiAgd28, loAgd6, and loAgd28 sets

The hiAgd28, loAgd6, and loAgd28 sets were generated by the UVic Genome BC Proteomics Centre when I was writing my thesis. The UPLC-MS analysis method was applied instead of the FTICR-MS method. Each of them contained four sets: HT3.Neg, HT3.Pos, C8.Neg, and C8.Pos. Note that HT3 and C8 refered to the column types used and had no relation to the chemical treatments. Sample allocation of those three sets is given in Table 3.2.

Unlike the FTICR-MS data, the UPLC-MS data did not have many missing values and the magnitudes of the observed values were quite large.

3.2.2 Transcriptomic Data

We had two sets of transcriptomic data from liver and brain tissue samples using microarray technology. The sample tissues were exposed to thyroid hormone and/or 6.0 μg/L silver and extracted on day 6.

A DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, each containing picomoles (10-12 moles) of a specific DNA

Data sets		NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Total
hiAgd6	Aqu.pos	12	12	12	12	11	12	71
	Aqu.neg	12	11	12	12	11	12	70
	Org.pos	11	10	11	12	11	12	67
	Org.neg	12	11	12	12	11	11	70
loAgd6	HT3.pos	12	12	11	11	12	11	69
	HT3.neg	12	12	11	11	12	11	69
	C8.pos	12	12	11	11	12	11	69
	C8.neg	12	12	10	11	12	11	68
hiAgd28	HT3.pos	9	11	10	9	10	8	57
	HT3.neg	9	11	10	9	10	8	57
	C8.pos	9	11	10	9	10	8	57
	C8.neg	9	11	10	9	10	8	57
loAgd28	HT3.pos	11	12	12	11	11	12	69
	HT3.neg	11	12	12	11	11	12	69
	C8.pos	11	12	11	11	11	12	68
	C8.neg	11	12	12	11	11	12	69

Table 3.2: Sample allocations of the metabolomics data sets.

sequence. These can be a short section of a gene or other DNA elements that are used to hybridize a cDNA or cRNA sample under high-stringency conditions. Hybridization is usually detected and quantified by detection of fluorophore-labeled, silver-labeled, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the sample. In our experiment, there were three spots per gene of which the medians were calculated for each array. Six biological replicates were run per treatment for the microarrays.

The liver and brain microarray data sets we received had been normalized, background corrected and ready for analysis. The microarray data sets were relatively simple compared to the metabolomics data. A total of 490 genes observed for each set. There were no missing values in the liver microarray data set and just a small amount of missing values in the brain set.

3.2.3 Proteomic Data

Bullfrog tadpole (*Rana catesbeiana*) liver samples from individual tadpoles as part of the frogSCOPE project were sent to the University of Victoria Proteomics Centre for

analysis. These animals have been exposed to thyroid hormone and/or 6ug/L silver for 6 days.

The iTraq technique, which is short for isobaric tags for relative and absolute quantization, was applied to study the quantitative changes in the proteome. iTraq is a non-gel-based technique used to quantify proteins from different sources in a single experiment. It is a relative quantization technique. In each run, one of the sources is used as a reference sample to provide relative quantization between the other sources. The method is based on the covalent labeling of peptides from protein digestions with tags of varying mass. There are currently two commonly used reagents (tags): 4-plex and 8-plex, which can be used to label all peptides (in theory) from different samples. The tagged samples are then pooled, fractionated and analyzed by tandem mass spectrometry.

An 8-plex kit iTraq was used at the Proteomics Centre for the frogSCOPE data. An inter-run standard liver (LIS) sample was used as the reference sample of each run, so each full run of the iTraq facility quantified 7 samples at the same time. A total of 72 samples were sent to the Proteomics Centre for iTraq analysis. It took 11 iTraq runs to finish the quantification of all the 72 samples.

Finally, for each iTraq run, a database search of the iTraq data (.wiff files) was performed to identify the labeled peptides and hence the corresponding proteins. $ProteinPilot^{TM}$ Software was used to identify proteins in each of the mass spectrometry files by searching certain databases. The software identifies proteins based on the spectra they explain, and proteins with similar evidence are organized into"protein groups" where the most likely protein in each group is marked as a winner protein (more detailed description about the $Pro\ Group^{TM}\ algorithm$ can be found in the release notes of $ProteinPilot^{TM}$ Software). Since the database search was done on one iTraq data at a time, there were 11 separate data files generated from the software

after the database searching. Each of the iTraq runs were independent, so the protein groups that came from the $ProteinPilot^{TM}$ were presented slightly differently for each iTraq data set even though the information might be the same (Cohen Freue et al, 2005). Therefore, it was necessary to regroup across all the 11 iTraq runs to unify the groups before further analysis (Cohen Freue et al, 2005). During the regrouping, groups containing bacterial or fungal ID only were removed from the overall group list manually.

Of all the 72 samples, it turned out that one of the samples in treatment NaOH.nano was tail fin tissue rather than liver tissue. That sample was removed for analysis. Therefore, the total sample size of the proteomic data was 71.

Chapter 4

Missing value analysis

Missing value analysis is very important to both proteomics centers and statisticians. Statisticians can find a robust way to analyze the data, while technicians can tune parameters in mass spectrometry.

It is normal for metabolomics data sets to have missing values. In searching for relevant literature regarding this issue, imputation methods were the most frequently used solutions. There were also papers that replaced missing values with zeros (Mueller et al, 2007) and arithmetic means (Denkert et al, 2006). However, for blanks in a treatment group where the other available observations are large, replacing blanks with zeros produces outliers that would affect later analysis. Also, filling in the blanks using imputation methods was not preferred by the technicians in the Proteomics Centre. They suggested that we analyze the data sets as they were, without adding in any observations that were not from the proteomics facility. Their concern was valid. Also, Table 4.1 shows that there were many masses with less than 2 observations within each treatment. Any imputation on such data sets might completely change the data sets.

	number of masses			
	aqu.pos	aqu.neg	org.pos	org.neg
Total	870	3627	1305	3952
at least **1** observation in each treatment	492	2567	594	2603
at least **2** observations in each treatment	357	1992	427	2026
at least **3** observations in each treatment	291	1614	322	1658
at least **4** observations in each treatment	236	1371	254	1365
at least **5** observations in each treatment	206	1173	209	1166
at least **6** observations in each treatment	181	992	160	997
at least **7** observations in each treatment	162	852	130	882
at least **8** observations in each treatment	148	742	107	758
at least **9** observations in each treatment	124	615	85	604
at least **10** observations in each treatment	102	488	50	446
at least **11** observations in each treatment	80	369	NA	301

Table 4.1: General data structure of the FTICR-MS hiAgd6 sets by the number of observations in each treatments. The org.pos set only has 10 samples in the NaOH.nano treatment (Table 3.2), so the number of masses is NA when we use 11 as a cutoff value.

4.1 Homogeneity between the Number of Missing Values and the Treatments

The data set aqu.pos was used as a test data set for the missing value analysis. Frequencies of missing and non-missing values by treatment, together with percentages in each treatment, are given in Table 4.2 through to Table 4.5.

Pearson's chi-square test was conducted for the test of homogeneity. We wanted to determine if there were any differences with respect to missing patterns among 6 treatments.

Null hypothesis: the proportions of missing values are identical through the 6 treatments.

1. Missing value analysis for all the 870 masses in the aqu.pos set (Table 4.2);

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	5529	5448	5925	4682	4568	5563	31715
	Percentage	52.96	52.18	**56.75**	**44.85**	47.73	53.29	51.34
Non-missing	Count	4911	4992	4515	5758	5002	4877	30055
	Percentage	47.04	47.82	43.25	55.15	52.27	46.71	48.66
Sum	Count	10440	10440	10440	10440	9570	10440	61770
	Percentage	100	100	100	100	100	100	100

Table 4.2: P-value from homogeneity test is 1.442e-79. So there is very strong evidence against null hypothesis. Treatments **T3.ionic** and **NaOH.ww** have the lowest and highest missing percentages.

2. Missing value analysis for the 206 masses left after deleting masses with less than 5 observations within any of the 6 treatments in the aqu.pos set (Table 4.3);

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	242	223	361	254	196	278	1554
	Percentage	9.79	9.02	**14.60**	10.28	**8.65**	11.25	10.62
Non-missing	Count	2230	2249	2111	2218	2070	2194	13072
	Percentage	90.21	90.98	85.40	89.72	91.35	88.75	89.38
Sum	Count	2472	2472	2472	2472	2266	2472	14626
	Percentage	100	100	100	100	100	100	100

Table 4.3: P-value=1.027e-11. Treatments **T3.nano** and **NaOH.ww** have the lowest and highest missing percentages. One of the T3.nano samples was removed from the original data set (refer to Section 3.2.1) because of poor data quality. This could be the reason of the low missing percentage in the T3.nano treatment.

3. Missing value analysis for the 124 masses left after deleting masses with less than 9 observations within any of the 6 treatments in the aqu.pos set (Table 4.4);

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	35	50	80	53	27	50	295
	Percentage	2.35	3.36	**5.38**	3.56	1.98	3.36	3.35
Non-missing	Count	1453	1438	1408	1435	1337	1438	8509
	Percentage	97.65	96.64	94.62	96.44	98.02	96.64	96.65
Sum	Count	1488	1488	1488	1488	1364	1488	8804
	Percentage	100	100	100	100	100	100	100

Table 4.4: P-value=7.261e-06. Treatment NaOH.ww has the highest missing percentage among the treatments.

4. Missing value analysis for the 102 masses left after deleting masses with less than 10 observations within any of the 6 treatments in the aqu.pos set (Table 4.5).

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	19	26	32	29	11	20	137
	Percentage	1.55	2.12	**2.61**	2.37	0.98	1.63	1.89
Non-missing	Count	1205	1198	1192	1195	1111	1204	7105
	Percentage	98.45	97.88	97.39	97.63	99.02	98.37	98.11
Sum	Count	1224	1224	1224	1224	1122	1224	7242
	Percentage	100	100	100	100	100	100	100

Table 4.5: P-value=0.04194. Although there are only 102 masses left in this data set, the high percentages of non-missingness give us confidence in the analysis results based this subset. Treatment NaOH.ww has the highest missing percentage among the treatments.

The missing and non-missing tables for the other FTICR-MS data sets (aqu.neg, org.pos, org.neg) are in Appendix A.2. The tables show that treatment NaOH.ww always had a high percentage of missing values in aqueous sets, while there were no specific patterns in the organic sets. This feature should be examined carefully.

4.2 Missing Value Analysis

There were many missing values in the FTICR-MS data sets, and it was difficult to say whether they were technically missing values or biologically missing values. On the one hand, if they were missing because of failure to detect them during the mass spectrometry process, it was reasonable to treat them as NAs. In this case, we needed to delete masses with too many missing values within any of the treatment combinations in order to obtain reliable analysis results. On the other hand, if the missing values were missing because they did not exist in some of the treatment combinations, those masses were actually the ones we were interested in. Deleting masses with no or few observed values would filter out many important and interesting masses. In this case, we replaced the missing values with zeroes and kept all the masses for later analysis. In frogSCOPE, a cutoff value $S/N = 3$ was used at the Proteomics Centre in the FTICR-MS experiment. Missing observations could happen when the ion peaks on mass spectra with S/N's below 3. In this case, a blank could be a technically missing value because of low peak intensities, lots of noise, or overlapping peaks, while it was also possible that there were no peaks at all (biologically missing values). How we treat the missing values was critical to the entire analysis.

4.2.1 Biologically Missing Values

Biologically missing values are more important and interesting than technically missing values since they are a clear sign of differential expression among treatment combinations. Tadpoles and frogs are different. A metabolite may not be expressed under one treatment and expressed under others when metamorphosis of tadpoles into frogs is affected by one of the treatments. If we omitted that metabolite because of missing

values, we would not detect this differential expression.

Logistic regression models were used to determine if detectable and undetectable features differed among treatment combinations. The masses were treated as factors at two levels, 0 and 1, where each missing observation was denoted by 0 and non-missing observation was denoted by 1. A family-wise error rate $\alpha = 0.05$ was used in the data analysis and Bonferroni criterion α/M was used to get the test level for each comparison with M comparisons in total. Usually, Shapiro-Wilk tests and Anderson-Darling test can be used to test for residual normality test, but there are cases when the residuals are identical since there are lots of zeros in the data set. Also, residual plots are not too useful for binary data. So we didn't perform normality tests here. We used boxplots and interaction plots to help us visualize the data structure after logistic regression analysis. Further selection was done based on the plots.

294 masses had significant detectable and undetectable features among treatments by logistic regression at significance level $0.05/870 = 5.747126e-05$. In the aqu.pos set, there were 177 masses which had at least 10 observations in one group and zero observations in at least one other group. The logistic regression picked all the 177 masses but mass "344.2328". The p-value of this mass from logistic regression was 0.0001167. Although it was larger than the Bonferroni criterion, the p-value itself was far smaller than the significance level we usually use. To summarize, the logistic regression results were consistent with the missing data features, and the results from logistic regression were reliable.

Interaction plots of proportions of non-missing values were used to illustrate the missing and non-missing features among groups. The approximate 95% confidence interval was calculated by

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}, \tag{4.1}$$

where n was the sample size of each treatment.

Interaction plots of proportions of some of the masses picked by logistic regression were plotted in Figure 4.1 to illustrate the idea.

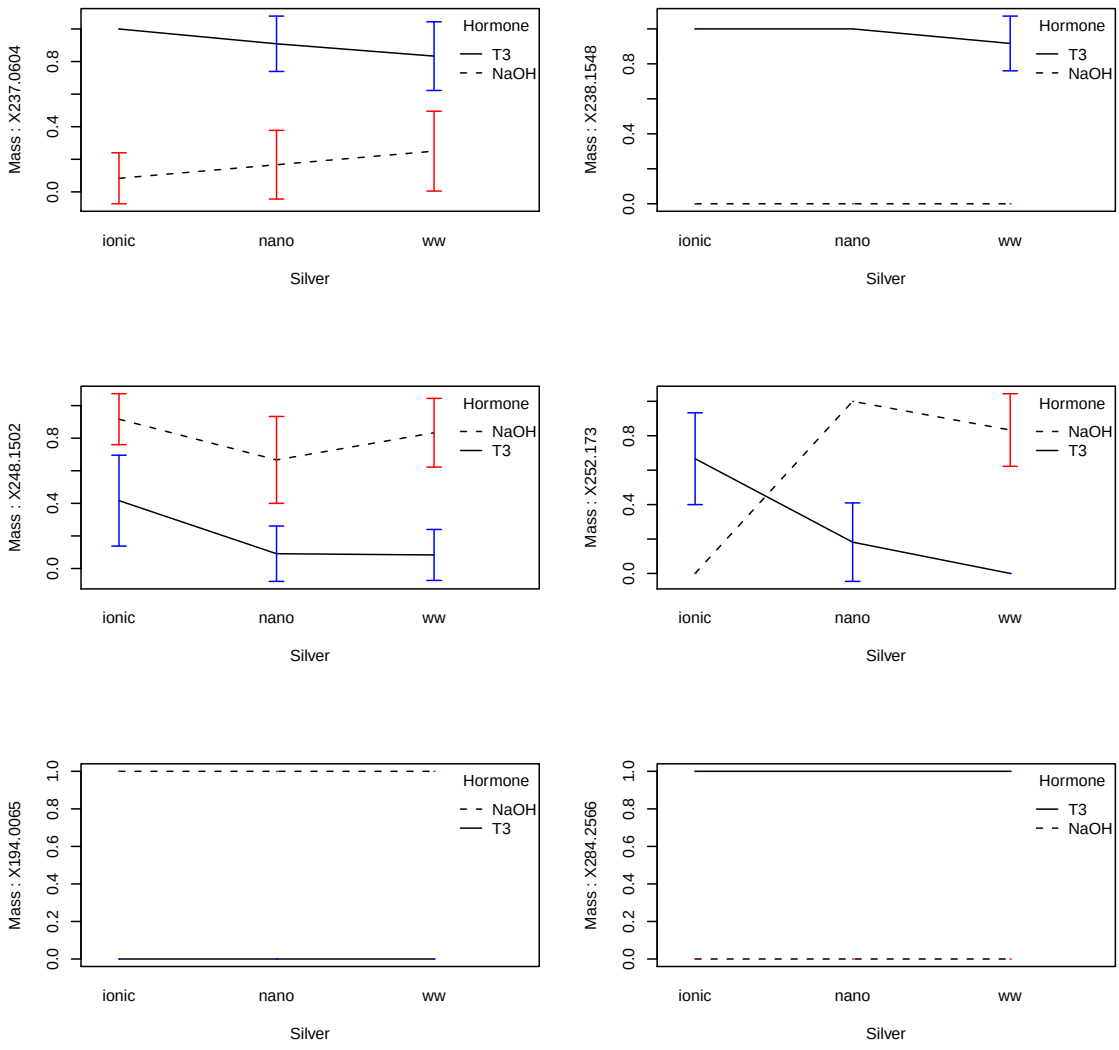

Figure 4.1: Interaction plots of proportions of non-missing observations for significant masses by logistic regression. The detectable and undetectable features of these masses differ among the treatments.

To assist in the interpretation of the interaction plots, Table 4.6 lists the number of non-missing observations within each treatment for these masses, and boxplots of

these masses are in Figure 4.2.

	NaOH.ionic(12)	NaOH.nano(12)	NaOH.ww(12)	T3.ionic(12)	T3.nano(11)	T3.ww(12)
Mass:237.0604	1	2	3	12	10	10
Mass:238.1548	0	0	0	12	11	11
Mass:248.1502	11	8	10	5	1	1
Mass:252.173	0	12	10	8	2	0
Mass:194.0065	12	12	12	0	0	0
Mass:284.2566	0	0	0	12	11	12

Table 4.6: Numbers of non-missing values within each treatment for masses from aqu.pos set. The number in parentheses is the sample size in that treatment.

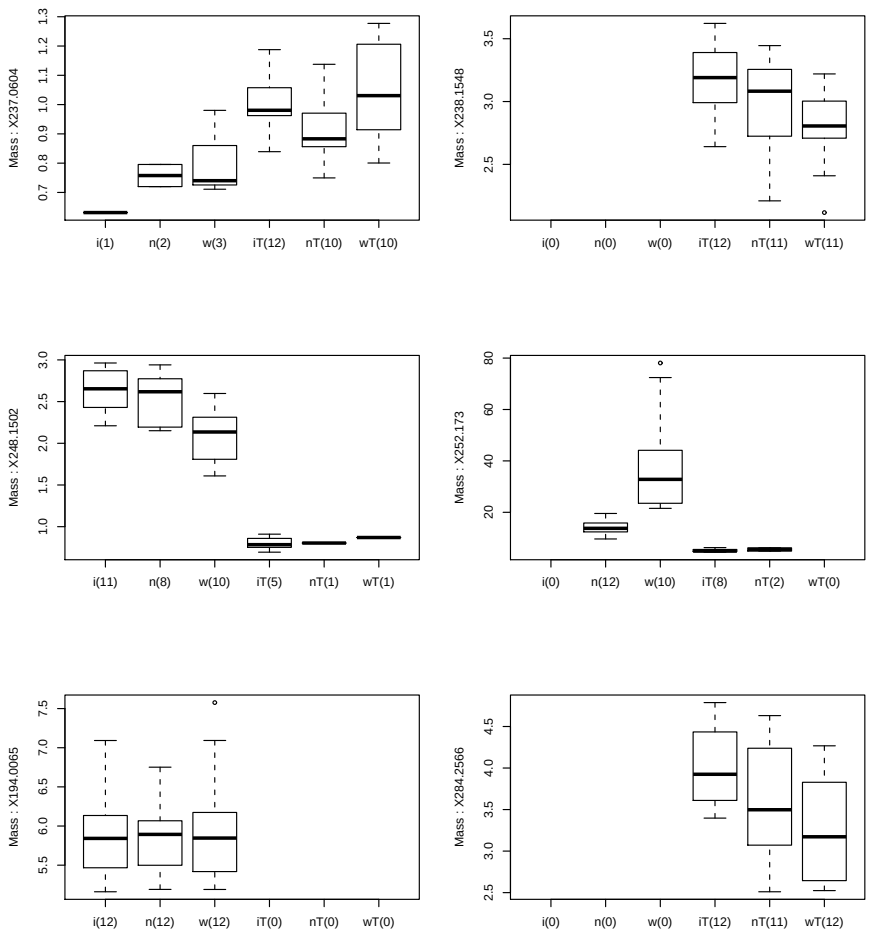

Figure 4.2: Boxplots of masses listed in Table 4.6.

Table 4.7 lists the numbers of masses with significant missing and non-missing features among treatments (i.e., frequencies of missing masses significantly associated with treatments) for each of the FTICR-MS sets by two-factor (missing and non-missing) factorial logistic regression model. Results for interesting treatment contrasts amongst the significant 5 degrees of freedom overall test metabolites are also included. The significance level for each one degree of freedom contrast was 0.01. The percentages are for the number of significant contrasts out of the total number of masses which were significant in the 5 degrees of freedom test of no overall treatment effect, e.g., 13/294=4.42%.

	Test of treatment effect (df5)	NaOH.ww vs T3.ww (df1)	NaOH.ww vs NaOH.ionic (df1)	NaOH.ww vs NaOH.nano (df1)	T3.ww vs T3.ionic (df1)	T3.ww vs T3.nano (df1)	NaOH.nano vs T3.ww (df1)
Aqu Pos (870 masses)	294 ($p < 5.75e-05$)	13 (4.42%)	5 (1.70%)	3 (1.02%)	55 (18.71%)	14 (4.76%)	16 (5.44%)
Aqu Neg (3627 masses)	526 ($p < 1.38e-05$)	29 (5.51%)	24 (4.56%)	16 (3.04%)	19 (3.61%)	9 (1.71%)	24 (4.56%)
Org pos (1305 masses)	641 ($p < 3.83e-05$)	45 (7.02%)	22 (3.43%)	4 (0.62%)	19 (2.96%)	2 (0.31%)	44 (6.86%)
Org Neg (3952 masses)	588 ($p < 1.27e-05$)	12 (2.04%)	40 (6.80%)	13 (2.21%)	50 (8.50%)	39 (6.63%)	26 (4.42%)

Table 4.7: Results from logistic regression missing value analysis of Metabolomic FTICR-MS data.

4.2.2 Technically Missing Values

We analyzed the data sets based on observed data from proteomic facility. In order to conduct reliable statistical analysis, we needed to delete masses with too many missing values within any of the treatment combinations. For different cutoff points, Levene's test [Levene, 1960] was used for the equal variances test, and Kruskal-Wallis rank sum test [Corder and Foreman, 2009] was conducted on the masses that passed the equal variances test to determine the masses that had significantly different expressions among treatment. Of all the 870 masses in the aqu.pos set, 787 passed the equal variances test at the 0.05 significance level. We then performed Kruskal-Wallis rank sum test [Corder and Foreman, 2009] to determine which masses differed among the 6 treatments combinations. 89 masses were picked by this test at significance level $0.05/787 = 6.35324e-05$, i.e., 89 masses differed among treatments by Kruskal-Wallis rank sum test at this significance level. The table below lists the number of masses picked at different cutoff points.

	number of masses			$\frac{\text{Kruskal-Wallis test}}{\text{after filtering}}$ %
	after filtering	Levene's test $s.l. = 0.05$	Kruskal-Wallis test $s.l. = 0.05/M$	
Total	870	787	89	71.25%
at least **1** observation in each treatment	492	439	88	17.89%
at least **2** observations in each treatment	357	316	86	24.09%
at least **3** observations in each treatment	291	253	85	29.21%
at least **4** observations in each treatment	236	201	84	35.59%
at least **5** observations in each treatment	206	173	81	39.32%
at least **6** observations in each treatment	181	151	77	42.54%
at least **7** observations in each treatment	162	136	74	45.68%
at least **8** observations in each treatment	148	123	72	48.65%
at least **9** observations in each treatment	124	104	63	50.81%
at least **10** observations in each treatment	102	86	57	55.88%
at least **11** observations in each treatment	80	68	57	71.25%

Table 4.8: The number of masses picked at different cutoff points by Kruskal-Wallis rank sum test for the aqu.pos set.

Significance level $\alpha = 0.05$ was used in Levene's test instead of the Bonferroni criterion to carry out a stricter test. The same analysis table for the remaining three

FTICR-MS data set are in Appendix A.3. From the four tables, we noted that when we only considered masses with at least 10 observations in each treatment, there were still a moderate number of masses picked for further analysis. Furthermore, those masses had relatively more observations in each treatment, and the analysis results based on those observations were more reliable. Therefore, a 10 cutoff point was used in our analysis for the FTICR-MS data sets later.

Chapter 5

Simulation Study of FTICR-MS Metabolomics Data

The application of FTICR-MS to analyze metabolomics data was new (Han et al, 2008), and it is more sensitive than other standard MS techniques. However, there were usually many missing values in the FTICR-MS data sets. This feature limited the application of standard statistical analysis methods. In this chapter, we simulate FTICR-MS data based on the data sets from the frogSCOPE project. Different experiments are then conducted to compare imputation methods and feature selection algorithms.

5.1 Simulation

We had four sets of FTICR-MS data sets from the frogSCOPE project (Section 3.2.1). Aqu.pos, Aqu.neg, Org.pos and Org.neg, where Aqu and Org standed for the extraction methods Aqueous and Organic respectively, and pos and neg standed for positive and negative ion modes respectively. In each data set, the normalized peak intensities of neutral (monoisotopic) masses were recorded for each sample.

In the FTICR-MS frogSCOPE data sets, there were biological replicates in each treatment (Table 3.2). Our study was based on medians of the replicates, that is, in each data set, the intensity of a mass in a treatment was represented by the median of the replicates for that treatment. The aqu.pos FTICR-MS set was chosen as the study set. First, histograms by treatment were used to examine the distribution of the masses in this set. The shape of frequency distribution for mass intensities was not well displayed on the histograms in Figure 5.1 because of the large range of the x-axis. Masses 414.204 and 436.1861 had extremely high intensities compared to the other masses, and they could be contaminated. These two masses were removed, and histograms were plotted again for the aqu.pos study set. In Figure 5.2, we see that without those two masses the range of the x-axis shrinks considerably.

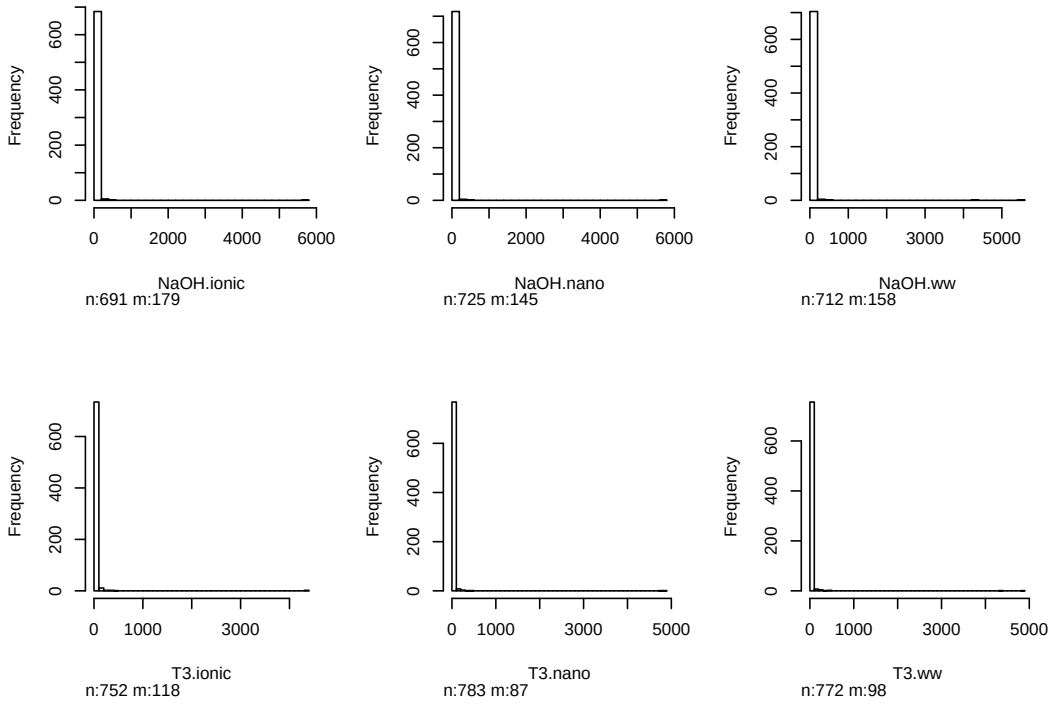

Figure 5.1: Histograms of intensities of the aqu.pos study data by treatment. The numbers of missing and non-missing observations are shown by default in R statistical software.

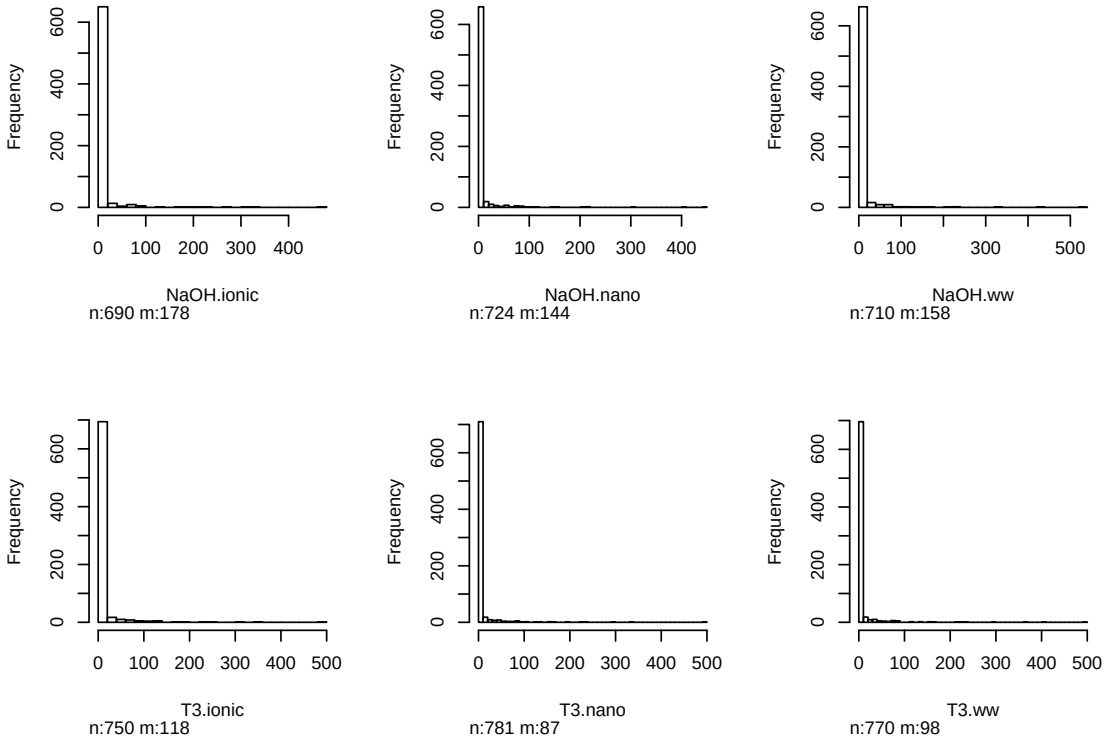

Figure 5.2: Histograms of intensities of the aqu.pos study data by treatment after two extreme outliers, 414.204 and 436.1861, are removed from the data. The numbers of missing and non-missing observations are shown by default in R statistical software.

To further examine the distribution of the intensities in each treatment, the aqu.pos set was separated into three sets for each treatment. Figure 5.3 and Figure 5.4, histograms of the study set split by the cutoffs given in Table 5.1, show that the low intensity sets were approximately log-normally distributed, and uniform distributions were appropriate for the moderate/high intensity sets considering the small sample sizes. The maximum likelihood estimates (MLEs) of the parameters were calculated and listed in Table 5.1. Intensity 100 was always used to separate the moderate intensity set and the high intensity set. Different cutoff points were used for the low intensity set as given in Table 5.1. Those values were picked by a crude approximation of the histograms, followed by Chi-square goodness-of-fit tests on the fit of the low intensity sets to adjust the cutoff points.

	Range	Number of masses	Percentage	MLE
NaOH.ionic	[0, 17]	647	93.768%	LogN($\mu = 0.524, \sigma = 0.731$)
	(17, 100)	34	4.928%	U(a=17.026, b=94.718)
	[100, -]	9	1.304%	U(a=137.362, b=478.251)
NaOH.nano	[0, 17]	677	93.508%	LogN($\mu = 0.514, \sigma = 0.731$)
	(17, 100)	37	5.110%	U(a=21.223, b=98.165)
	[100, -]	10	1.381%	U(a=104.740, b=444.917)
NaOH.ww	[0, 16]	659	92.817%	LogN($\mu = 0.527, \sigma = 0.690$)
	(16, 100)	40	5.634%	U(a=16.277, b=85.922)
	[100, -]	11	1.549%	U(a=111.051, b=524.556)
T3.ionic	[0, 19]	693	92.400%	LogN($\mu = 0.709, \sigma = 0.748$)
	(19, 100)	41	5.467%	U(a=19.290, b=89.315)
	[100, -]	16	2.133%	U(a=101.532, b=499.794)
T3.nano	[0, 18]	725	92.830%	LogN($\mu = 0.619, \sigma = 0.700$)
	(18, 100)	43	5.506%	U(a=18.390, b=96.565)
	[100, -]	13	1.665%	U(a=100.031, b=498.246)
T3.ww	[0, 14]	708	91.948%	LogN($\mu = 0.566, \sigma = 0.680$)
	(14, 100)	48	6.234%	U(a=14.493, b=86.718)
	[100, -]	14	1.818%	U(a=110.451, b=490.557)

Table 5.1: The study data was examined by treatment. For each treatment, cutoff points and numbers of masses as well as the percentages are given. MLEs are also listed in the last column.

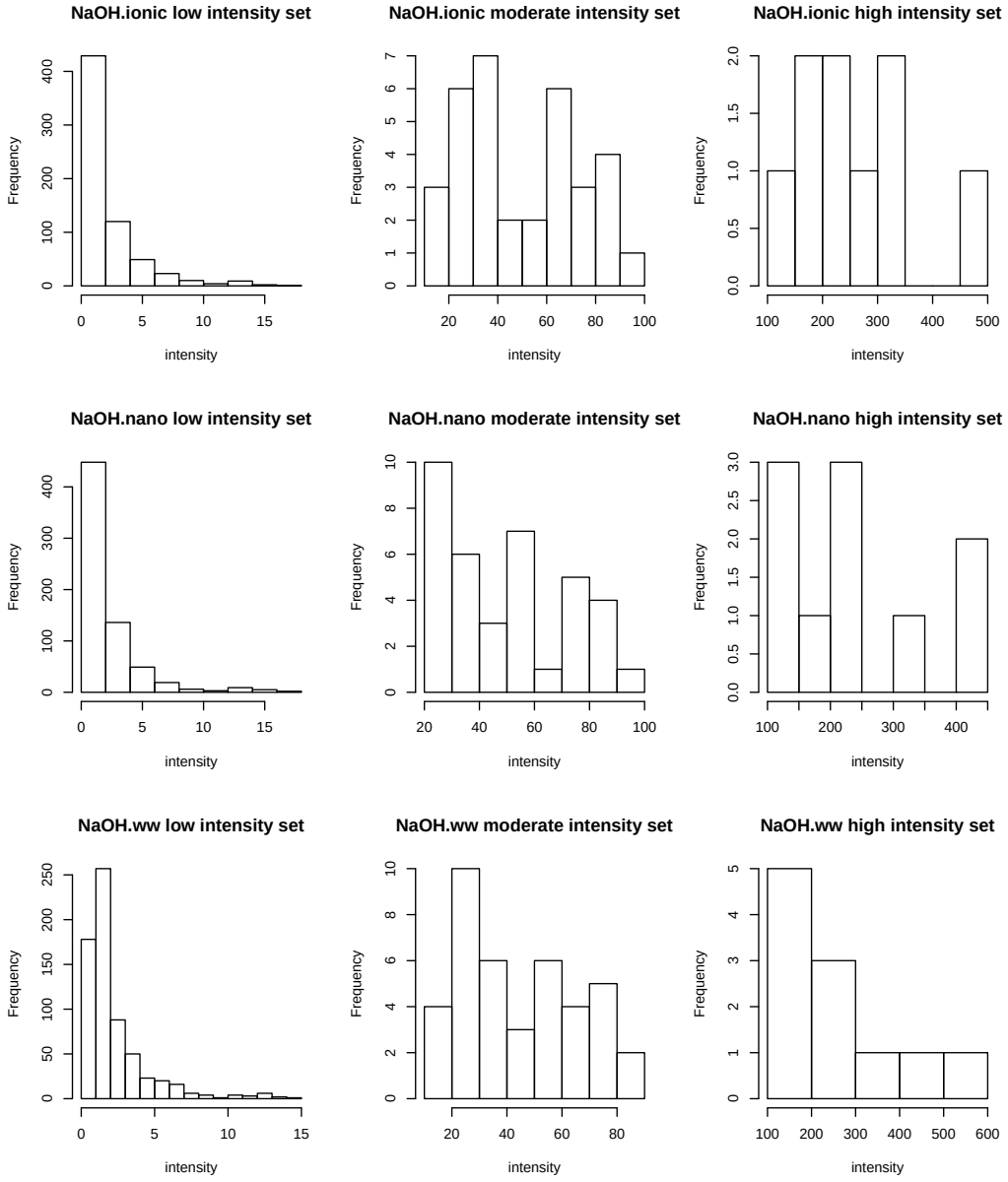

Figure 5.3: Histograms of the aqu.pos study data by intensities for the NaOH treatments.

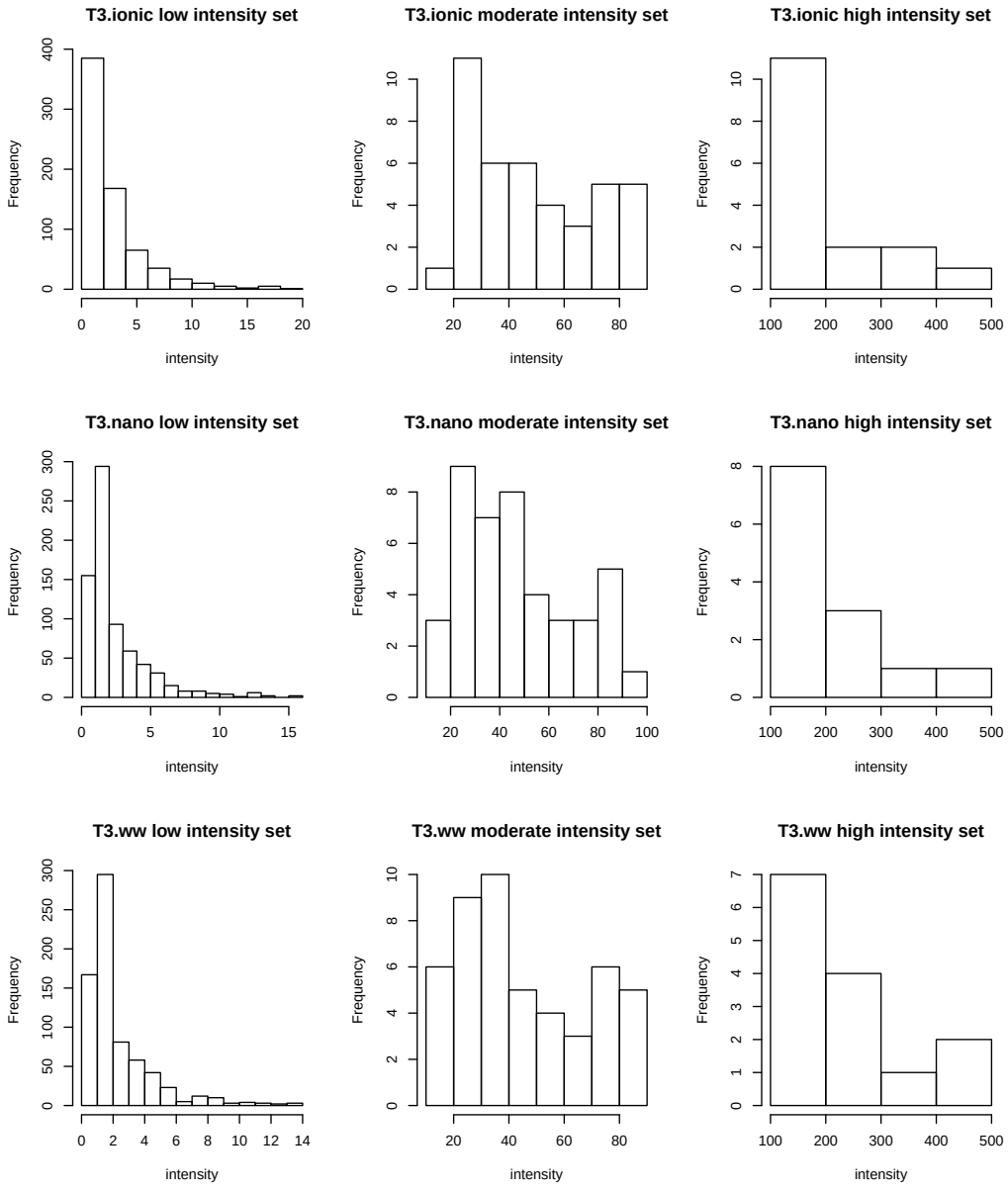

Figure 5.4: Histograms of the aqu.pos study data by intensities for the T3 treatments.

The Chi-square goodness-of-fit tests of the Log-normal distributions resulted in small p-values, but graphis in Figure 5.5 give us confidence in the fitted models. The curves are the fitted curves based on the MLEs given in Table 5.1.

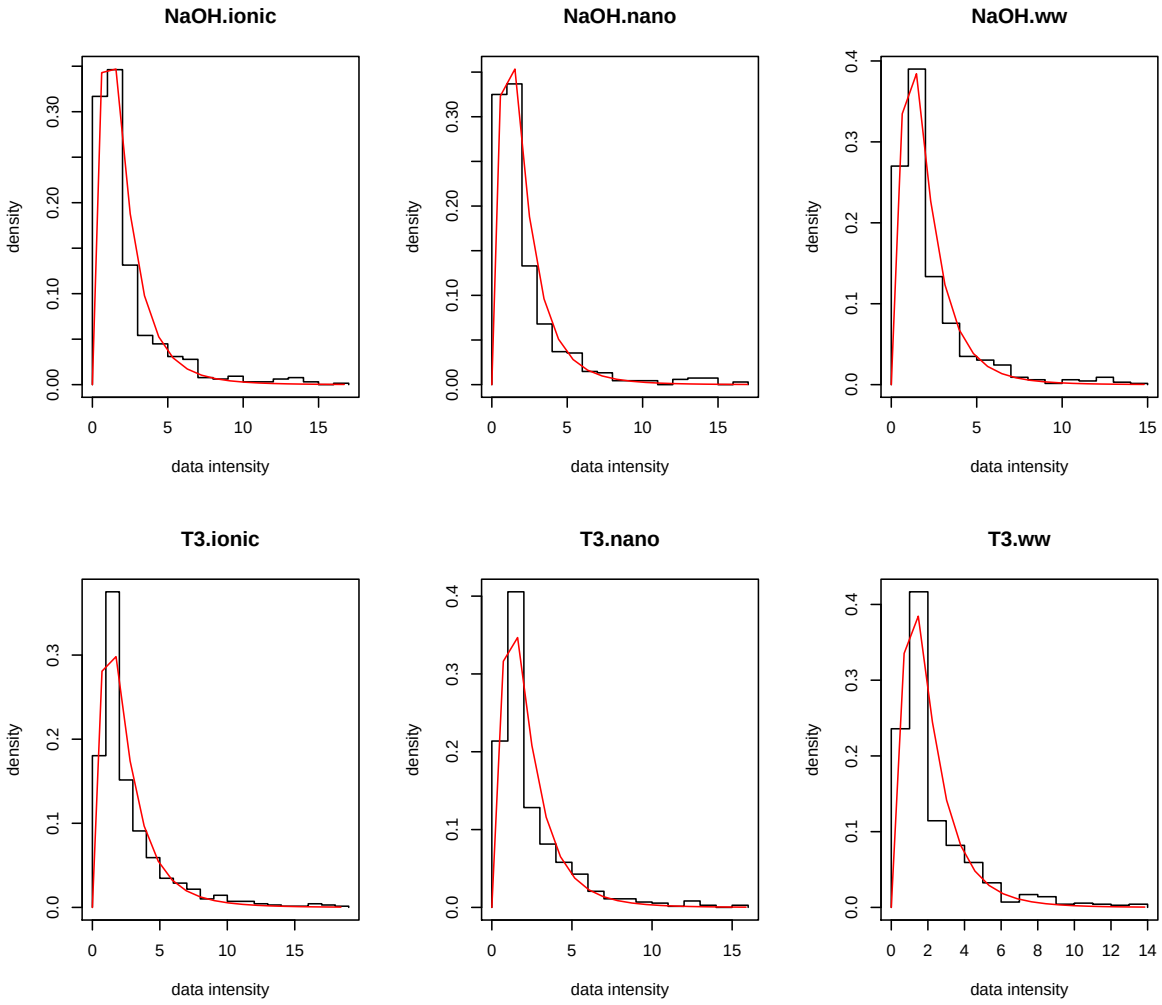

Figure 5.5: Density histograms of the low intensity sets and the fitted curves.

The simulation of differentially expressed (DE) masses was based on the structure set up by Mu (2008) in her Master thesis. Masses were randomly selected to be DE masses, where some of them were up-regulated and others were down-regulated. When simulating more than 2 treatments (including the control), differential expressions could happen in one treatment or multiple treatments. Let P_{DE} be the pro-

portion of DE masses, P_{MDE} be the proportion of multiply differentially expressed masses among the DE masses, and P_{DRE} be the proportion of down-regulated masses among the DE masses. DE masses were picked randomly and proportionally using these parameters. The values in the control treatment, NaOH.ww, were as a starting value for the generation of DE effects. We added or subtracted a percentage (α) of the control values to or from the corresponding control values to generate the DE masses. For a given α, if the control values are small and the variance among replicates is large, the DE regulation after adding or subtracting a percentage α may not be significant. A tuning parameter was added to complete the regulation. The formula was as follows where the DE percentage α and the tuning parameters were fixed for all the DE masses:

$$\begin{cases} \text{up-regulated:} & \text{Intensity}_{DE} = \text{Intensity}_{Control} \times (1+\alpha) + \text{tune.para;} \\ \text{down-regulated:} & \text{Intensity}_{DE} = \text{Intensity}_{Control} \times (1-\alpha) - \text{tune.para.} \end{cases} \tag{5.1}$$

To simulate replicates for each treatment, analysis was carried out to examine the consistency within replicates in the FrogSCOPE data aqu.pos set. Table 5.2 lists median variances among replicates for the aqu.pos sets by treatment. Variances were also calculated for the data set with at least 10 non-missing values within each treatment. The median variances in the 10 non-missing set were larger than the full set because there were more observations within each treatment.

	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww
aqu.pos (all masses)	0.074	0.221	0.095	0.194	0.207	0.200
aqu.pos (10 cutoffs)	0.206	0.448	0.423	0.727	0.549	0.589

Table 5.2: The median variances between replicates based on the FTICR-MS frogSCOPE data aqu.pos set.

Further analysis based on the aqu.pos 10 cutoff set was done to investigate the

relation between intensities and variances. In Figure 5.6, the first plot is the plot of all the 102 masses, where intensity is the median intensity of each mass in the NaOH.ww group and variance is the corresponding variance among replicates. There are extreme outliers in this plot. Subsequent plots reduce the vertical and horizontal scales, filtering out large variance masses and zooming in on low median intensity masses. There is no clear pattern between intensities and variances.

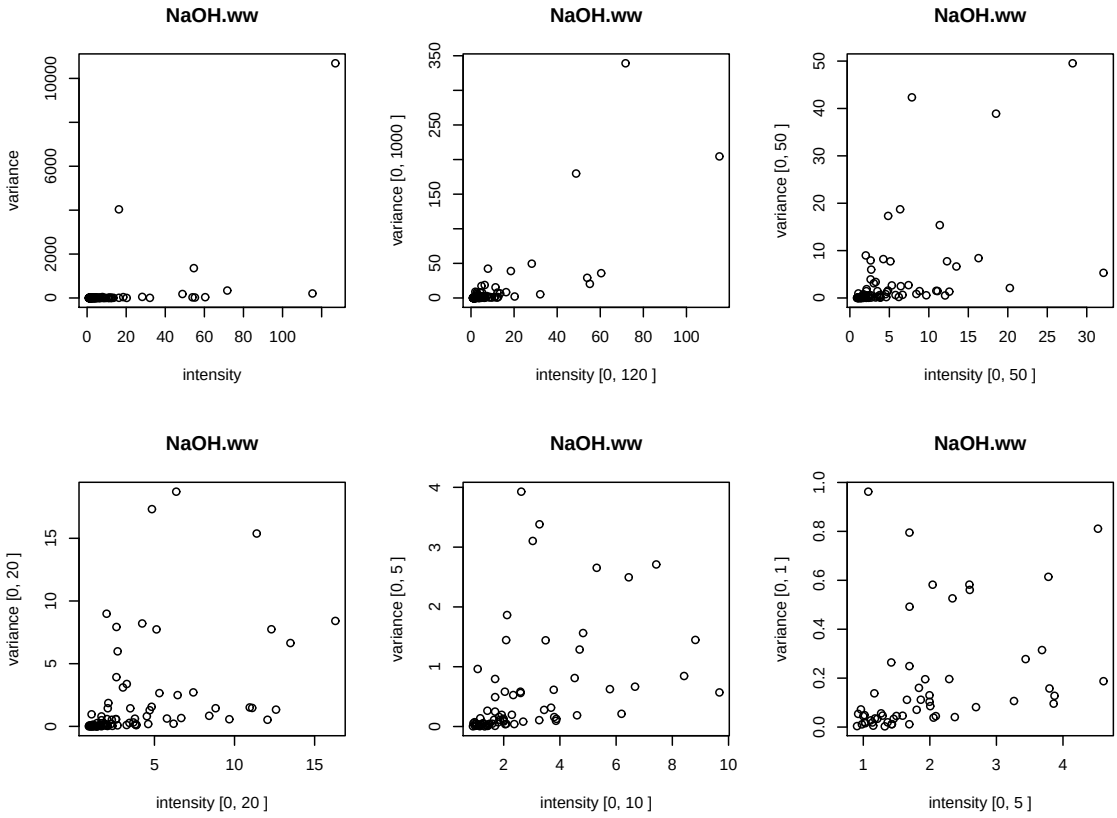

Figure 5.6: Plots of variances against median intensities among replicates of each mass.

To get the distribution of variances from the study data, subsequent histograms of variances were plotted, zooming in to small variances (Figure 5.7).

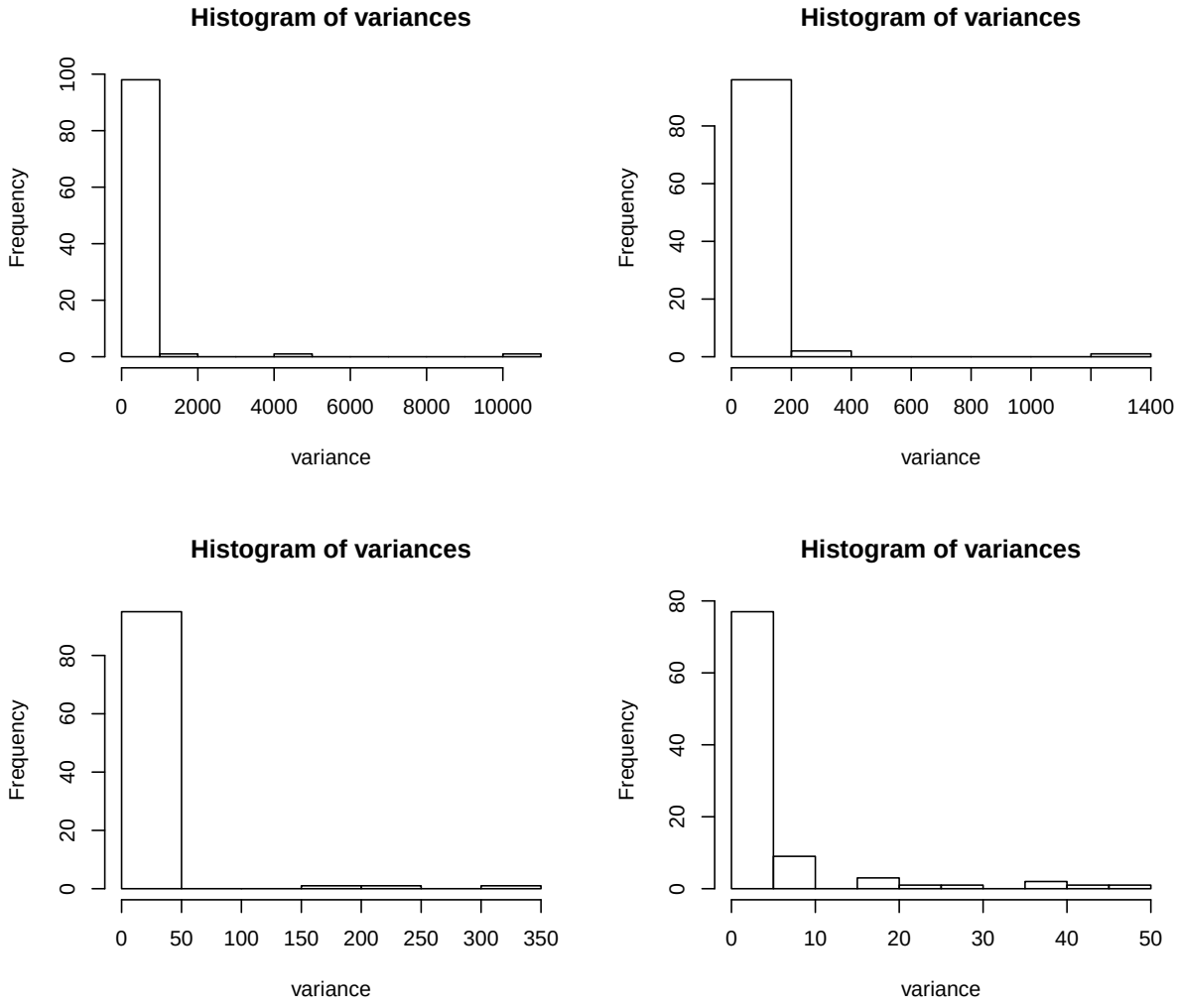

Figure 5.7: Histograms of variances.

Data with small variation between replicates is considered reliable and consistent. Histograms in Figure 5.7 show that there is no obvious outliers in range from 0 to 50.

Range	(0, 5]	(5, 10]	(10, 15]	(15, 20]	(20, 25]	(25, 30]	(30, 35]	(35, 40]	(40, 45]	(45, 50]
Count	77	9	0	3	1	1	0	2	1	1
	77 (81.05%)	18 (18.95%)								

Table 5.3: The number of counts inside each cell in the last histogram in Figure 5.7.

Based on the above study results (Figure 5.7 and Table 5.3), variances among replicates were sampled from $\mathrm{U}(0.01, 5)$ and $\mathrm{U}(5, 50)$ with model:

$$81.05\% * \mathrm{U}(0.01, 5) + 18.95\% * \mathrm{U}(5, 50). \tag{5.2}$$

To summarize, a data set with 1,000 masses, 6 treatments and 12 replicates each was generated in five steps:

1. Firstly, we simulated the control. A matrix of size 1000×1 was simulated based on the distribution of the NaOH.ww group (which is the control in FrogSCOPE experimental design) given in Table 5.1. The simulation model was:

   $$92.82\% * \mathrm{LogN}(0.527, 0.690) + 5.63\% * \mathrm{U}(16.277, 85.922) + 1.55\% * \mathrm{U}(111.051, 524.556)$$

2. Table 5.1 and Figure 5.3 - 5.4 show that all the treatments had approximately the same distribution. So the other 5 treatments were simulated from normal distribution with the corresponding control intensity as mean and 0.01 as standard deviation. After simulation, a 1000×6 matrix was generated, where each column represents one treatment.

3. DE masses were randomly selected and regulated before generating replicates.

In the simulation of DE masses, fixed parameters $P_{DE} = 0.01$, $P_{MDE} = 0.2$ and $P_{DRE} = 0.5$ were applied. When a mass was selected to be regulated in more than one treatment, the number of treatments containing DE masses was 3. The scale of DE masses were as defined in Eq 5.1. The regulation percentage parameter α and the tuning parameter were 2 and 1 respectively.

4. After the simulation of DE masses, generating replicates was the last step. Replicates were simulated from normal distributions with the simulated values as means. Variances were sampled according to Eq 5.2.

5. There were negative values in the simulated data set which is not appropriate for intensity data. The whole data set was shifted from negative to positive by adding the absolute value of the min value generated.

For a given regulation parameter α, expressions of a randomly selected DE component after Eq. 5.1 might not act very differently among treatments when the variation within each treatment was very large. In figure 5.8, a large regulation parameter $\alpha = 15$ is used to demonstrate the simulation result of the DE components.

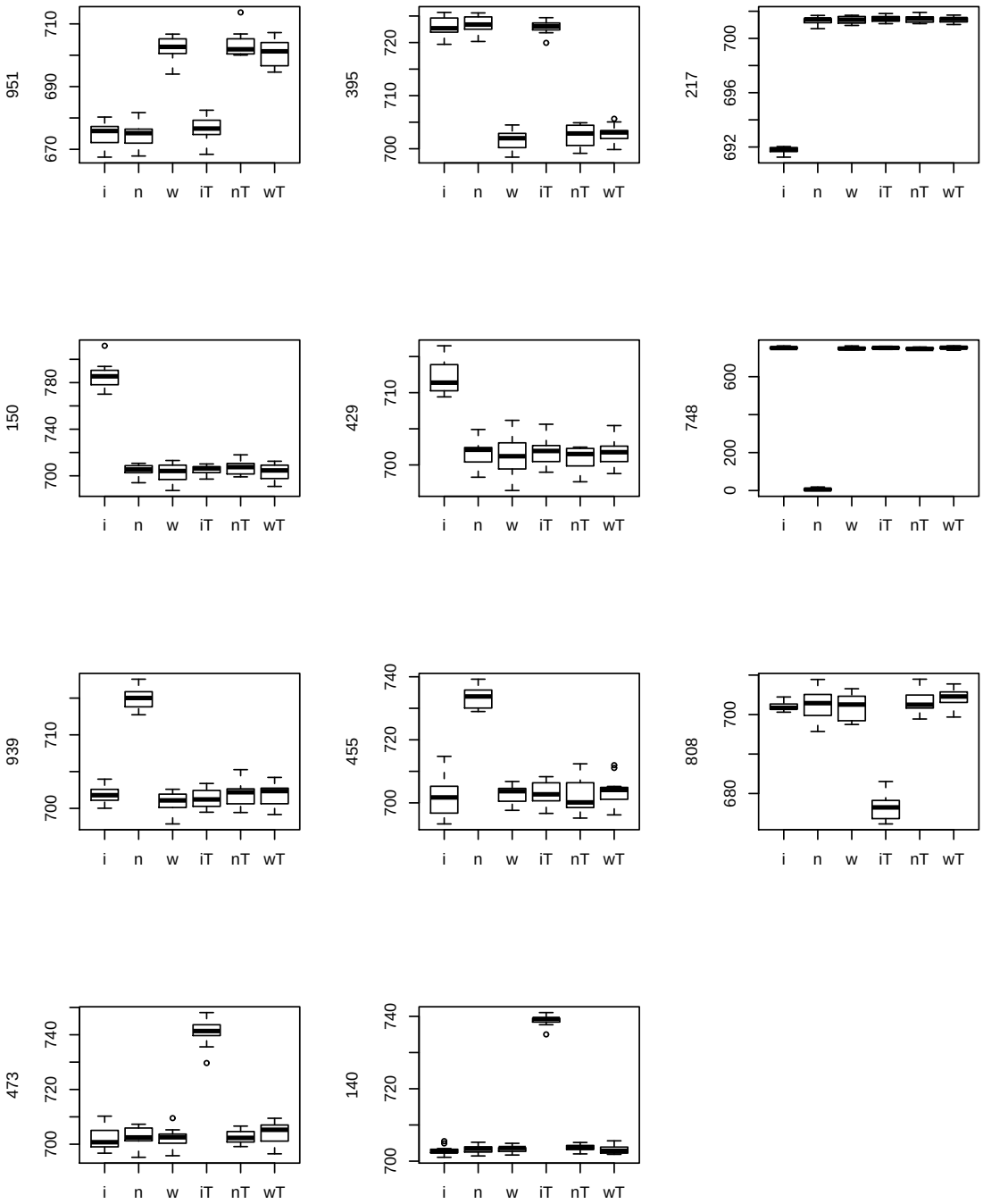

Figure 5.8: Boxplots of the 10 DE components of a 1000×72 simulated set when $P_{DE} = 0.01$, $P_{MDE} = 0.2$ and $P_{DRE} = 0.5$.

5.2 Comparing Different Selection Methods

For hypothesis testing, the increase in type I error occurs when statistical tests are used repeatedly. Multiple testing procedures (MTP) (Pollard et al, 2005) in package multtest from Bioconductor is designed to control a broad class of Type 1 error rates. MTP is a function to perform resampling-based multiple hypothesis testing. Single-step and step-down procedures with combination of minima of unadjusted p-values (minP) and maxima of test statistics (maxT) methods are used to control the chosen type I error rate (the family-wise error rate (FWER), the generalized family-wise error rate (gFWER), the tail probabilities for the proportion of false positives (TPPFP), or the false discovery rate (FDR)).

Another package from Bioconductor is Limma. Limma (Smyth, 2004) is designed to analyze gene expression microarray data, particularly, the use of linear models for analyzing designed experiments and the assessment of differential expression. Although Limma is created for analyzing microarray data, the ability of this package to analyze comparisons between many RNA targets simultaneously is very handy. We include the Limma functions here to assess their stability and reliability on the FTICR-MS data.

There are other packages from Bioconductor which are also set up for identification of differentially expressed genes simultaneously. For example, the HEM package (Cho and Lee, 2005) fits an error model with heterogeneous experimental and/or biological error variances for analyzing microarray data which enables identification of DE genes by simultaneously estimating a large number of heterogeneous variance parameters, while the DEDS (Xiao and Yang, 2007) package ranks genes in evidence of differential expression via distance synthesis. Those functions were not widely used, therefore we didn't include them in the study.

A non-parametric method, the Kruskal-Wallis test (Section 2.4), is also included

in the study. Bonferroni correction (Section 2.1) is applied to control the type I error rate, and Levene's test (Section 2.3) is conducted to assess the equality of variances among treatments at significance level 0.05. From the way the data was simulated, most of the masses passed Levene's test.

The data sets from the frogSCOPE project are high throughput data with thousands of measurements per sample. When choosing a selection method, we prefer a strict one with low false positive rate to get a more concise and accurate list of DE components. To test the performance of the selection methods, especially the false positive rate, we simulated 9 data sets following the steps as described in the previous section, where we kept all the parameters as defined at the end of the previous section except the regulation percentage parameter α. Note that P_{DE} was 0.01, so there are 10 true DE masses in each simulation.

	Kruskal-Wallis test		MTP		Limma	
	DE selected	true DE selected	DE selected	true DE selected	DE selected	true DE selected
$\alpha = 50\%$	3	3	6	5	3	3
	1	1	2	2	2	1
	0	0	3	1	0	0
$\alpha = 100\%$	3	3	6	4	5	4
	2	2	3	3	3	3
	3	3	6	5	4	3
$\alpha = 200\%$	6	6	9	7	5	5
	5	4	7	6	5	5
	3	3	6	4	3	3

Table 5.4: Comparison of three selection methods at different regulation percentages α.

Table 5.4 shows that Kruskal-Wallis test had the highest true positive rate among the three selection methods. MTP was more sensitive to variation among treatments, and that was why it had a relatively higher false positive rate. The performance of Limma was between Kruskal-Wallis test and MTP regarding to the false positive rate and sensitivity to differently expressed components.

5.3 Comparing Different Imputation Methods

An important feature of the DIFT-MS data was that there were usually many missing values. Tables in Section 4.1 and Appendix A.2 confirm that the missing value percentage was often as high as 50%.

To simulate a data set with missing values, we first simulated a 1000×72 data set using the procedure given in Section 5.1. Half of the elements were randomly sampled and replaced with NAs.

The Bayesian Principle Component Analysis (BPCA) missing value estimator uses an expectation maximization approach together with a Bayesian model to approximate the principal axes (eigenvectors of the covariance matrix in PCA) (Oba et al, 2003). The imputation is done iteratively. The algorithm stops if either the maximum number of iterations is reached or if the estimated increase in precision falls below e^{-4} (Oba et al, 2003). The Singular Value Decomposition (SVD) impute algorithm was proposed by Troyanskaya et al (2001). It starts by replacing the missing values with zeros. Then it calculates the mean for each of the rows, and replaces the missing values with the row means. SVD is then performed on the newly formed set to replace row means with its output. The algorithm works iteratively until the changes in the estimated solutions fall below a certain threshold. Missing value estimation using local least squares (LLS) was first described by Kim et al (2005). The general idea of this algorithm is selecting k variables (masses) which have the highest correlation with the target protein by one of Pearson, Spearman or Kendall correlation coefficients. Then missing values are imputed by a linear combination of the k selected variables, where LLS regression is conducted to find the optimal combination.

In this experiment, we tested the performance of three different imputation methods when the missing value percentage was 50%. The Kruskal-Wallis test was conducted to select DE components.

			DE selected	true DE selected
Experiment 1	Before simulating missing values		5	4
	After randomly taking out 50% values	BPCA	2	2
		SVD	1	1
		LLS (k=5)	1	1
		LLS (k=10)	1	1
Experiment 2	Before simulating missing values		4	4
	After randomly taking out 50% values	BPCA	2	2
		SVD	1	1
		LLS (k=5)	2	2
		LLS (k=10)	2	2
Experiment 3	Before simulating missing values		6	6
	After randomly taking out 50% values	BPCA	3	3
		SVD	3	3
		LLS (k=5)	3	3
		LLS (k=10)	3	3

Table 5.5: Test of performance of three different imputation methods.

Table 5.5 shows that variation among treatments could disappear after imputation. We also noticed that all the masses identified as DE masses after imputation were true DE masses, i.e., there were no new DE masses generated after imputation. From this perspective, the use of imputation method is secure. For analysis methods which don't allow missing value, say MTP and PCA, imputation method could be applied before analysis. To illustrate what happened after imputation, boxplots of two DE masses before and after missing value simulation are given in Figure 5.9. By Kruskal-Wallis test, mass 854 is still differentially expressed after imputation methods, but mass 996 is no longer different among treatments. Figure 5.9 gives us some idea about why DE patterns disappeared after imputation.

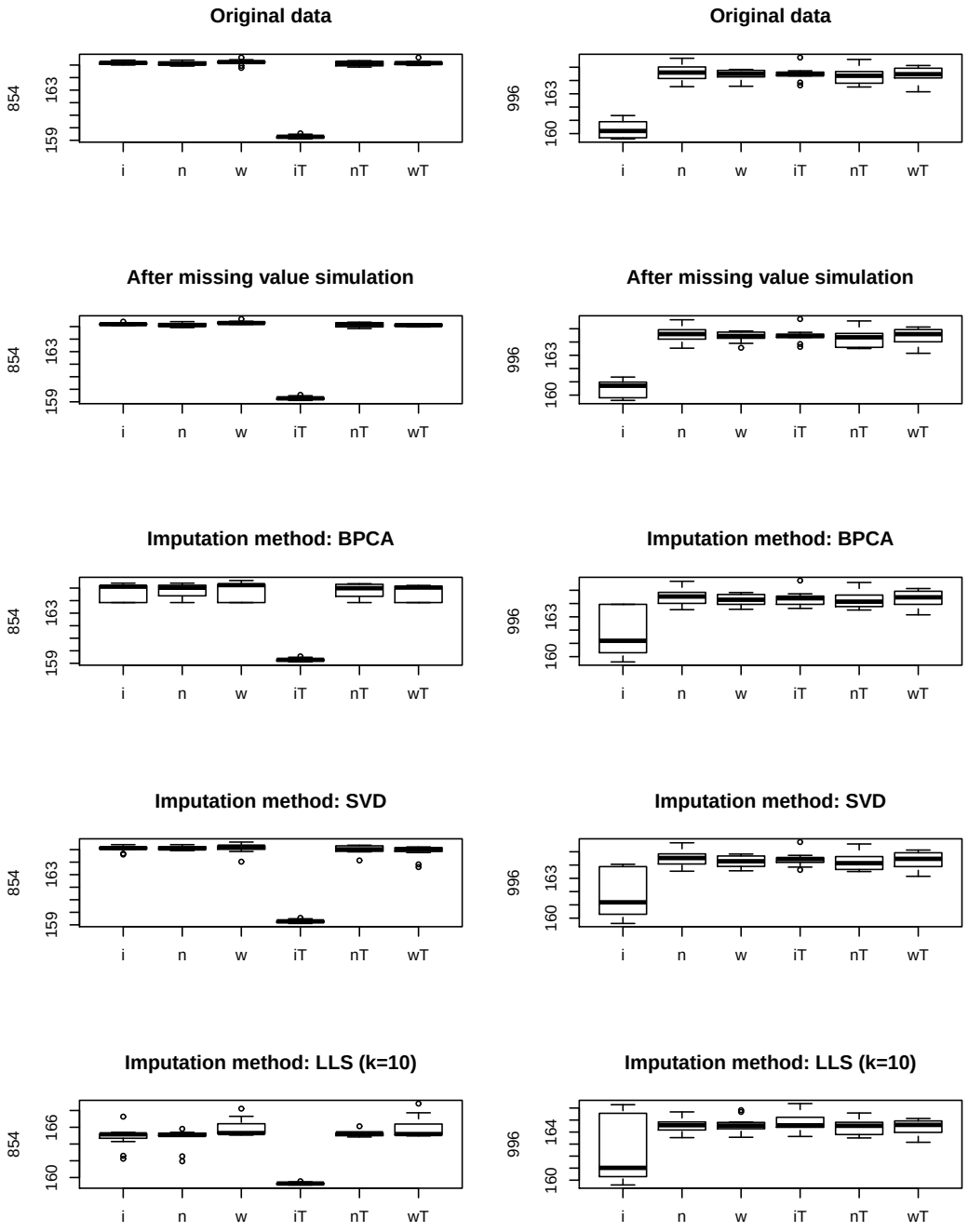

Figure 5.9: Boxplots of 2 DE masses, 854 and 996, before missing value simulation and after imputation methods. The first two boxplots are plots of the original data, i.e., the data set before randomly removing values. The second two boxplots are plots of the data set after randomly removing values. The remaining boxplots are plots of the data sets after imputation methods.

Chapter 6

Results

6.1 DI-FTICR: aqu pos/neg, org pos/neg

All the four sets aqu.pos, aqu.neg, org.pos and org.neg were very similar. We only give a detailed description for the aqu.pos set here. Summary results for the other sets are listed in Table 6.1.

	Remained after filtering (cutoff=10)	Equal variances test (Levene test)	Kruskal-Wallis test
Aqu Pos (870 masses)	102	99 (s.l.= 0.00049)	67 (s.l.= 0.00051)
Aqu Neg (3627 masses)	488	470 (s.l.= 0.0001)	448 (s.l.= 0.00011)
Org pos (1305 masses)	50	46 (s.l.= 0.001)	33 (s.l.= 0.00109)
Org Neg (3952 masses)	446	437 (s.l.= 0.00011)	179 (s.l.= 0.00012)

Table 6.1: Results from Levene's test and Kruskal-Wallis test of Metabolomic data.

For the aqu.pos set from the DI-FTICR technique, there were 870 masses observed (Table 4.1). Bonferroni corrections were used to obtain the test level for each comparison for a family-wise error rate 0.05 throughout for different total comparisons.

Before applying any statistical analysis, masses with less than 10 observations

in any of the six treatments were deleted. 102 masses remained. Levene's test was applied first to assess the equality of variances in different samples for each mass. 99 of the 102 masses passed the Levene's test at a statistical significance level of 0.05/102=0.0004901961. Of these 99 masses, 67 of them were significant by the Kruskal-Wallis test with 5 degrees of freedom (significance level=0.0005050505), i.e., 67 masses differed among treatment combinations by the Kruskal-Wallis test. Box plots and interaction plots of those 67 masses were given for further selection (Appendix A.4). Figure 6.1 illustrates the notation and order of treatments for the boxplots in the appendix. The numbers in parentheses are the numbers of non-missing observations in each treatment.

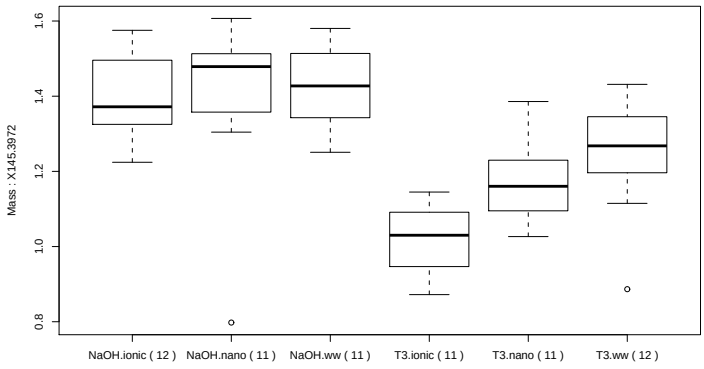

Figure 6.1: An illustration of notation and order of treatments for boxplots in the appendix.

Correspondence Analysis, CA, was conducted on the 102 masses to identify the associations between masses and the treatment combinations. The component names in the original data sets were relatively long. In order to make the texts on CA plots readable, numbers were used to label the points instead of the real mass/gene/protein names.

On the CA plot of the fold changed medians, the further away a given point is relative to the origin, the more the observed response deviates from the control response.

If the point lies close to or along the line representing a treatment condition, then the more strongly related the component response is to that treatment. Components that are differentially expressed in more than one treatment lie midway between the respective treatments. Numbers close to the origin are indecipherable, but points far away from the origin are the ones we are interested in.

Similarly, for CA plots of normalized data, if a point lies close to or along a line representing a treatment, then this component is strongly related to this treatment. The further away this point lies along the line away from the origin, the more this component is up-regulated/down-regulated compared to the other treatments. For example, in the second plot in Figure 6.3, component 26, which is mass 294.1109 in the aqu.pos data set, lies close to the line representing treatment T3.ionic, and it is on the positive side, far away from the origin. This indicates that the expression level of component 26 was up-regulated by treatment T3.ionic compared to the other two treatments. To help with interpretation of the CA plot, the medians of this component by treatment from the 10-cutoff aqu.pos data set were calculated:

index 26 (mass 294.1109)			
	ionic	nano	ww
NaOH	1.2821	3.22745	2.0136
T3	13.516	2.0236	1.5084

Table 6.2: Medians by treatment for mass 294.1109 (index 26) in the aqu.pos set.

Table 6.2 shows that the median expression of this mass in treatment T3.ionic is larger than it is in the other two T3 treatments, which agrees with the observation from the CA plot. Also, from this table, we note that this component is down-regulated by treatment NaOH.ionic compared to the other treatments in the NaOH group with a slightly higher expression in NaOH.nano than NaOH.ww. As we have expected, in the first CA plot in Figure 6.3, point 26 is on the negative side of the line representing treatment NaOH.ionic and closer to the line representing the NaOH.nano

treatment than to the line representing the NaOH.ww treatment. The display on CA plots explains the data very well.

Note that index 26 (mass 294.1109) is not in the list of DE masses by the Kruskal-Wallis test because of the large variance of the T3.ionic treatment (Figure 6.6).

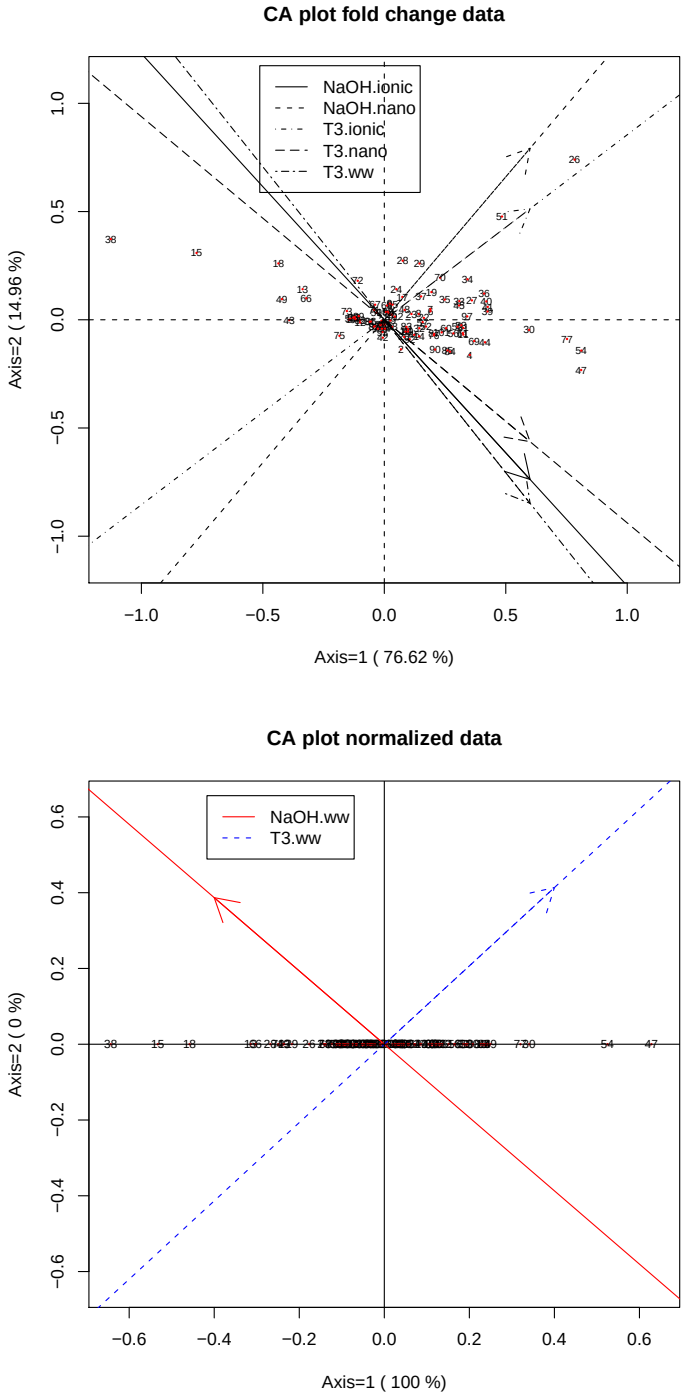

Figure 6.2: CA plots of the aqu.pos set for all the 102 masses after missing value filtering.

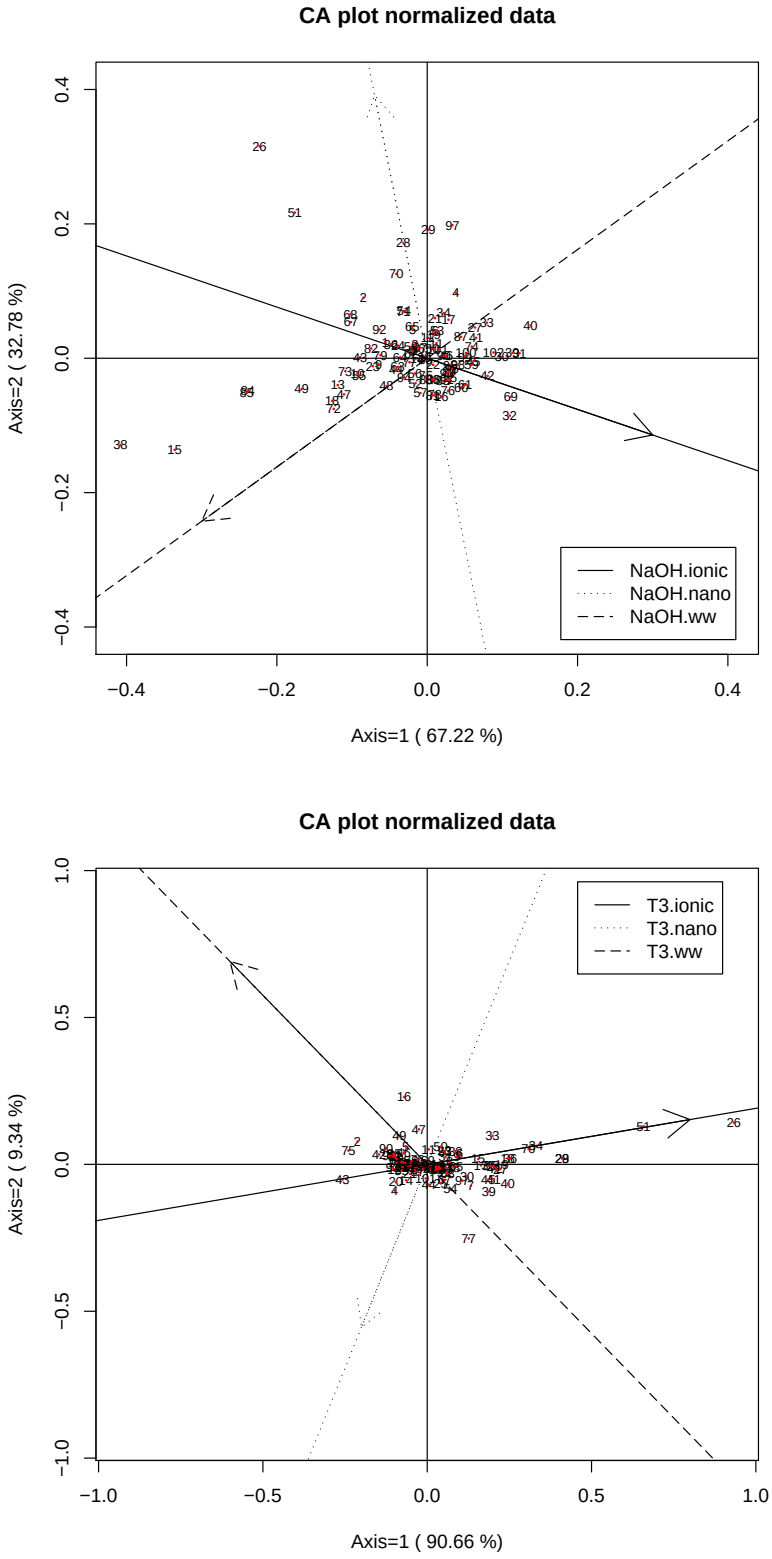

Figure 6.3: CA plots of the aqu.pos set for all the 102 masses after missing value filtering.

To improve visibility of the CA plots, numbers on the plots are indexes instead of real mass notations. In the following description, we put the real mass notation in parentheses after each index so that findings based on the CA plots can be used directly.

In the first CA plot in Figure 6.2, the positive directions of all the lines representing treatments are on the negative side of axis 1. The display clearly shows that compared to the control (the origin), components 26 (294.1109), 30 (300.1707), 47 (359.3166), 51 (368.3445), 54 (380.2903) and 77 (524.2811) were up-regulated by at least one of the treatments, and components 13 (246.1233), 15 (256.1443), 18 (270.1811), 38 (320.1606), 43 (348.1063), 49 (360.2643) and 66 (450.2021) were down-regulated by at least one of the treatments. In the second CA plot in Figure 6.2, components 15 (256.1443), 18 (270.1811) and 38 (320.1606) were down-regulated by T3.ww, while components 30 (300.1707), 47 (359.3166), 54 (380.2903) and 77 (524.2811) were up-regulated by T3.ww. In Figure 6.3, the first CA plot shows that compared to NaOH.ww, components 15 (256.1443), 38 (320.1606), 49 (360.2643), 84 (613.174) and 85 (615.1692) were down-regulated by NaOH.ionic and NaOH.nano, components 26 (294.1109) and 51 (368.3445) were up-regulated by NaOH.nano. The second CA plot in Figure 6.3 shows that compared to T3.ww, components 26 (294.1109) and 51 (368.3445) were up-regulated by T3.ionic, and component 77 (524.2811) was up-regulated by T3.ionic and T3.nano, component 43 (348.1063) was down-regulated by T3.ionic.

DE components 15 (256.1443), 26 (294.1109) and 77 (524.2811) were not in the final DE list by selection method Kruskal-Wallis test because they did not pass equal variances test (see Figure 6.6). But those components were up/down-regulated compared to the control from CA plots Figure 6.2 and Figure 6.3. Therefore, CA can be applied to unequal variances groups as a graphical selection method.

Plots from principal component analysis (PCA) reveal that there was a very clear separation between the NaOH group and the T3 group for the aqu.pos set (Figure 6.4).

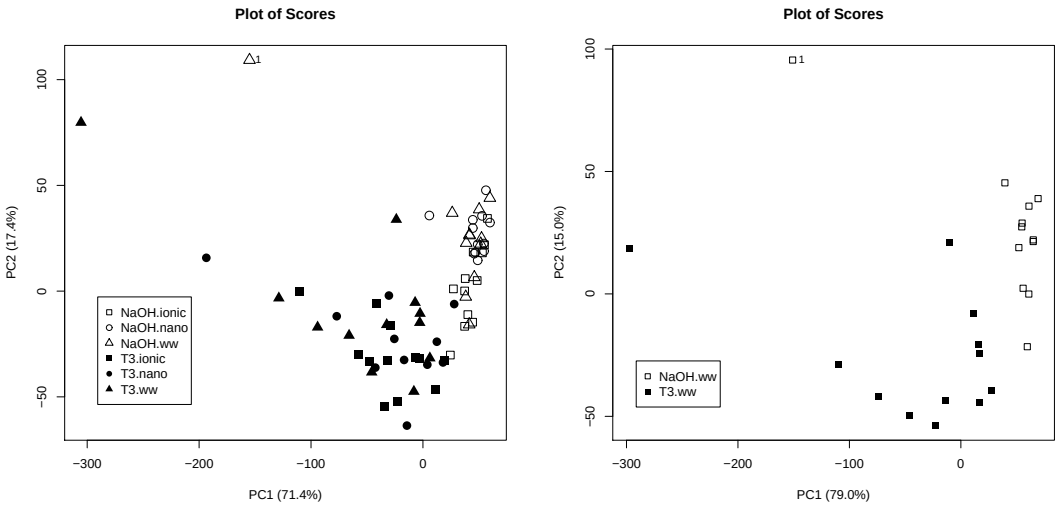

Figure 6.4: PCA plots of the aqu.pos set based on all the masses that were picked by Kruskal-Wallis tests except mass 436.1861 which is an extreme outlier in the data set (Section 5.1).

The loadings of a given principal component (PC) represent the relative extent to which the masses influence the PC. The component loadings can be interpreted as the derived relative weightings of the masses in the derived linear combination that constitutes each PC. Thus, from the scatter plot of loadings of the first two PCs, we see the relative extent to which the masses influence the two PCs (Figure 6.5).

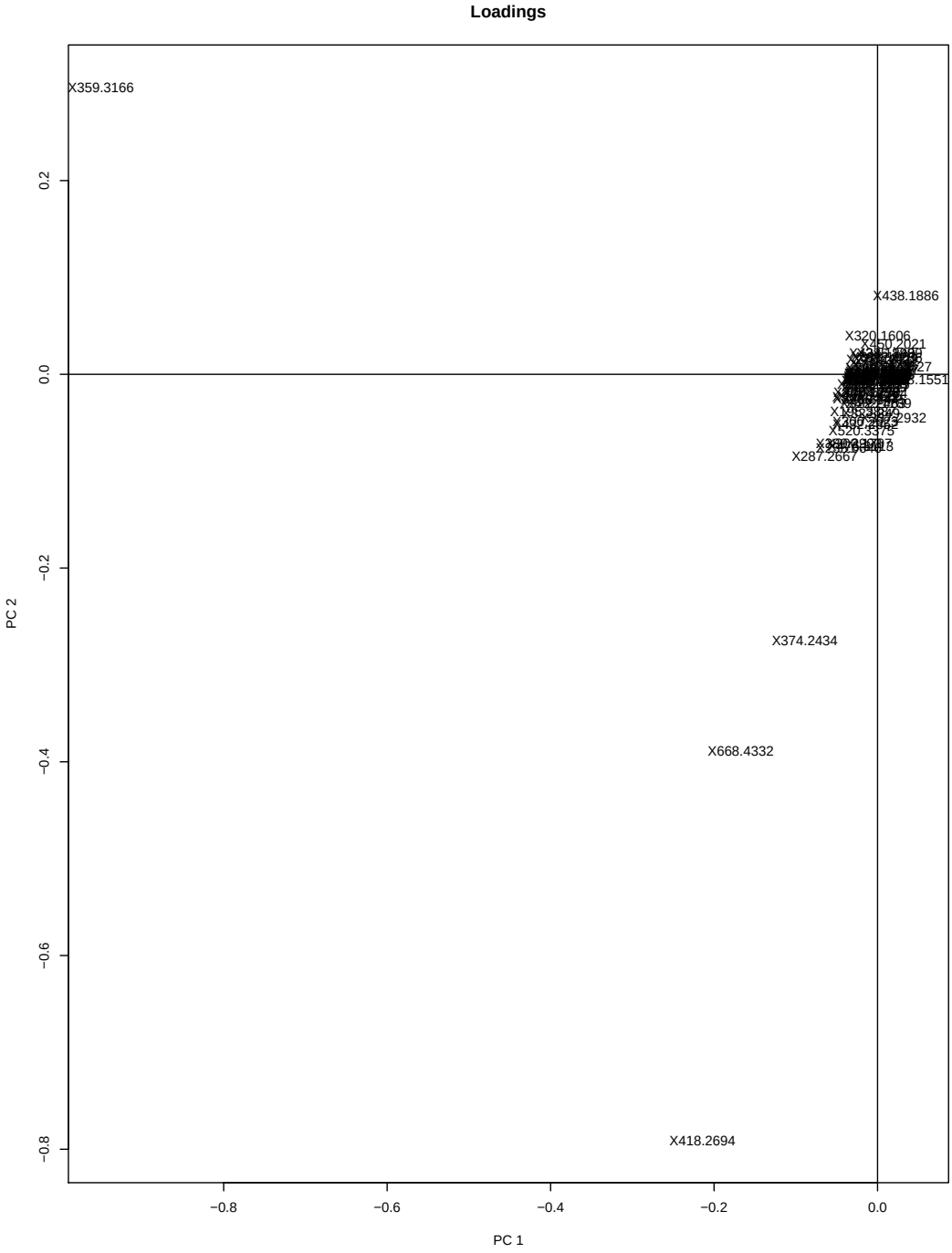

Figure 6.5: Scatter plot of loadings of the first two PCs of the aqu.pos set corresponding to the scores of the first plot in Figrue 6.4.

Figure 6.5 shows that masses 418.2694, 668.4332, 374.2434 and 359.3166 weighted more in the construction of the first two PCs than the other masses. These masses were the ones strongly associated with at least one of the treatments since there is a clear separation between the NaOH group and the T3 group in the PCA plots. Masses picked from Figure 6.5 were all significant by Kruskal-Wallis tests. However, we note that expression levels of masses standing out from the scattor plot of loadings tend to have large magnitude. Hence, ranking DE masses based on loadings needs further study.

In Table 6.3, detailed information of masses identified by CA and PCA are listed. Note that medians are used in CA calculation. CA plots only display differences among treatment medians. It doesn't check whether the differences are statistically significant. When treatment medians differ (numerically), the differences may not be significant by statistical tests. Therefore, masses that are identified by CA plots may not significant statistically. There were 57 masses significant by Kruskal-Wallis tests which were not identified by the CA plots.

Index(Mass)	Figure	Findings based on CA plots	p-value K-W test (5df)
26(294.1109)	first plot in Figure 6.2 first plot in Figure 6.3 second plot in Figure 6.3	up-regulated by at least one treatment up-regulated by NaOH.nano up-regulated by T3.ionic	fail equal variances test
30 (300.1707)	first plot in Figure 6.2	up-regulated by at least one treatment	8.8e-13
54 (380.2903)	second plot in Figure 6.2	up-regulated by T3.ww	9.653e-11
51 (368.3445)	first plot in Figure 6.2 first plot in Figure 6.3 second plot in Figure 6.3	up-regulated by at least one treatment up-regulated by NaOH.nano up-regulated by T3.ionic	1.034e-05
77 (524.2811)	first plot in Figure 6.2 second plot in Figure 6.2 second plot in Figure 6.3	up-regulated by at least one treatment up-regulated by T3.ww up-regulated by T3.ionic and T3.nano	fail equal variances test
13 (246.1233)	first plot in Figure 6.2	down-regulated by at least one treatment	9.693e-10
66 (450.2021)			2.787e-10
15 (256.1443)	first plot in Figure 6.2 second plot in Figure 6.2 first plot in Figure 6.3	down-regulated by at least one treatment down-regulated by T3.ww down-regulated by NaOH.ionic and NaOH.nano	fail equal variances test
18 (270.1811)	first plot in Figure 6.2 second plot in Figure 6.2	down-regulated by at least one treatment down-regulated by T3.ww	1.118e-08
38 (320.1606)	first plot in Figure 6.2 second plot in Figure 6.2 first plot in Figure 6.3	down-regulated by at least one treatment down-regulated by T3.ww down-regulated by NaOH.ionic and NaOH.nano	6.678e-11
43 (348.1063)	first plot in Figure 6.2 second plot in Figure 6.3	down-regulated by at least one treatment down-regulated by T3.ionic	1.864e-07
49 (360.2643)	first plot in Figure 6.2 first plot in Figure 6.2	down-regulated by at least one treatment down-regulated by NaOH.ionic and NaOH.nano	8.085e-10
84 (613.174)	first plot in Figure 6.3	down-regulated by NaOH.ionic and NaOH.nano	0.0719
85 (615.1692)			0.0826
47 (359.3166)	first plot in Figure 6.2 second plot in Figure 6.2 PCA Figure 6.5	up-regulated by at least one treatment up-regulated by T3.ww	1.352e-08
(418.2694)	PCA Figure 6.5	not notable by CA	8.94e-10
(668.4332)	PCA Figure 6.5	not notable by CA	6.085e-09
(374.2434)	PCA Figure 6.5	not notable by CA	4.166e-10
(438.1886)	PCA Figure 6.5	not notable by CA	7.802e-08

Table 6.3: List of masses identified by the CA plots and PCA loadings plot.

Boxplots of all the masses significant by Kruskal-Wallis tests are in Figure A.1 - Figure A.4. Boxplots of masses with unequal variances in Table 6.3 are given in Figure 6.6.

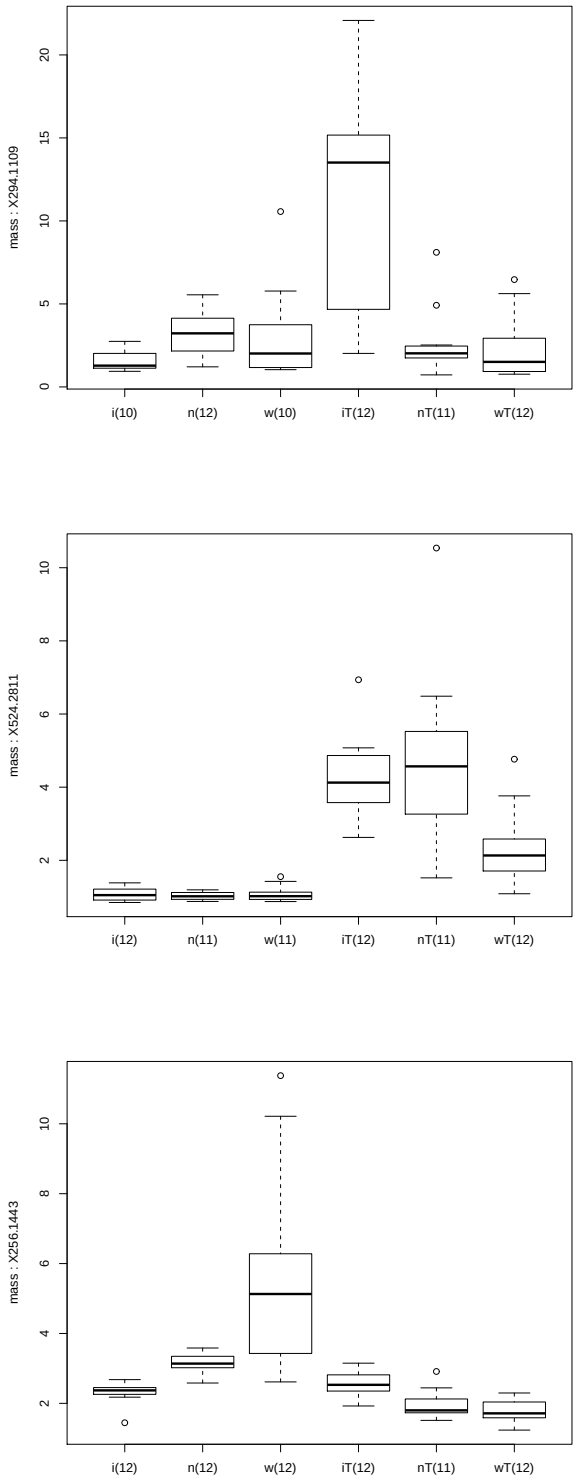

Figure 6.6: Boxplots of masses with unequal variances among treatments (Table 6.3).

6.2 Microarray: Liver, Brain

For the liver and brain microarray data set, there were 490 genes observed for each set. A significance level 0.05 was applied throughout the analysis.

There were no missing values in the liver microarray data set and just a small number of missing values in the brain set. We conducted Levene's test on each of the 490 genes observed to assess the equality of variances. For the liver set, 460 genes passed the test, and 25 of them were significant by the Kruskal-Wallis tests with 5 degrees of freedom, i.e., 25 gene transcripts differed among treatment combinations by the Kruskal-Wallis test. For the brain set, 462 genes passed the Levene's test, and 45 of them were significant by the Kruskal-Wallis test. Boxplots, interaction plots and CA plots of those significant genes are provided.

	Equal variances test (Levene test)	Kruskal-Wallis test
Liver (490 gene IDs)	460 (s.l.= 0.05)	25 (s.l.= 0.05)
Brain (490 gene IDs)	462 (s.l.= 0.05)	45 (s.l.= 0.05)

Table 6.4: Results from Levene test and Kruskal-Wallis test of Microarray data.

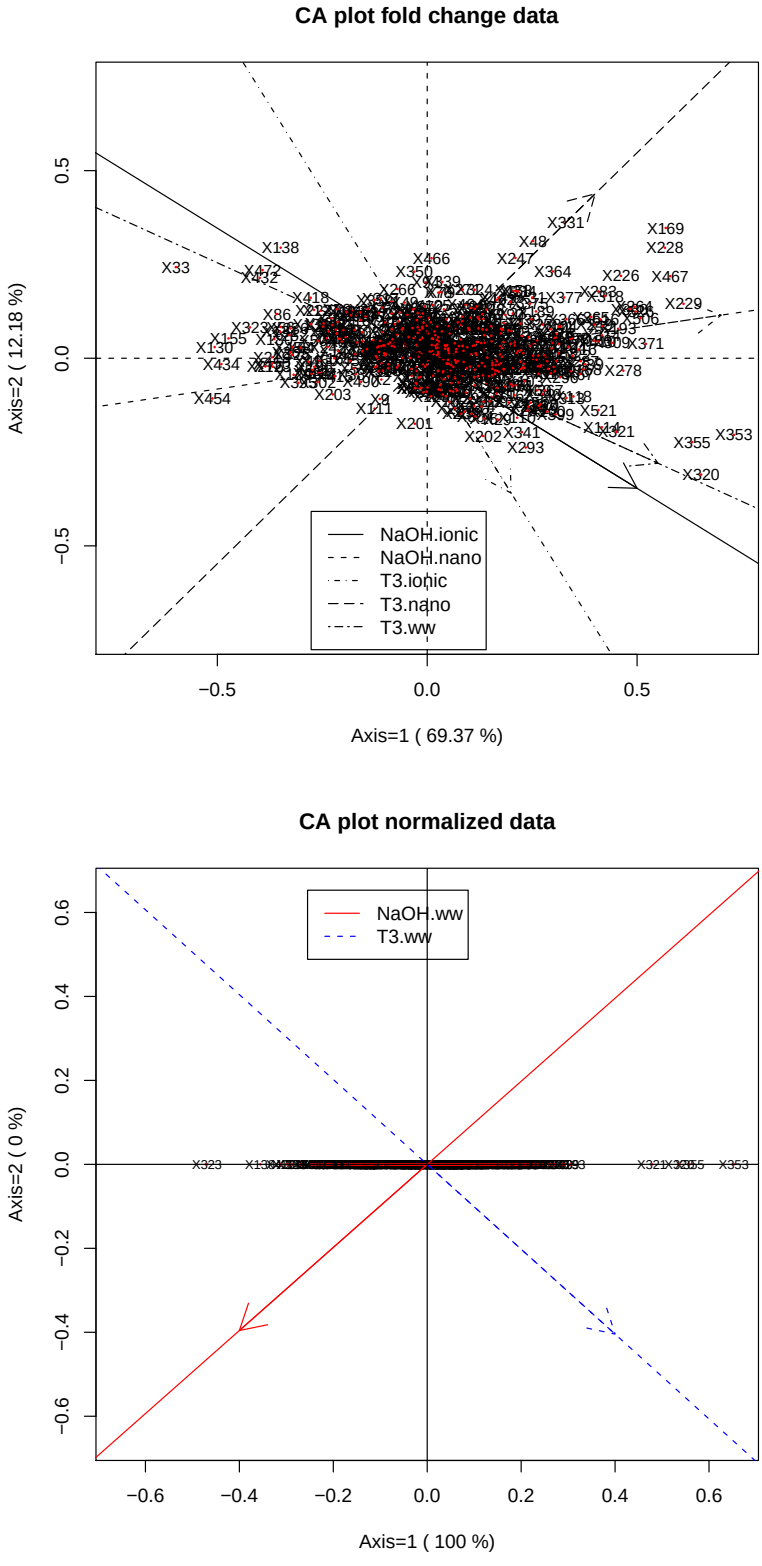

Figure 6.7: CA plots of the liver microarray set for all the 490 genes.

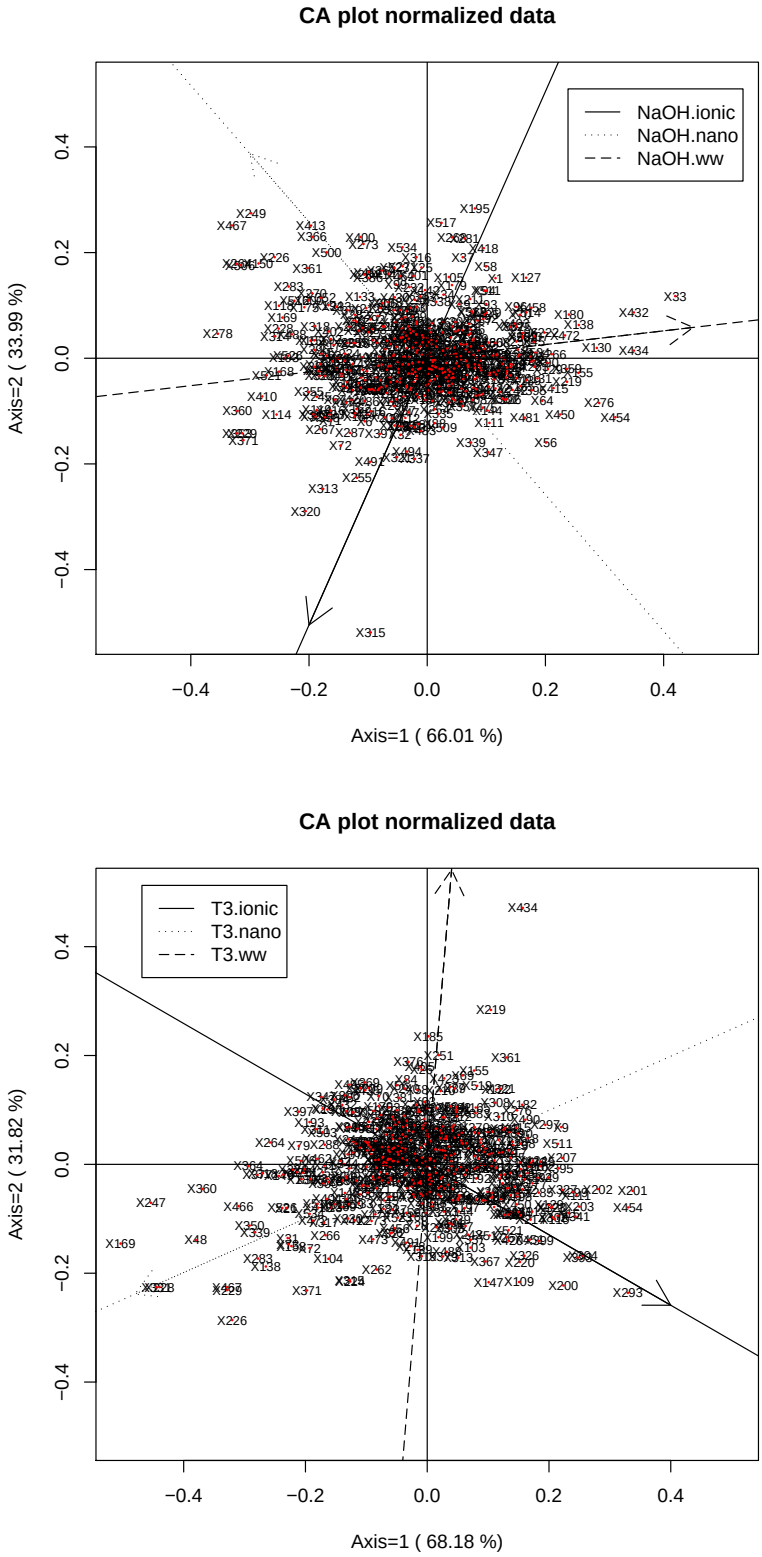

Figure 6.8: CA plots of the liver microarray set for all the 490 genes.

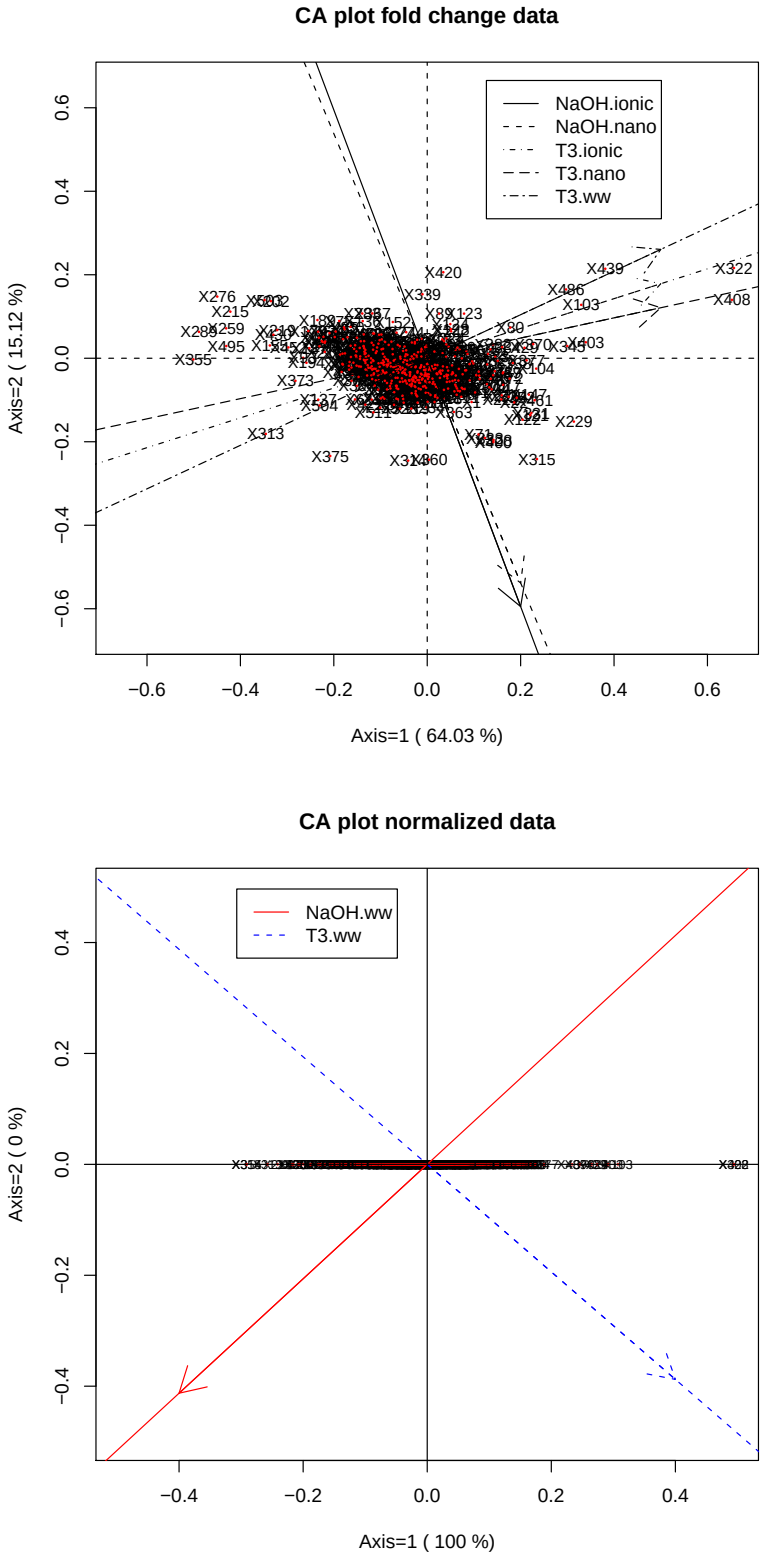

Figure 6.9: CA plots of the brain microarray set for all the 490 genes.

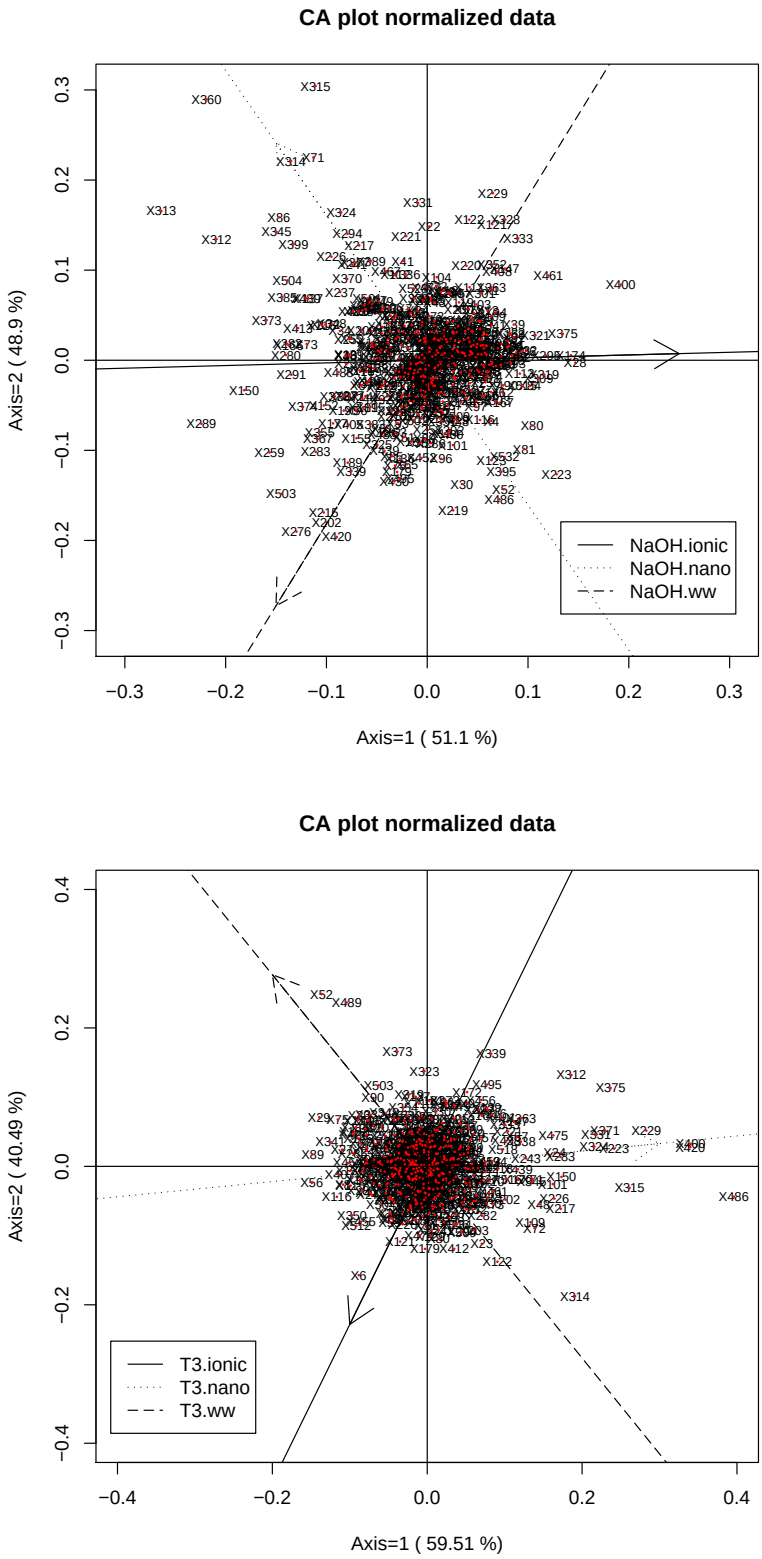

Figure 6.10: CA plots of the brain microarray set for all the 490 genes.

Unlike the CA plots of the FTICR-MS data set, gene label IDs were used in the CA plots for liver and brain microarray data instead of indexes. The gene label IDs are the same as the IDs used in Figure C.1 - Figure B.10.

In the first CA plot in Figure 6.7, the display shows that compared to the control, gene label IDs 48, 247, 331, 364, 226, 169, 228, 467, 229, 278, 521, 321, 114, 355, 353 and 320 were up-regulated by at least one of the treatments, and gene label IDs 138, 472, 432, 33, 155, 130, 434 and 454 were down-regulated by at least one of the treatments. In the second CA plot in Figure 6.7, when considering the effect of T3, gene label ID 323 was down-regulated by T3.ww, while gene label IDs 353 and 321 were up-regulated by T3.ww. In Figure 6.8, the first CA plot shows that compared to NaOH.ww, gene label IDs 130, 432, 33, 434, 450, 276, 454, 56 and 64 were down-regulated by NaOH.ionic and NaOH.nano, while gene label IDs 278, 410, 114, 360 and 371 were up-regulated by NaOH.ionic and NaOH.nano. Also, gene label IDs 249, 467, 413, 366, 226, 150, 500 and 361 were up-regulated by NaOH.nano, gene label IDs 315, 320, 313, 491 and 255 were up-regulated by NaOH.ionic, and gene label IDs 195, 517, 418, 37, 58, 281 and 268 were down-regulated by NaOH.ionic. The second CA plot in Figure 6.8 shows that compared to T3.ww, gene label IDs 293, 200, 109, 220, 326, 454, 201 and 202 were up-regulated by T3.ionic, gene label IDs 434, 219, 185, 251 and 361 were down-regulated by T3.ionic and T3.nano, and gene label IDs 169, 247, 48, 360, 466, 350, 339, 283, 138, 229, 467 and 104 were up-regulated by T3.nano.

For the brain set, in the first CA plot in Figure 6.9, the display shows that compared to the control, gene label IDs 439, 486, 103, 322, 408, 403, 229 and 315 were up-regulated by at least one of the treatments, and gene label IDs 276, 215, 289, 259, 495, 355 and 313 were down-regulated by at least one of the treatments. In the second CA plot in Figure 6.9, when considering the effect of T3, a couple of

components were up-regulated by T3.ww. In Figure 6.10, the first CA plot shows that compared to NaOH.ww, gene label IDs 150, 289, 259, 503, 276, 215, 202 and 420 were down-regulated by NaOH.ionic and NaOH.nano, while gene label IDs 229, 122, 121, 328, 333, 461 and 400 were up-regulated by NaOH.ionic and NaOH.nano. Also, gene label IDs 315, 360, 314, 71, 313, 312, 86, 345, 399, 294, 324 and 217 were up-regulated by NaOH.nano, while 80, 81, 532, 123, 395, 30, 52, 486, 219 and 223 were down-regulated by NaOH.nano. The second CA plot in Figure 6.10 shows that compared to T3.ww, gene label ID 6 was up-regulated by T3.ionic, gene label IDs 314 and 122 were up-regulated by T3.ionic and T3.nano, while gene label IDs 52, 489 and 373 were down-regulated by T3.ionic and T3.nano. Also, there are components lying along the positive side of the line representing treatment T3.nano. The expression levels of these components were up-regulated by treatment T3.nano. This indicates clear T3.nano effect in the brain set.

PCA plots of all the 36 samples with respect to the first two principal components were given for both the liver set and the brain set (Figure 6.11 and 6.12). The differences between the two levels of factor Hormone, NaOH and T3, showed clearly on the plots.

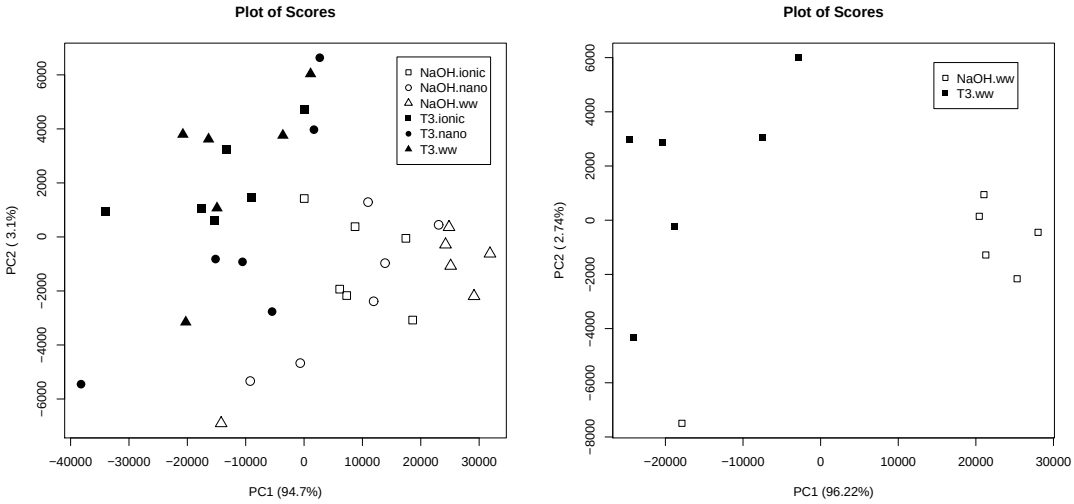

Figure 6.11: PCA plots of the liver microarray set based on the 25 transcripts that were picked by Kruskal-Wallis tests.

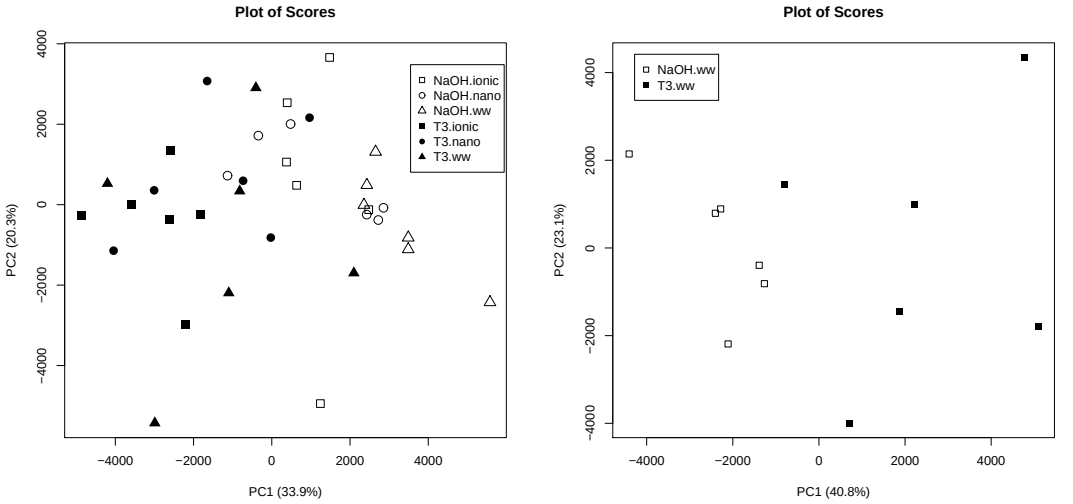

Figure 6.12: PCA plots of the brain microarray set based on the 45 transcripts that were picked by Kruskal-Wallis tests.

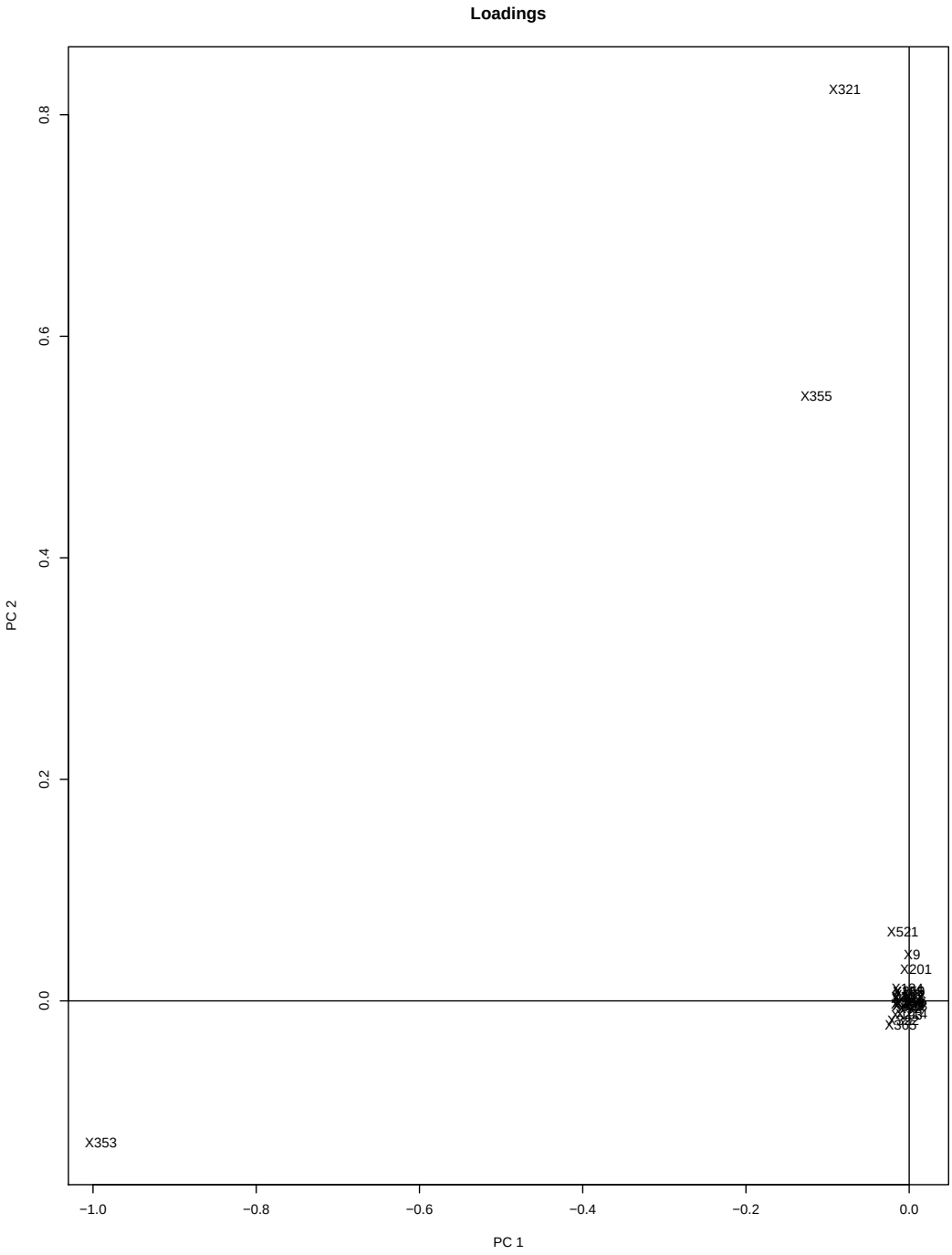

Figure 6.13: Scatter plot of loadings of the first two PCs of the liver set corresponding to the scores of the first plot in Figure 6.11.

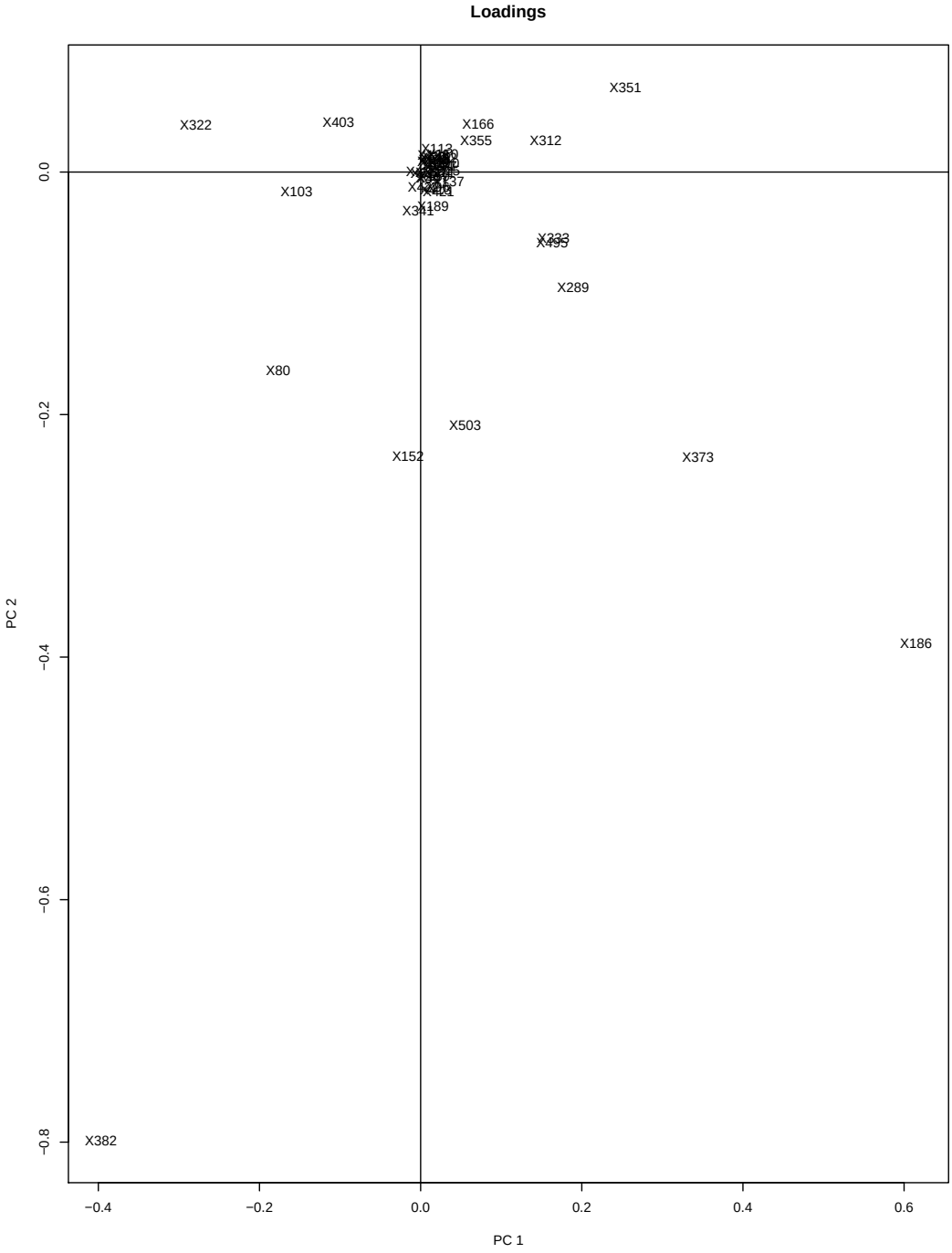

Figure 6.14: Scatter plot of loadings of the first two PCs of the brain set corresponding to the scores of the first plot in Figure 6.12.

As noted in Section 6.1, CA detects DE components by comparing medians. When sample sizes are small, it is easy to have different medians while components may not be significant statistically. For the liver and brain data, most of the genes identified by PC plots are not significant by Kruskal-Wallis tests because of small sample sizes. Gene label IDs picked from scatter plots of loadings were significant by Kruskal-Wallis tests for both liver and brain sets. Similarly to the table in Section 6.1, summary tables are given to compare CA/PCA results with Kruskal-Wallis tests using some of the prominent genes from the CA and PCA loadings plots.

Genes	Figure	Findings based on CA plots	p-value K-W test (5df)
33	first plot in Figure 6.7 first plot in Figure 6.8	down-regulated by at least one treatment down-regulated by NaOH.ionic and NaOH.nano	0.1353
353	first plot in Figure 6.7 second plot in Figure 6.7 PCA Figure 6.13	up-regulated by at least one treatment up-regulated by T3.ww	0.0011
355	first plot in Figure 6.7 PCA Figure 6.13	up-regulated by at least one treatment	0.0007
320	first plot in Figure 6.7 first plot in Figure 6.8	up-regulated by at least one treatment up-regulated by NaOH.ionic	fail equal variances test
229	first plot in Figure 6.7 second plot in Figure 6.8	up-regulated by at least one treatment up-regulated by T3.nano	0.4043
169	first plot in Figure 6.7 second plot in Figure 6.8	up-regulated by at least one treatment up-regulated by T3.nano	0.1049
323	second plot in Figure 6.7	down-regulated by T3.ww	fail equal variances test
315	first plot in Figure 6.8	up-regulated by NaOH.ionic	0.4300
434	second plot in Figure 6.8	down-regulated by T3.ionic and T3.nano	0.2219
293	second plot in Figure 6.8	up-regulated by T3.ionic	0.1166
321	PCA Figure 6.13	not notable by CA	0.0004
521	PCA Figure 6.13	not notable by CA	0.0277
9	PCA Figure 6.13	not notable by CA	0.0272
201	PCA Figure 6.13	not notable by CA	0.0293

Table 6.5: List of genes identified by the CA plots and PCA loadings plot of the liver set.

Genes	Figure	Findings based on CA plots	p-value K-W test (5df)
289	first plot in Figure 6.9 first plot in Figure 6.10 PCA Figure 6.14	down-regulated by at least one treatment down-regulated by NaOH.ionic and NaOH.nano	0.0218
355	first plot in Figure 6.9 PCA Figure 6.14	down-regulated by at least one treatment	0.0011
322	first plot in Figure 6.9 PCA Figure 6.14	up-regulated by at least one treatment	0.0009
408	first plot in Figure 6.9	up-regulated by at least one treatment	fail equal variances test
314	first plot in Figure 6.10 second plot in Figure 6.10	up-regulated by NaOH.nano up-regulated by T3.ionic and T3.nano	0.2175
360	first plot in Figure 6.10	up-regulated by NaOH.nano	0.1311
315	first plot in Figure 6.10 second plot in Figure 6.10	up-regulated by NaOH.nano up-regulated by T3.nano	0.4677
503	first plot in Figure 6.10 PCA Figure 6.14	down-regulated by NaOH.ionic and NaOH.nano	0.0355
52	second plot in Figure 6.10	down-regulated by T3.ionic and T3.nano	0.3218
489	second plot in Figure 6.10	down-regulated by T3.ionic and T3.nano	0.8319
373	second plot in Figure 6.10 PCA Figure 6.14	down-regulated by T3.ionic and T3.nano	0.0028
486	first plot in Figure 6.9 second plot in Figure 6.10	up-regulated by at least one treatment up-regulated by T3.nano	fail equal variances test
382	PCA Figure 6.14	not notable by CA	0.0478
186	PCA Figure 6.14	not notable by CA	0.0227
152	PCA Figure 6.14	not notable by CA	0.0116
80	PCA Figure 6.14	not notable by CA	0.0499
103	PCA Figure 6.14	not notable by CA	0.0006
403	PCA Figure 6.14	not notable by CA	0.0499
351	PCA Figure 6.14	not notable by CA	0.0075
166	PCA Figure 6.14	not notable by CA	0.0146
312	PCA Figure 6.14	not notable by CA	0.0275

Table 6.6: List of genes identified by the CA plots and PCA loadings plot of the brain set.

Boxplots of all the significant genes by Kruskal-Wallis tests are given in Appendix B.1. Boxplots of genes with unequal variances in Table 6.5 and Table 6.6 are given in Figure 6.15 and Figure 6.16 respectively.

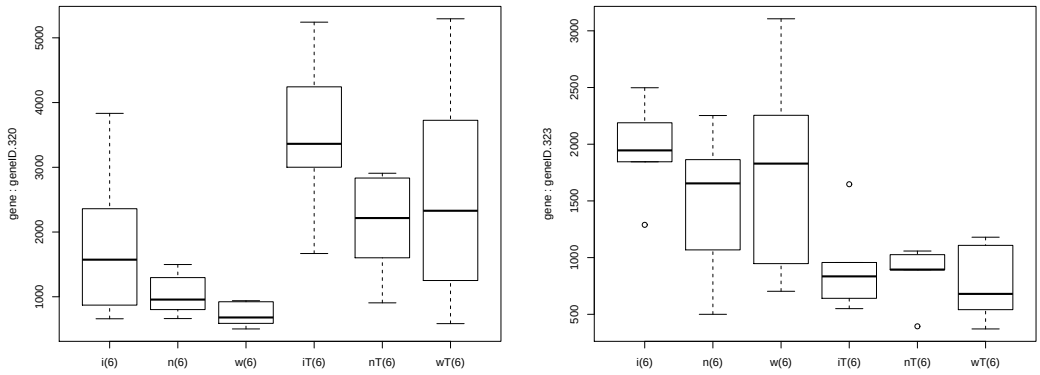

Figure 6.15: Boxplots of genes with unequal variances among treatments (Table 6.5).

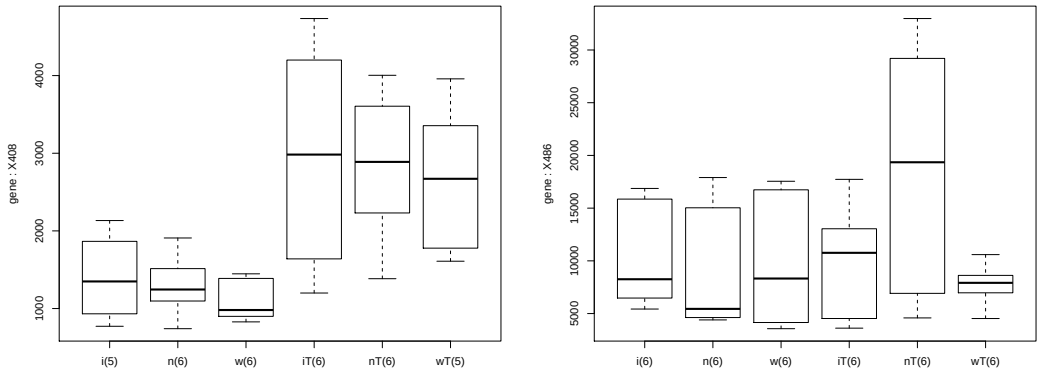

Figure 6.16: Boxplots of genes with unequal variances among treatments (Table 6.6).

6.3 iTraq

There were 1051 protein groups in the iTraq set after regrouping (Section 3.2.3). A significance level 0.05 was applied throughout the analysis. There were many missing values in the data set. However, missing value analysis using the two-factor factorial logistic regression model yielded no significant groups. In order to keep most of the protein groups in the analysis while maintaining stable Kruskal-Wallis tests, we deleted group IDs with less than 5 observations in any of the six treatments. 280 groups remained. 272 of the 280 groups passed Levene's test at a statistical significance level of 0.05, and 10 of them were significant by the Kruskal-Wallis test with 5 degrees of freedom (significance level=0.05). Box plots and interaction plots of those 10 protein groups are given in Appendix C.1.

	Remained after filtering (cutoff=5)	Equal variances test (Levene test)	Kruskal-Wallis test
iTraq (1051 groups)	280	272 (s.l.= 0.05)	10 (s.l.= 0.05)

Table 6.7: Results from Levene's test and Kruskal-Wallis test of iTraq data.

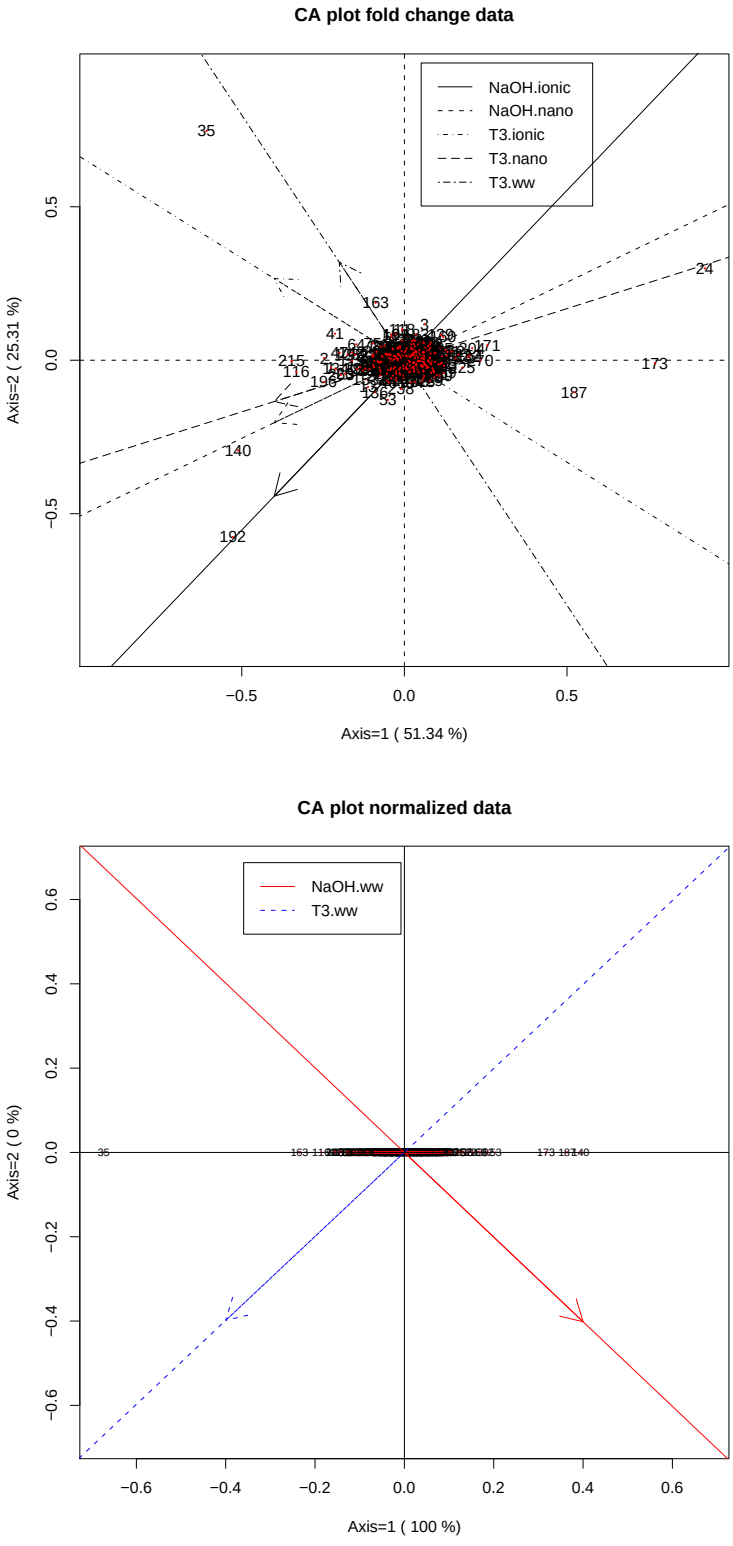

Figure 6.17: CA of the iTraq set for all the 280 protein groups after missing value filtering.

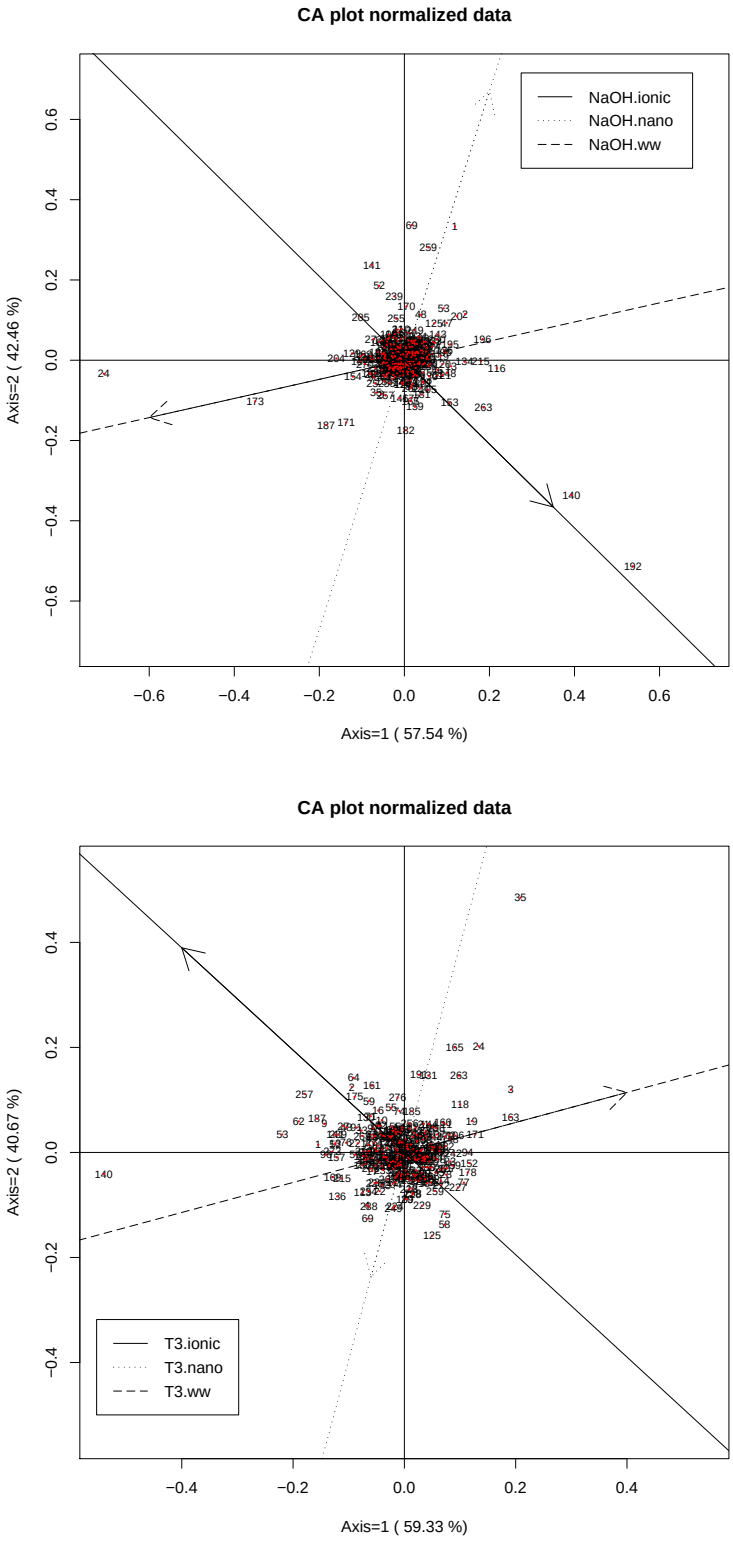

Figure 6.18: CA of the iTraq set for all the 280 protein groups after missing value filtering.

In the first CA plot in Figure 6.17, the display shows that compared to the control, components 35 (PGC46), 140 (PGC771) and 192 (PGC131) were up-regulated by at least one of the treatments, and components 24 (PGC32), 173 (PGC126) and 187 (PGC385) were down-regulated by at least one of the treatments. In the second CA plot in Figure 6.17, when considering the effect of T3, components 173 (PGC126) was down-regulated by T3.ww, while component 35 (PGC46)was up-regulated by T3.ww. In Figure 6.18, the first CA plot shows that compared to NaOH.ww, components 24 (PGC32), 187 (PGC385) and 173 (PGC126) were down-regulated by NaOH.ionic and NaOH.nano, components 69 (PGC123), 259 (PGC336) and 1 (PGC1) were up-regulated by NaOH.nano, and components 140 (PGC771) and 192 (PGC131) were up-regulated by NaOH.ionic. The second CA plot in Figure 6.18 shows that compared to T3.ww, component 140 (PGC771) was up-regulated by T3.ionic and T3.nano, and component 35 (PGC46) was down-regulated by T3.nano.

A PCA plot of all the 71 samples with respect to the first two principal components is given here. The samples are all mixed together. There are no clear divisions among the groups.

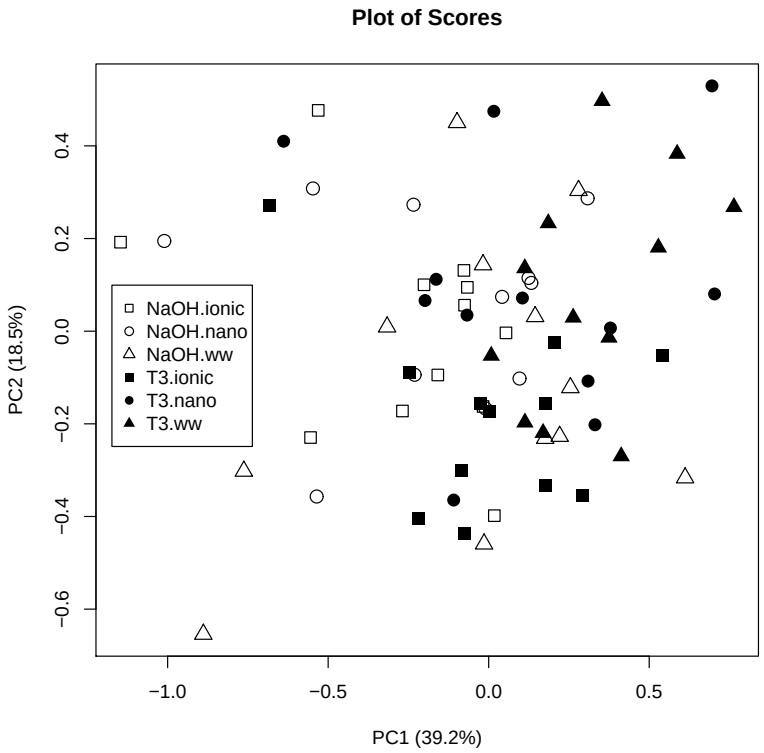

Figure 6.19: PCA plot of the iTraq data set based on the 10 protein groups that were picked by Kruskal-Wallis tests.

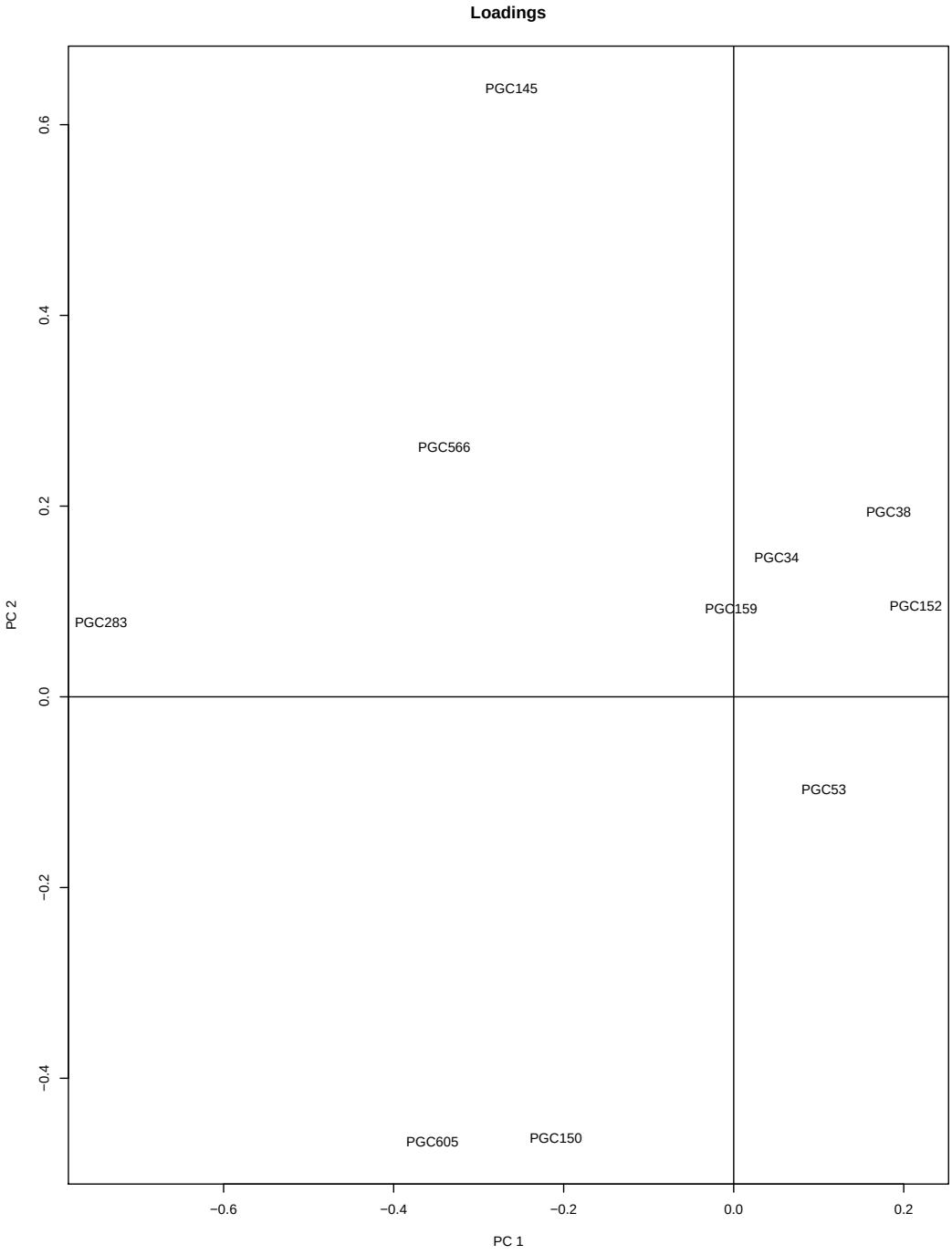

Figure 6.20: Scatter plot of loadings of the first two PCs of the iTraq set.

A table similar to those given previous sections is given to compare CA/PCA results with Kruskal-Wallis tests using some of the prominent protein groups from the CA and PCA loadings plots. None of the significant protein groups picked by Kruskal-Wallis tests are identified clearly by CA plots.

Protein groups	Figure	Findings based on CA plots	p-value K-W test (5df)
35 (PGC46)	first plot in Figure 6.17 second plot in Figure 6.17 second plot in Figure 6.18	up-regulated by at least one treatment up-regulated by T3.ww down-regulated by T3.nano	0.1850
140 (PGC771)	first plot in Figure 6.17 second plot in Figure 6.17 first plot in Figure 6.18 second plot in Figure 6.18	up-regulated by at least one treatment down-regulated by T3.ww up-regulated by NaOH.ionic up-regulated by T3.ionic and T3.nano	fail equal variances test
192 (PGC131)	first plot in Figure 6.17 first plot in Figure 6.18	up-regulated by at least one treatment up-regulated by NaOH.ionic	fail equal variances test
24 (PGC32)	first plot in Figure 6.17 first plot in Figure 6.18	down-regulated by at least one treatment down-regulated by NaOH.ionic and NaOH.nano	0.8876
173 (PGC126)	first plot in Figure 6.17 second plot in Figure 6.17 first plot in Figure 6.18	down-regulated by at least one treatment down-regulated by T3.ww down-regulated by NaOH.ionic and NaOH.nano	0.636
187 (PGC385) 69 (PGC123)	first plot in Figure 6.17 second plot in Figure 6.17 first plot in Figure 6.18 first plot in Figure 6.18	down-regulated by at least one treatment down-regulated by T3.ww down-regulated by NaOH.ionic and NaOH.nano up-regulated by NaOH.nano	0.2809 0.1043
259 (PGC336)	first plot in Figure 6.18	up-regulated by NaOH.nano	fail equal variances test
1 (PGC1)	first plot in Figure 6.18	up-regulated by NaOH.nano	0.1594
(PGC145)	PCA Figure 6.20	not notable by CA	0.0232
(PGC150)	PCA Figure 6.20	not notable by CA	0.0183
(PGC605)	PCA Figure 6.20	not notable by CA	0.0250
(PGC283)	PCA Figure 6.20	not notable by CA	0.0216
(PGC566)	PCA Figure 6.20	not notable by CA	0.0500

Table 6.8: List of protein groups identified by the CA plots and PCA loadings plot of the iTraq data.

Boxplots of all the significant protein groups by Kruskal-Wallis tests are given in Figure C.1. Boxplots of the three protein groups that had unequal variances but were identified by CA plots (Table 6.8) are given in Figure 6.21.

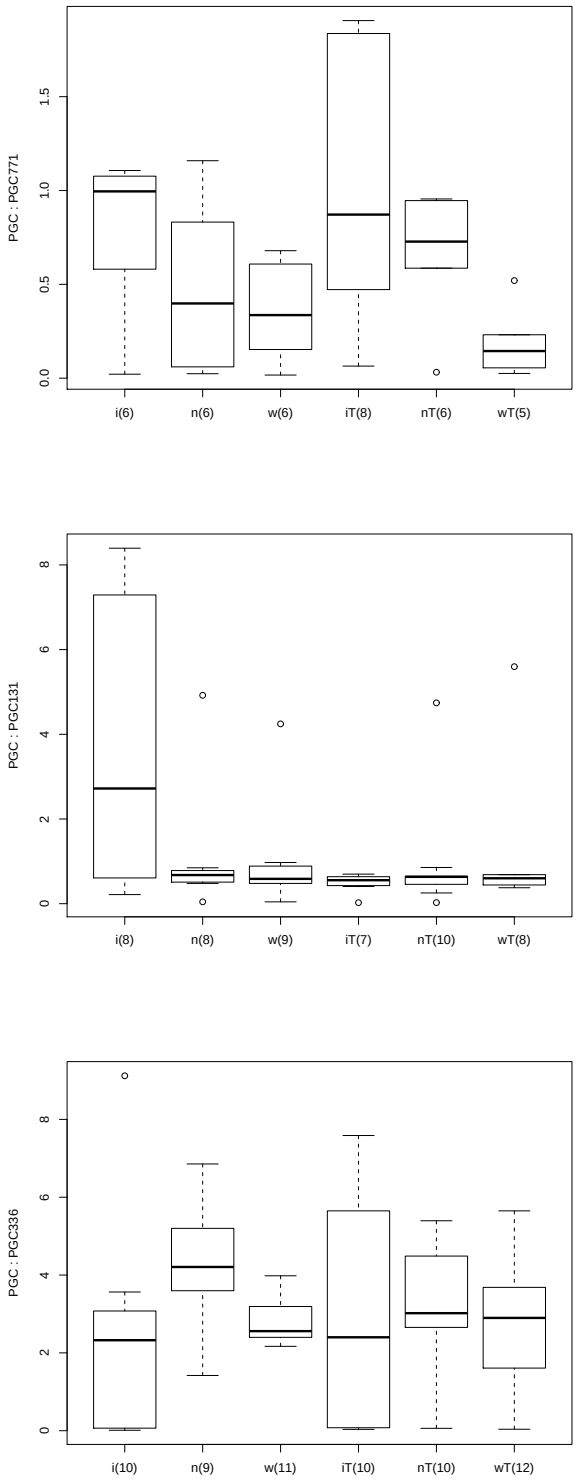

Figure 6.21: Boxplots of protein groups with unequal variances among treatments (Table 6.8).

Chapter 7

Conclusions

Data sets from the frogSCOPE project covered the identification of various components from genes to proteins to metabolites. New technologies allowed us to search and identify hundreds and thousands of genes/proteins/metabolites at the same time. When testing null hypotheses in component, the actual type I error rate explodes as the number of null hypotheses gets large. In the frogSCOPE project, Bonferroni correction was applied to control type I error rate. The adjusted p-value was defined by the desired family error rate divided by the total number of comparisons. The idea behind this is to require a stronger level of evidence to be observed in order for an individual comparison be deemed "significant", so as to compensate for the number of inferences being made. From the data analysis, we observed that the adjusted p-value from Bonferroni correction was too small for the microarrays array data sets and iTraq data. No components were significant after Bonferroni correction. When this happened, a significance level 0.05 was used for each comparison, and boxplots were provided for further analysis.

Two technologies, FTICR-MS and UPLC-MS, were applied to identify and quantify metabolites. The FTICR-MS data were high in missing value percentages. The

application of FTICR-MS technology to identify metabolites was fairly new. Analysis on the missing value pattern and performance of imputation methods were essential for potential biomarker detection.

- In Chapter 4, we calculated the missing value percentages within each treatment for each FTICR-MS set. Those tables showed that the control condition, NaOH.ww, consistently had the highest missing values percentage in the aqueous FTICR-MS data sets. This observation suggested that there might be a special chemical phenomena related to the control condition. Further analysis in lab and proteomic center was suggested. Also, the FTICR-MS sets were obtained following the workflow given in Appendix A.1. Missing values analysis tables could help technicians to improve the workflow or instrument tuning if necessary.

- In Chapter 5, we simulated FTICR-MS data sets for further study. Simulated data is a useful tool for comparing different selection methods and imputation methods. The studying of real data scales and variations was very important for data simulation afterward. It improved the quality of the simulated data.

  In this chapter, selection methods Kruskal-Wallis test, MTP and Limma were compared. The Kruskal-Wallis test is a nonparametric method. It is an extension of the Mann-Whitney U test to 3 or more groups. MTP and Limma are packages from the Bioconductor project, which is open source software for bioinformatics and provides tools for the analysis and comprehension of high-throughput genomic data. MTP and Limma are two selection methods created to facilitate the analysis of high-throughput data. The control of type I error rate is included in the functions as one of the parameters. The performance of the Bioconductor packages is hard to evaluate since different testing data sets

were used during the development of the package. In this study, those three functions were compared using the simulated data sets. As a result, MTP was more sensitive to variations among treatments, which in turn brings relatively more false positive selections, while Kruskal-Wallis test was quite conservative, and the false positive rate was pretty low. The performance of Limma was in the middle, between MTP and Kruskal-Wallis test.

Performance of three most common imputation methods: BPCA, SVD and LLS, was also evaluated. Analysis showed that differences among treatments lessened or disappeared after applying imputation methods to fill in blanks. We also noticed that imputation methods, in a sense, were conservative since no new DE mass were generated in our experiment. Therefore, it is safe to apply imputation methods when it is necessary.

After missing value analysis, in Chapter 6, multiple comparisons were carried out by applying Kruskal-Wallis test on components which have passed Levene's test. Significance levels were either defined by Bonferroni correction or using the common level 0.05 when the adjusted p-value by Bonferroni correction was too small. Boxplots and interaction plots were handy tools for further selection of potential biomarkers.

PCA plots indicated relative positions of the samples. There were always clear separations between NaOH.ww and T3.ww except the iTraq data set. During the data analysis, we discovered that principal component analysis worked better when only components differing among treatments were included. More components brought more noise than information when trying to group the samples. One example is given here to illustrate the idea (Figure 7.1). The first plot was based on PCA of masses picked by Kruskal-Wallis test, while the second plot was based on PCA analysis of masses with at least 10 non-missing values in each treatment.

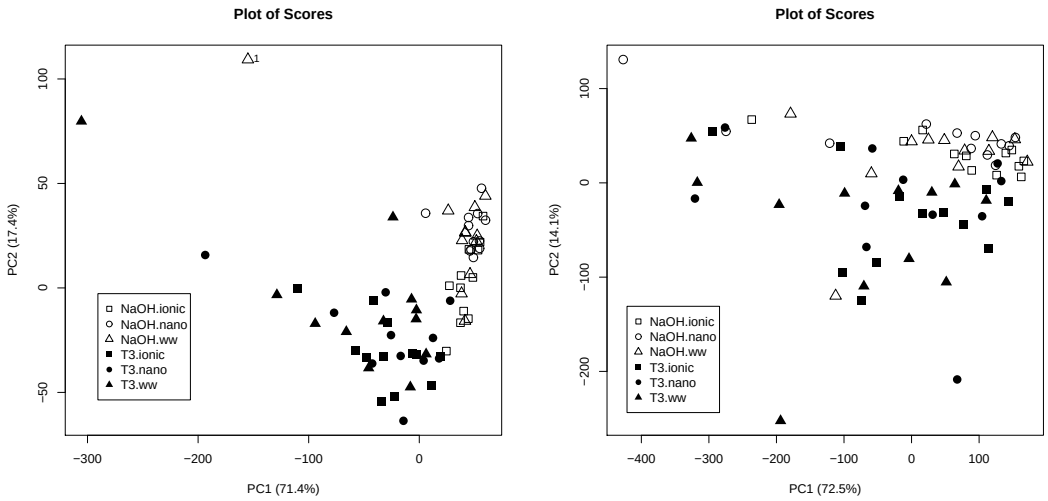

Figure 7.1: A comparision of two PCA plots based on different masses in the aqu.pos set.

CA plots were very neat in presenting the whole data set in two dimensions. Some of the plots were hard to read when there were many of components. But the points that lie far away from the origin and stayed closed to one of the treatment lines are the ones that can be interesting. So CA plots contains valuable information.

The frogSCOPE project proceeded with database searching of the potential biomarkers picked by Kruskal-Wallis tests to narrow down biomarker candidates. In the end, promising biomarkers will undergo rigorous validation.

The frogSCOPE project data are very typical sets of current powerful analytical technologies. The findings and results in this thesis are useful for other statisticians when encountering the same type of data.

Appendix A

FTICR-MS data (hiAgd6 set)

A.1 Workflow for FTICR-MS Data

For tadpole serum FTICR-MS data, raw mass spectra were batch-processed with a custom VBA script developed within DA as follows.

1. Automate internal mass calibration with the reference masses of the calibration standards spiked into each sample right before MS acquisitions.

2. Pick up monoisotopic peaks corresponding to the metabolite isotopic distribution patterns.

3. Sodium-adduct ions, $(M + Na)+$, are filtered out of the positive ESI spectra based on the expected mass differences for these ions within 3 ppm; for negative ESI spectra, Cl-adducted ions, (M+Cl)-, were filtered out within the sample mass error.

4. Generate a mass list of unique chemical component masses together with their peak intensities for each raw MS data.

The above generated mass list files from all the mass spectra in both positive and negative ion modes are then processed with another custom developed software program.

5. Convert the monoisotopic m/z's to neutral masses by subtracting 1.007276 for positive ESI mode, by adding 1.007276 Da for negative ion mode.

6. Normalize each peak's intensity of all the neutral masses with the intra-sample total ion intensity, then multiply by 10000 to make the value not too small.

7. Align the masses observed in at least 9 of 36 samples in the NaOH or T3 group, i.e., to combine the masses into unique metabolite features, matched within 3 ppm across all the samples in the same group.

8. Save the output in a tab delimited format for further data modeling and statistics.

A.2 Count: Missing vs. Non-missing

A.2.1 aqu.neg

1. For all the 3627 masses:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	19414	18794	23304	19280	17648	19248	117688
	Percentage	44.61	47.11	**53.54**	44.30	44.23	44.22	46.35
Non-missing	Count	24110	21103	20220	24244	22249	24276	136202
	Percentage	55.39	52.89	46.46	55.70	55.77	55.78	53.65
Sum	Count	43524	39897	43524	43524	39897	43524	253890
	Percentage	100	100	100	100	100	100	100

2. For 1173 masses left after deleting masses with less than 5 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	1616	1587	2930	1628	1400	1796	10957
	Percentage	11.48	12.30	**20.82**	11.57	10.85	12.76	13.34
Non-missing	Count	12460	11316	11146	12448	11503	12280	71153
	Percentage	88.52	87.70	79.18	88.43	89.15	87.24	86.66
Sum	Count	14076	12903	14076	14076	12903	14076	82110
	Percentage	100	100	100	100	100	100	100

3. For 615 masses left after deleting masses with less than 9 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	173	141	490	244	160	311	1519
	Percentage	2.34	2.08	**6.64**	3.31	2.37	4.21	3.53
Non-missing	Count	7207	6624	6890	7136	6605	7069	41531
	Percentage	97.66	97.92	93.36	96.69	97.63	95.79	96.47
Sum	Count	7380	6765	7380	7380	6765	7380	43050
	Percentage	100	100	100	100	100	100	100

4. For 488 masses left after deleting masses with less than 10 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	71	44	211	121	54	161	662
	Percentage	1.21	0.82	**3.60**	2.07	1.01	2.75	1.94
Non-missing	Count	5785	5324	5645	5735	5314	5695	33498
	Percentage	98.79	99.18	96.40	97.93	98.99	97.25	98.06
Sum	Count	5856	5368	5856	5856	5368	5856	34160
	Percentage	100	100	100	100	100	100	100

5. For 369 masses left after deleting masses with less than 11 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	19	0	86	51	0	66	222
	Percentage	0.43	0.00	**1.94**	1.15	0.00	1.49	0.86
Non-missing	Count	4409	4059	4342	4377	4059	4362	25608
	Percentage	99.57	100.00	98.06	98.85	100.00	98.51	99.14
Sum	Count	4428	4059	4428	4428	4059	4428	25830
	Percentage	100	100	100	100	100	100	100

A.2.2 org.pos

1. For all the 1305 masses:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	8962	7931	8055	7712	6404	6492	45556
	Percentage	62.43	60.77	56.11	49.25	44.61	41.46	52.10
Non-missing	Count	5393	5119	6300	7948	7951	9168	41879
	Percentage	37.57	39.23	43.89	50.75	55.39	58.54	47.90
Sum	Count	14355	13050	14355	15660	14355	15660	87435
	Percentage	100	100	100	100	100	100	100

2. For 85 masses left after deleting masses with less than 9 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	27	27	15	44	21	14	148
	Percentage	2.89	3.18	1.60	4.31	2.25	1.37	2.60
Non-missing	Count	908	823	920	976	914	1006	5547
	Percentage	97.11	96.82	98.40	95.69	97.75	98.63	97.40
Sum	Count	935	850	935	1020	935	1020	5695
	Percentage	100	100	100	100	100	100	100

3. For 50 masses left after deleting masses with less than 10 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	5.00	0.00	5.00	11.00	2.00	3.00	26.00
	Percentage	0.91	0.00	0.91	1.83	0.36	0.50	0.78
Non-missing	Count	545	500	545	589	548	597	3324
	Percentage	99.09	100.00	99.09	98.17	99.64	99.50	99.22
Sum	Count	550	500	550	600	550	600	3350
	Percentage	100	100	100	100	100	100	100

A.2.3 org.neg

1. For all the 3952 masses:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	22877	20898	24126	21691	19945	21313	130850
	Percentage	48.24	48.07	50.87	45.74	45.88	49.03	47.99
Non-missing	Count	24547	22574	23298	25733	23527	22159	141838
	Percentage	51.76	51.93	49.13	54.26	54.12	50.97	52.01
Sum	Count	47424	43472	47424	47424	43472	43472	272688
	Percentage	100	100	100	100	100	100	100

2. For 604 masses left after deleting masses with less than 9 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	321	169	255	344	112	246	1447
	Percentage	4.43	2.54	3.52	4.75	1.69	3.70	3.47
Non-missing	Count	6927	6475	6993	6904	6532	6398	40229
	Percentage	95.57	97.46	96.48	95.25	98.31	96.30	96.53
Sum	Count	7248	6644	7248	7248	6644	6644	41676
	Percentage	100	100	100	100	100	100	100

3. For 446 masses left after deleting masses with less than 10 observations within any of the 6 treatment:

status	treatment	NaOH.ionic	NaOH.nano	NaOH.ww	T3.ionic	T3.nano	T3.ww	Sum
Missing	Count	124	47	107	167	28	77	550
	Percentage	2.32	0.96	2.00	3.12	0.57	1.57	1.79
Non-missing	Count	5228	4859	5245	5185	4878	4829	30224
	Percentage	97.68	99.04	98.00	96.88	99.43	98.43	98.21
Sum	Count	5352	4906	5352	5352	4906	4906	30774
	Percentage	100	100	100	100	100	100	100

A.3 Kruskal-Wallis Results at Different Cutoffs

A.3.1 aqu.neg

	number of masses			
	after filtering	Levene's test $s.l. = 0.05$	Kruskal-Wallis test $s.l. = 0.05/M$	$\frac{\text{Kruskal-Wallis test}}{\text{after filtering}}$ %
Total	3627	2444	372	10.26%
at least **1** observation in each treatment	2567	1559	389	15.15%
at least **2** observations in each treatment	1992	1066	404	20.28%
at least **3** observations in each treatment	1614	753	405	25.09%
at least **4** observations in each treatment	1371	560	380	27.72%
at least **5** observations in each treatment	1173	443	325	27.71%
at least **6** observations in each treatment	992	329	259	26.11%
at least **7** observations in each treatment	852	259	212	24.88%
at least **8** observations in each treatment	742	204	171	23.05%
at least **9** observations in each treatment	615	151	132	21.46%
at least **10** observations in each treatment	488	106	95	19.47%
at least **11** observations in each treatment	369	83	75	20.33%

Table A.1: The number of masses picked at different cutoff points by Kruskal-Wallis rank sum test for the aqu.neg set.

A.3.2 org.pos

	number of masses			
	after filtering	Levene's test $s.l. = 0.05$	Kruskal-Wallis test $s.l. = 0.05/M$	$\frac{\text{Kruskal-Wallis test}}{\text{after filtering}}$ %
Total	1305	1141	91	6.97%
at least **1** observation in each treatment	594	481	105	17.68%
at least **2** observations in each treatment	427	335	106	24.82%
at least **3** observations in each treatment	322	248	100	31.06%
at least **4** observations in each treatment	254	186	92	36.22%
at least **5** observations in each treatment	209	151	84	40.19%
at least **6** observations in each treatment	160	112	71	44.38%
at least **7** observations in each treatment	130	92	59	45.38%
at least **8** observations in each treatment	107	75	51	47.66%
at least **9** observations in each treatment	85	62	42	49.41%
at least **10** observations in each treatment	50	37	29	58%
at least **11** observations in each treatment	0	0	0	0

Table A.2: The number of masses picked at different cutoff points by Kruskal-Wallis rank sum test for the org.pos set.

A.3.3 org.neg

	number of masses			
	after filtering	Levene's test $s.l. = 0.05$	Kruskal-Wallis test $s.l. = 0.05/M$	$\frac{\text{Kruskal-Wallis test}}{\text{after filtering}}$ %
Total	3952	3443	122	3.09%
at least **1** observation in each treatment	2603	2212	138	5.30%
at least **2** observations in each treatment	2026	1673	152	7.50%
at least **3** observations in each treatment	1658	1325	166	10.01%
at least **4** observations in each treatment	1365	1050	177	12.97%
at least **5** observations in each treatment	1166	875	178	15.27%
at least **6** observations in each treatment	997	723	177	17.75%
at least **7** observations in each treatment	882	630	166	18.82%
at least **8** observations in each treatment	758	534	153	20.18%
at least **9** observations in each treatment	604	424	132	21.85%
at least **10** observations in each treatment	446	307	105	23.54%
at least **11** observations in each treatment	301	214	73	24.25%

Table A.3: The number of masses picked at different cutoff points by Kruskal-Wallis rank sum test for the org.neg set.

A.4 Boxplots and Interaction Plots of Masses Picked by Kruskal-Wallis Tests of the aqu.pos Set.

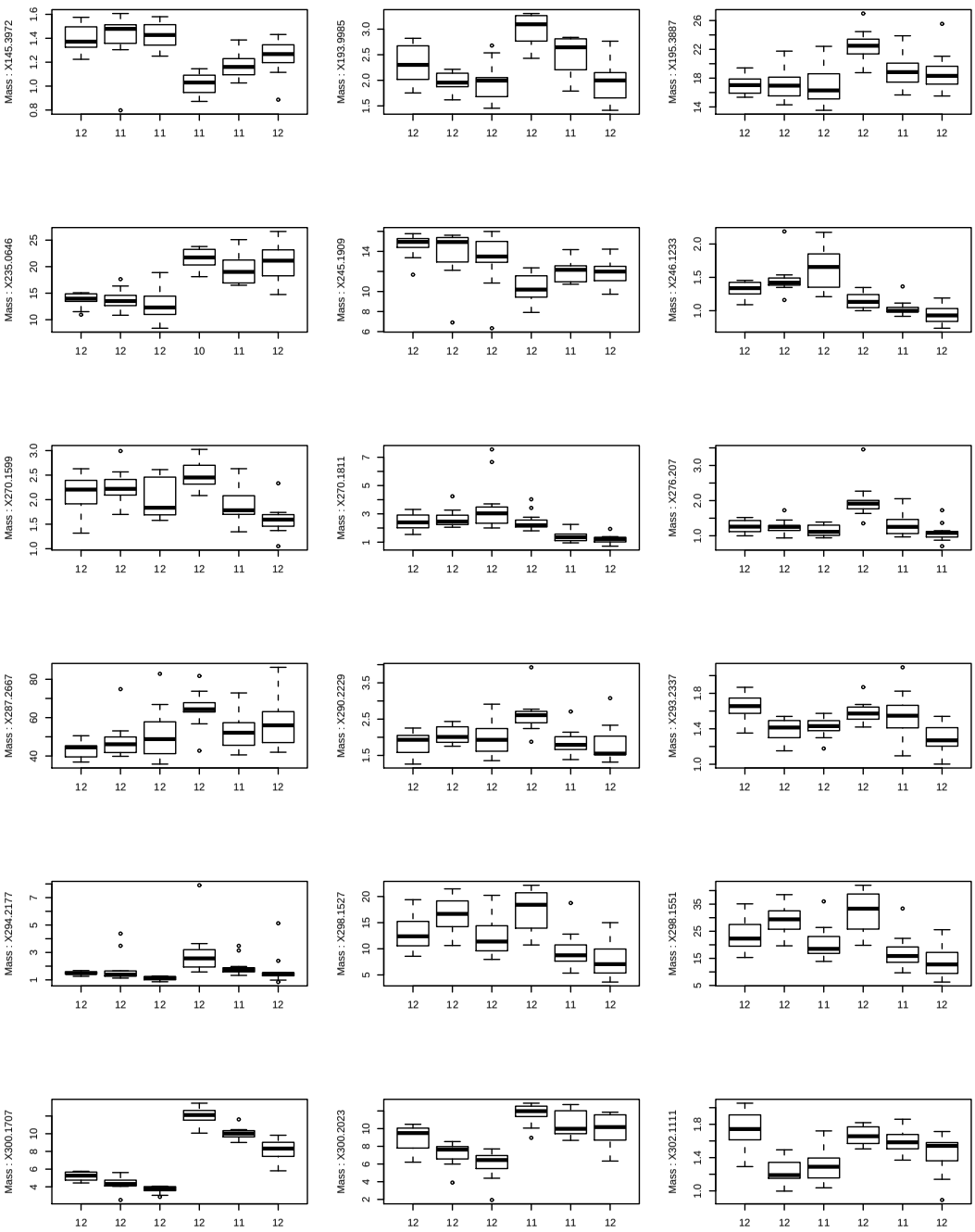

Figure A.1: (a) Boxplots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

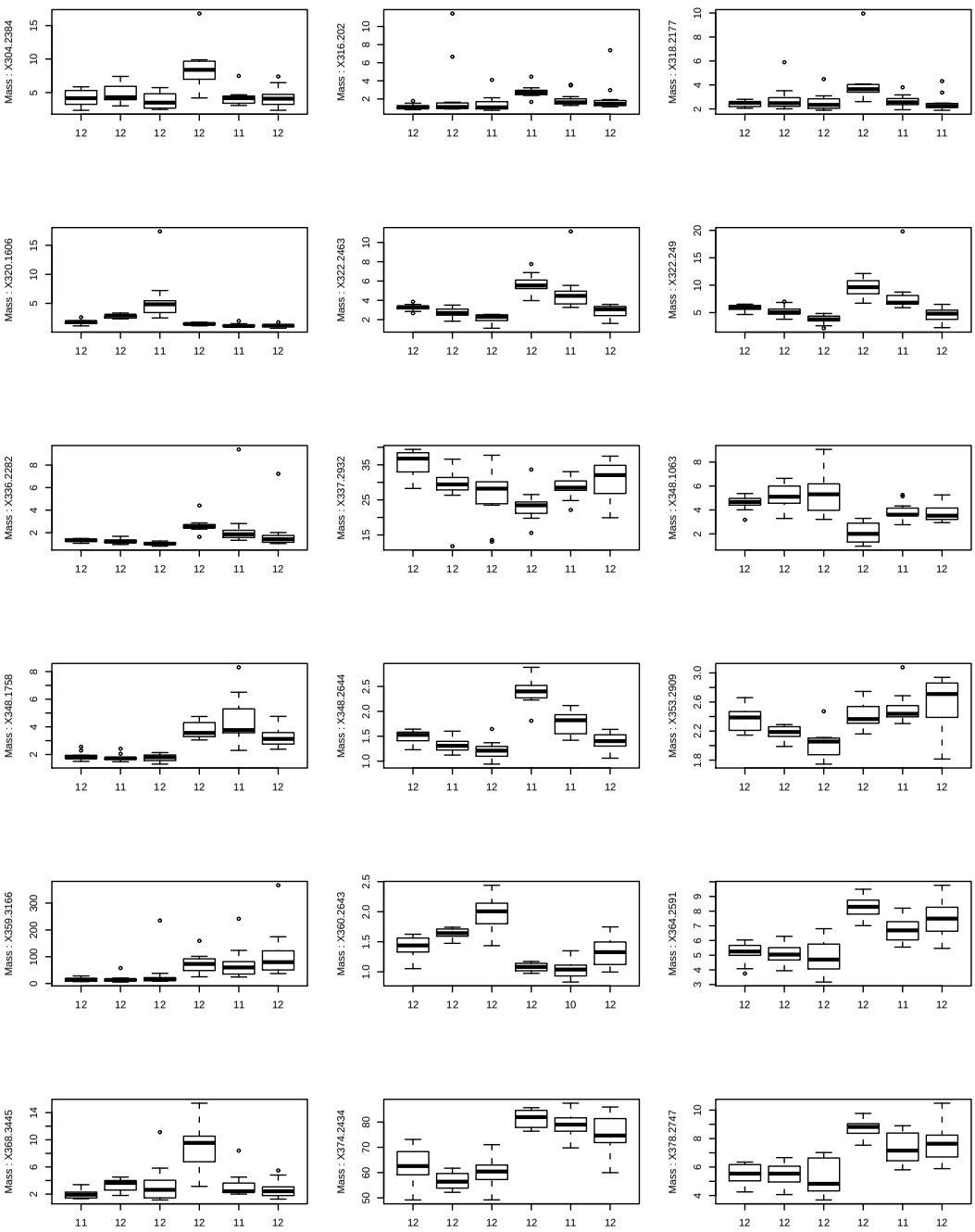

Figure A.2: (b) Boxplots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

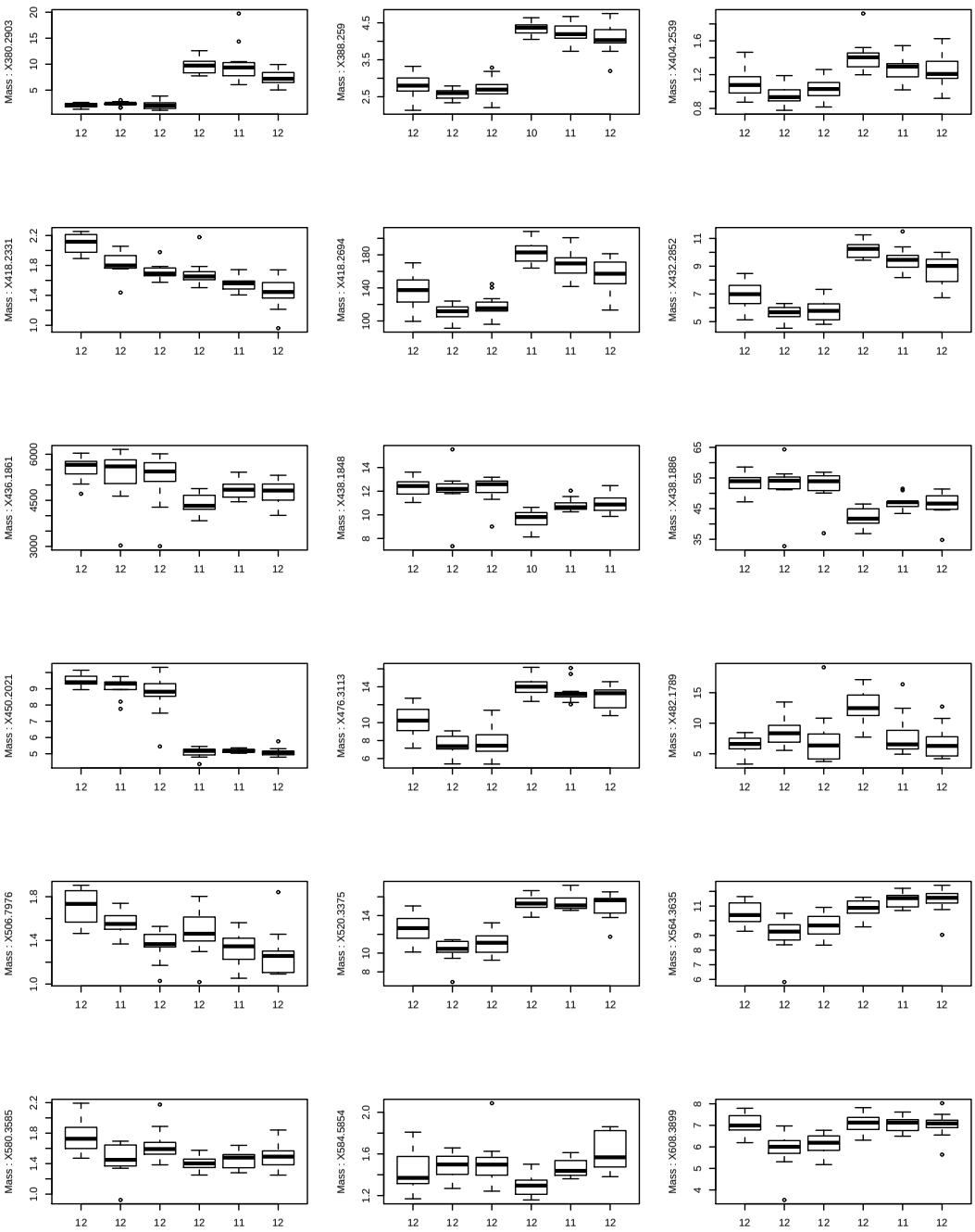

Figure A.3: (c) Boxplots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

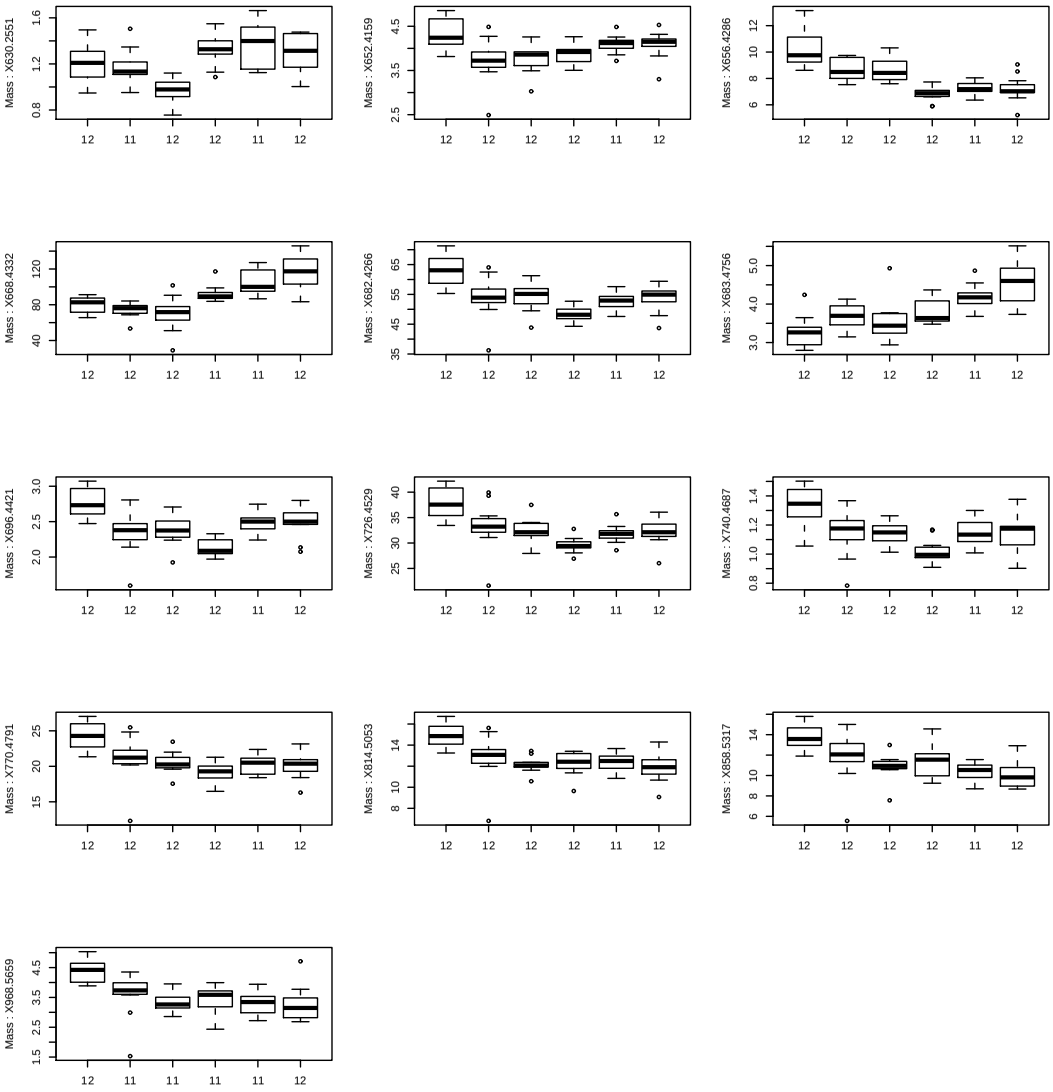

Figure A.4: (d) Boxplots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

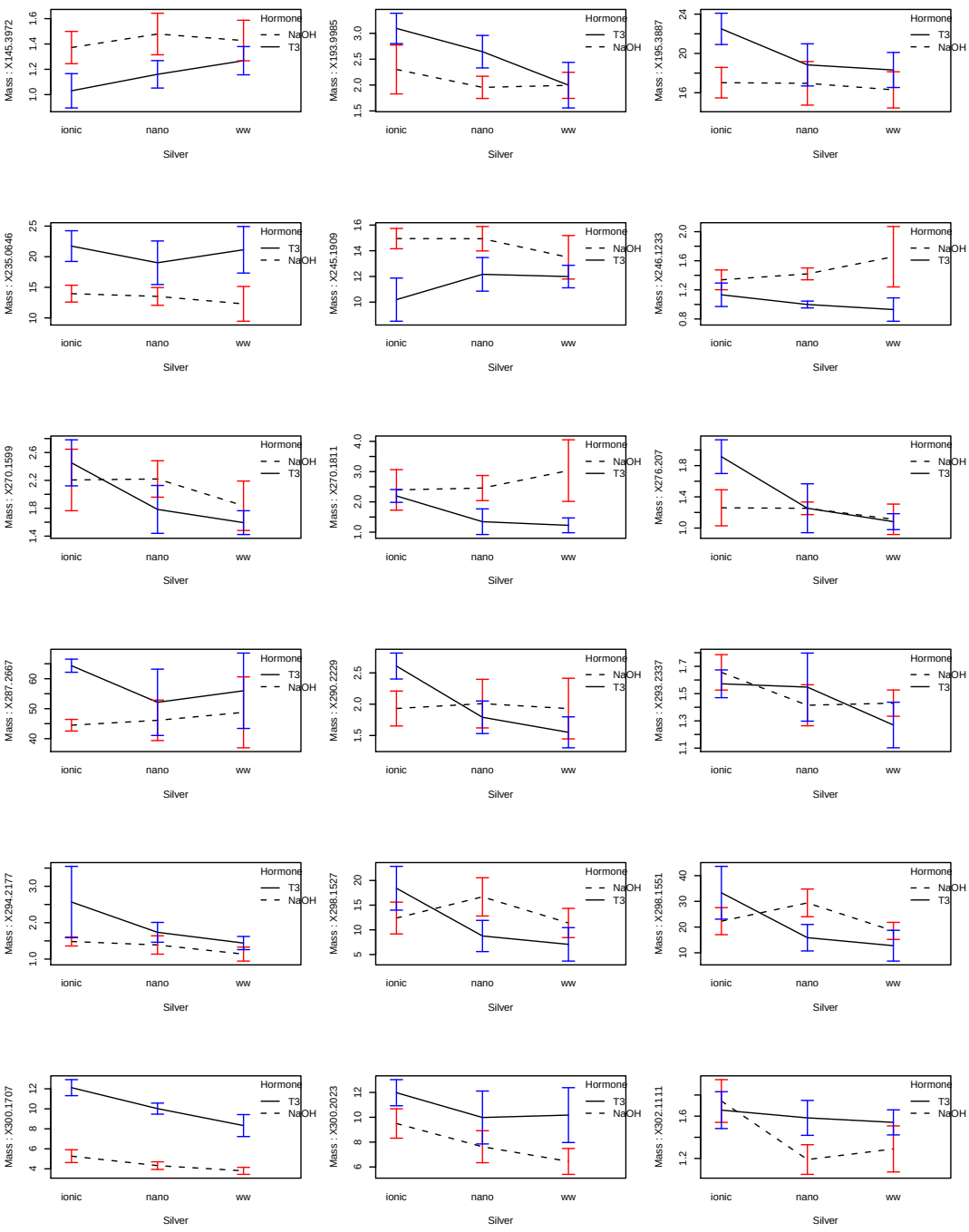

Figure A.5: (a) Interaction plots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

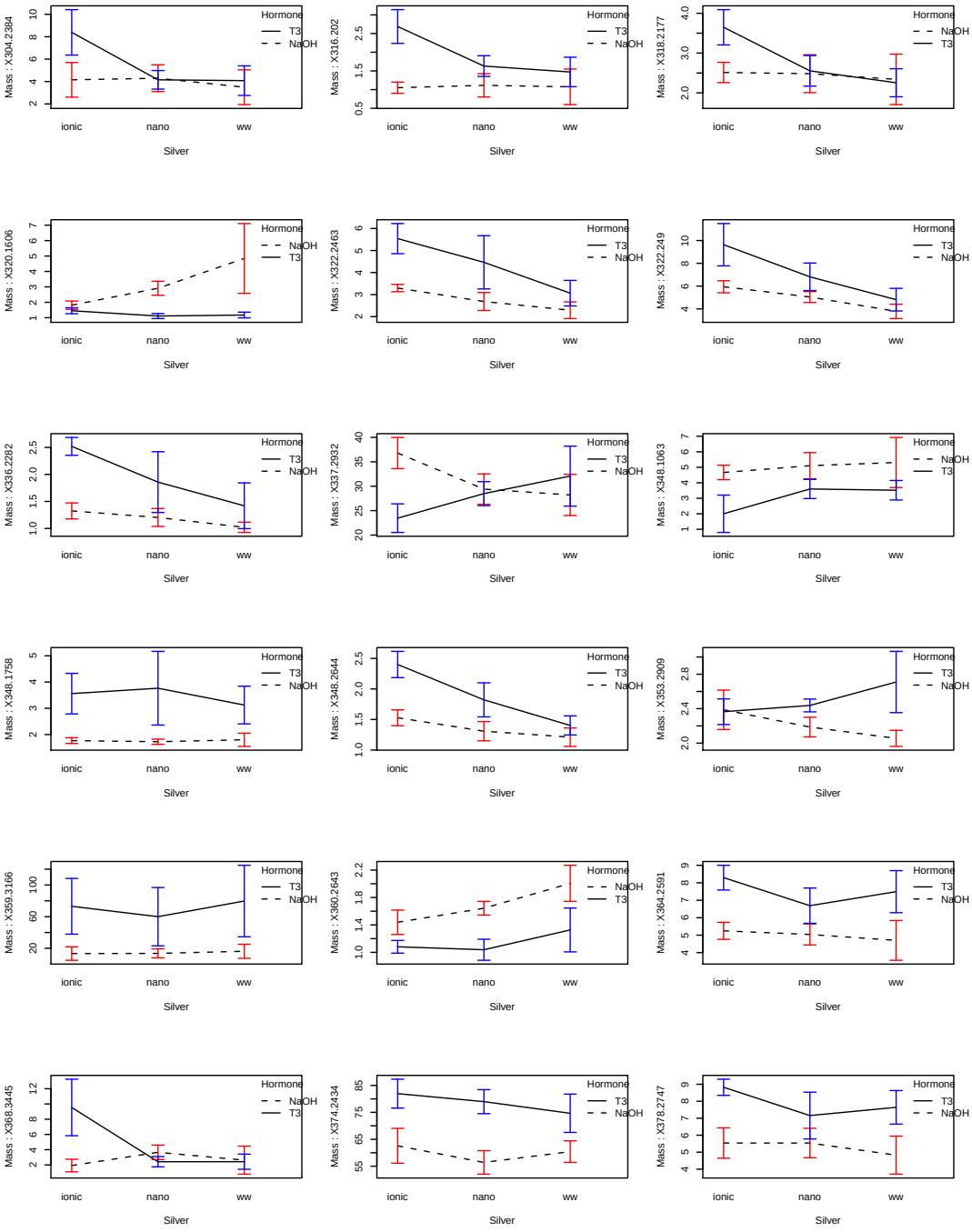

Figure A.6: (b) Interaction plots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

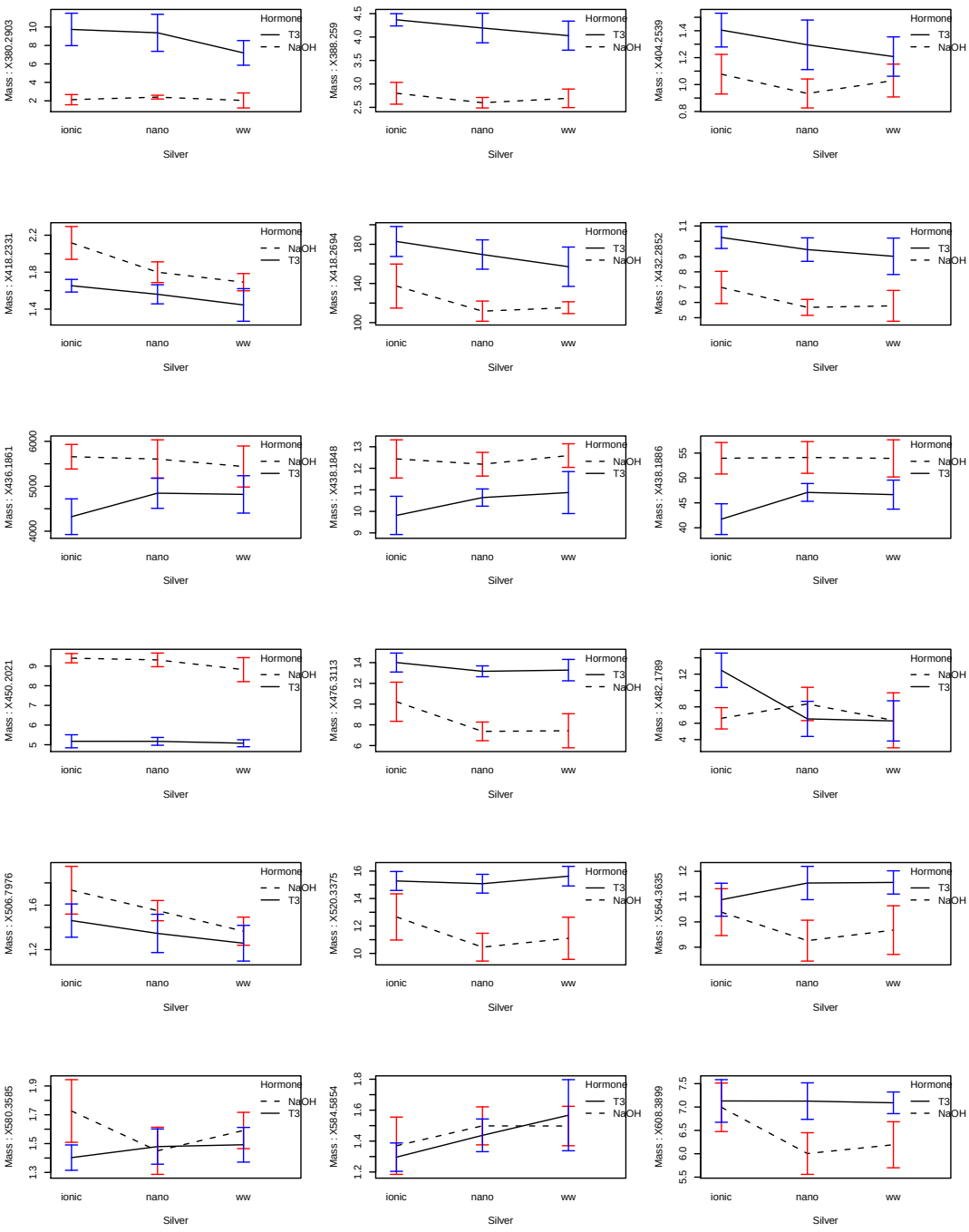

Figure A.7: (c) Interaction plots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

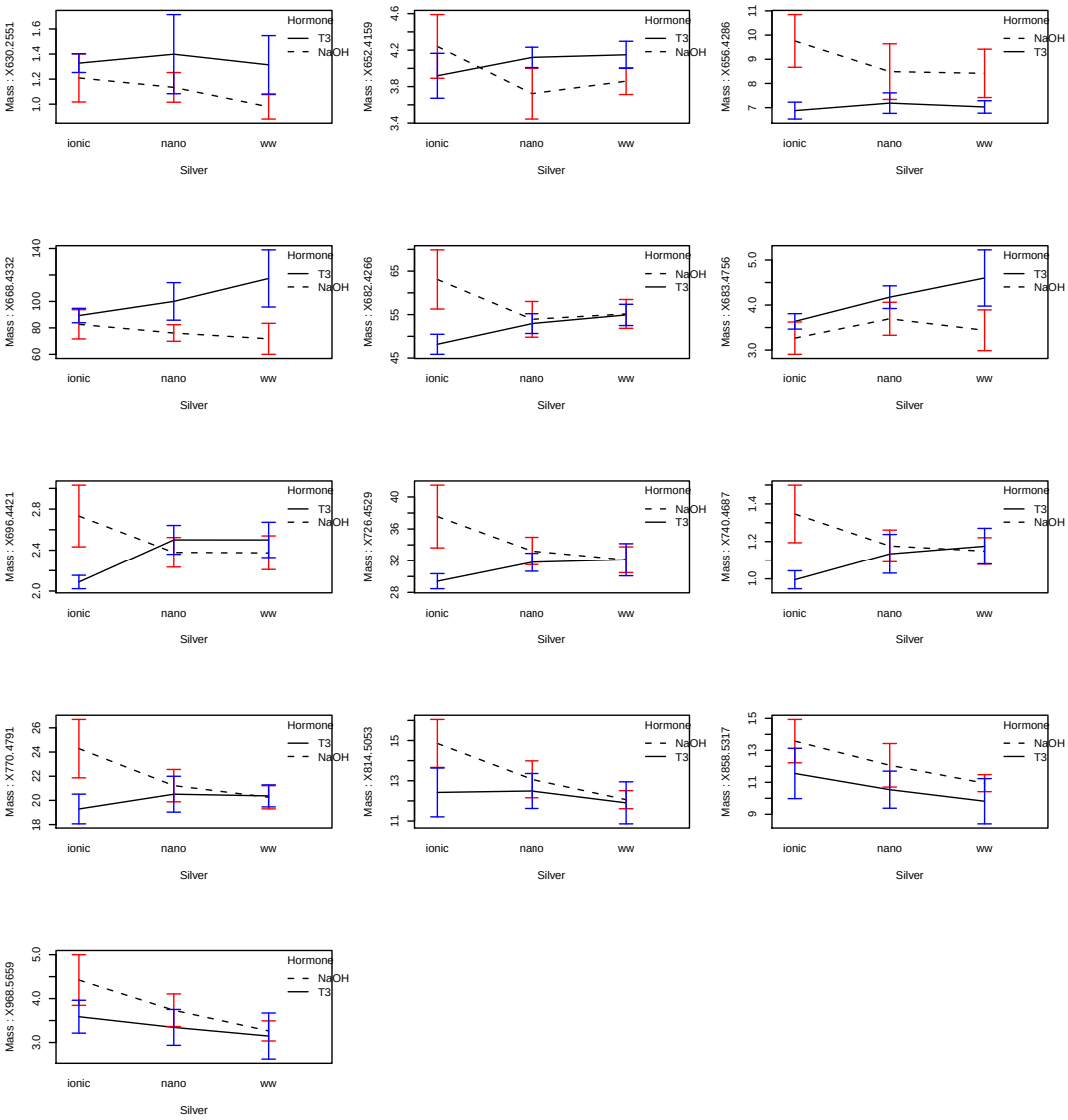

Figure A.8: (d) Interaction plots of masses in the aqu.pos set that are picked by Kruskal-Wallis tests.

Appendix B

Plots for microarray data

B.1 Boxplots and Interaction Plots of Genes Picked by Kruskal-Wallis Tests in the Microarray Sets.

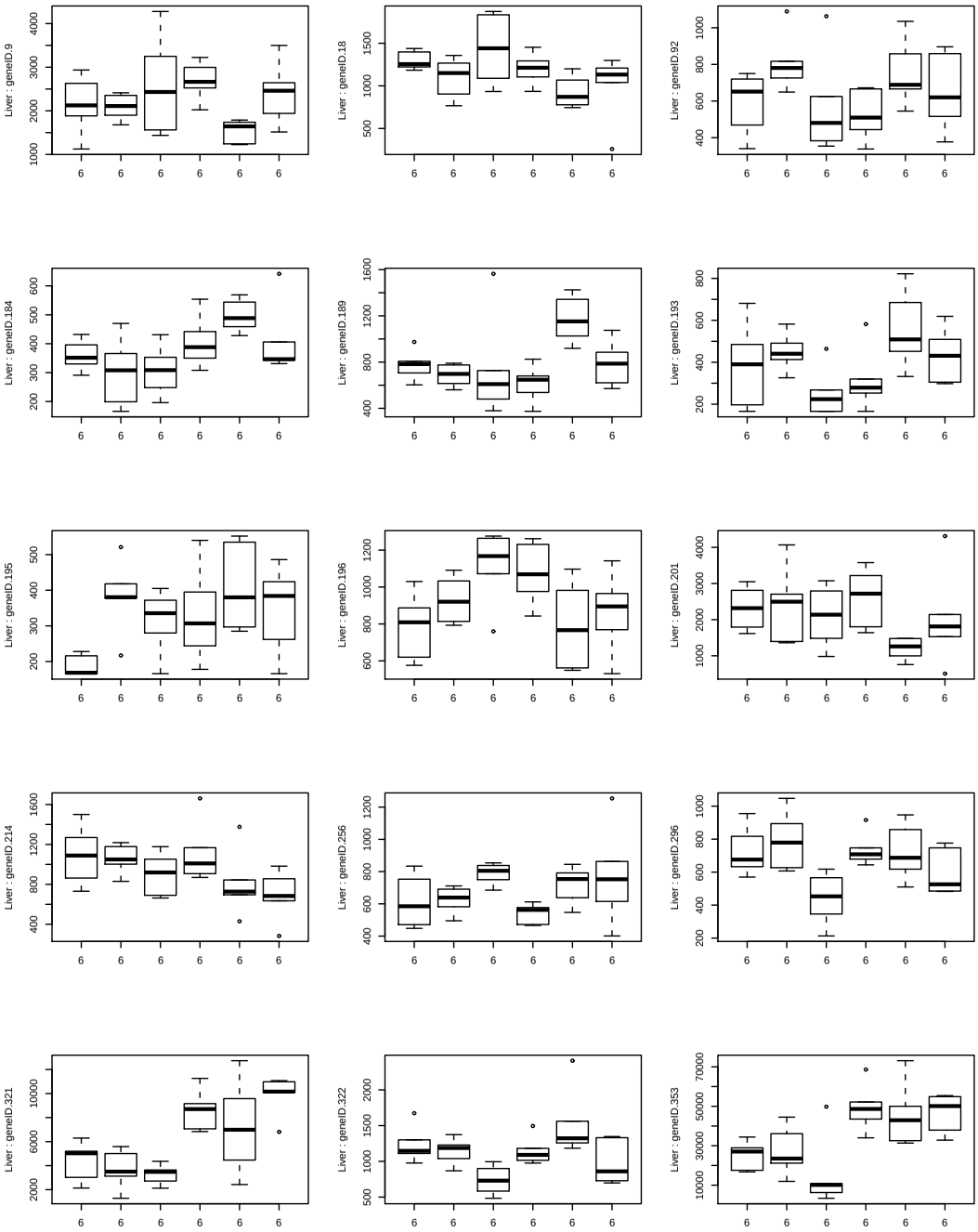

Figure B.1: (a) Boxplots of genes in the liver microarray set that are picked by Kruskal-Wallis tests.

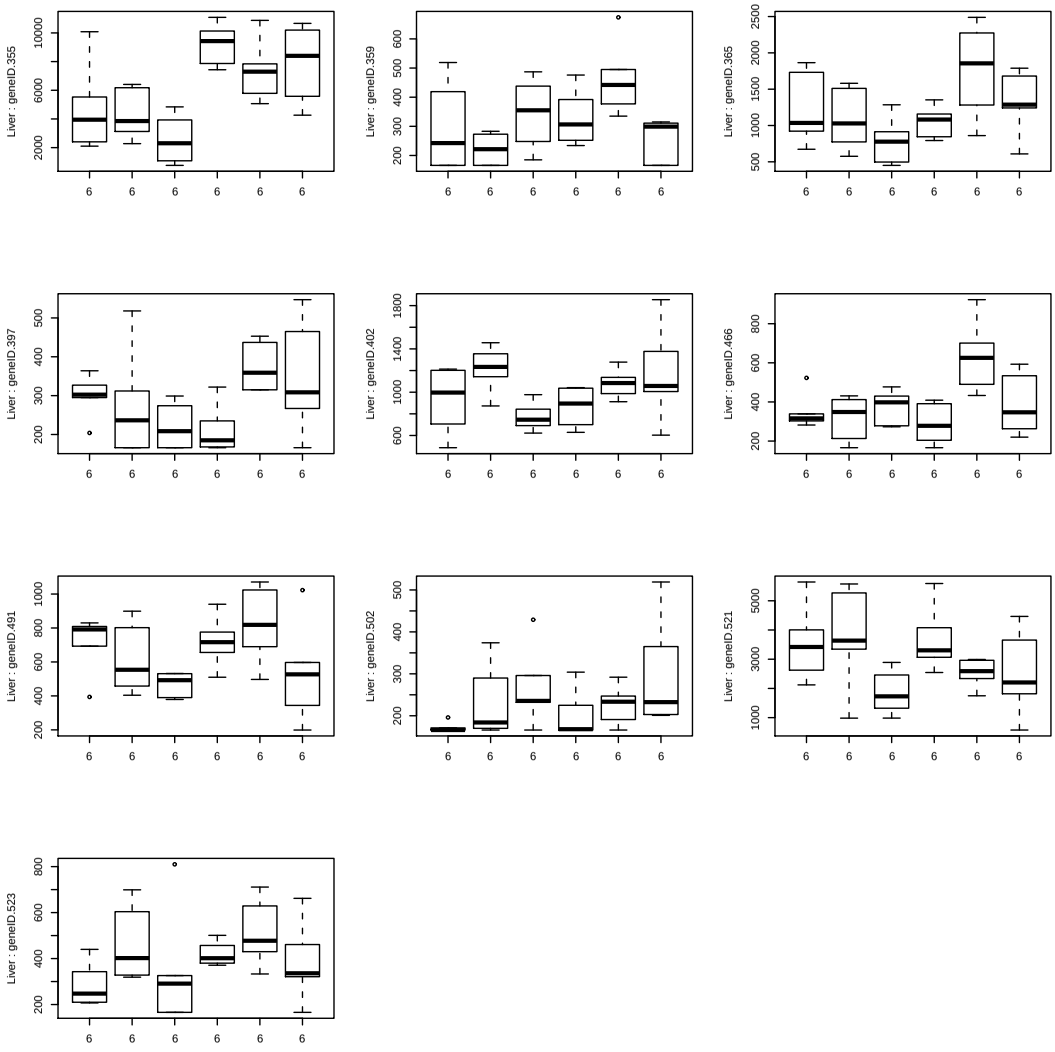

Figure B.2: (b) Boxplots of genes in the liver microarray set that are picked by Kruskal-Wallis tests.

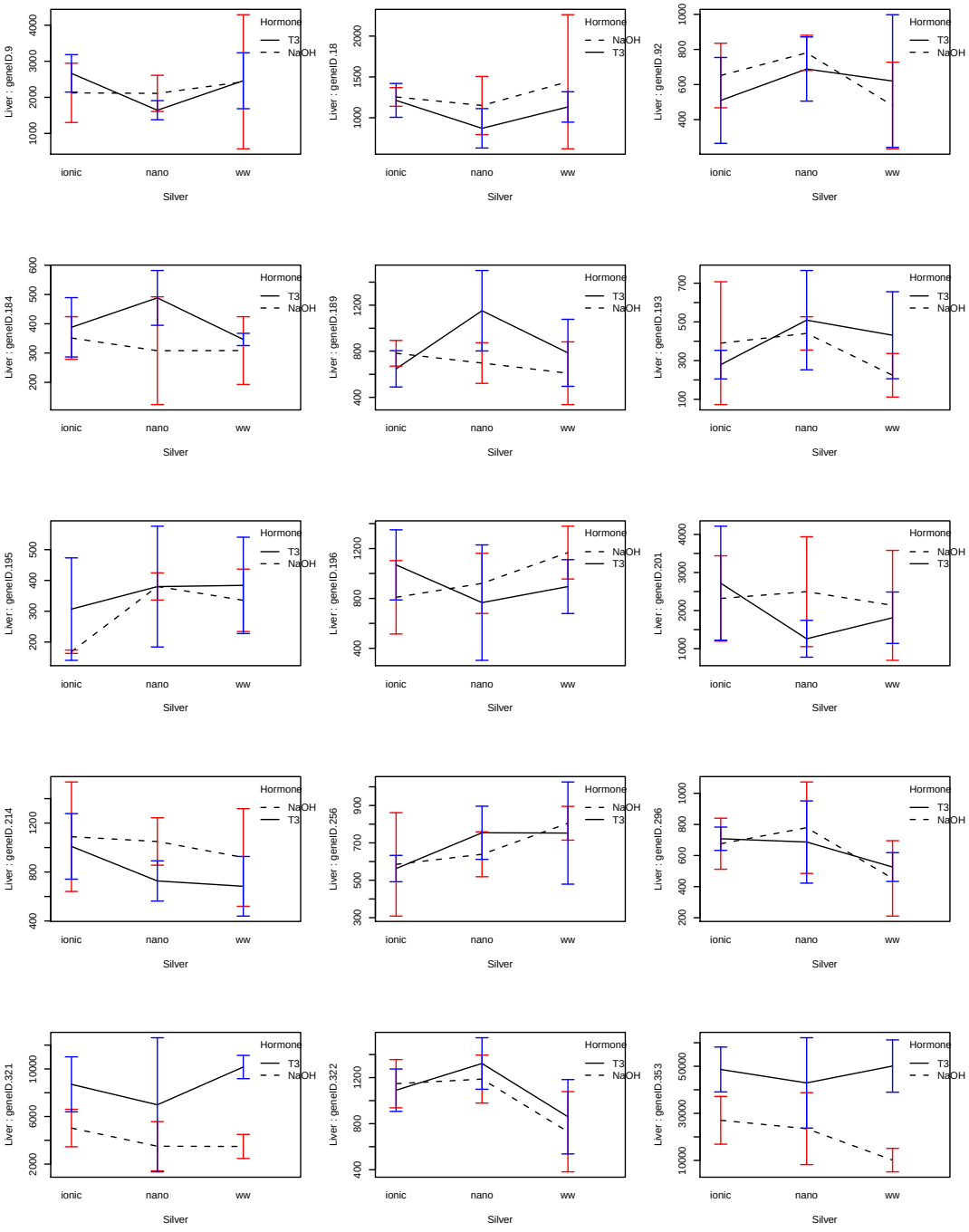

Figure B.3: (a) Interaction plots of genes in the liver microarray set that are picked by Kruskal-Wallis tests.

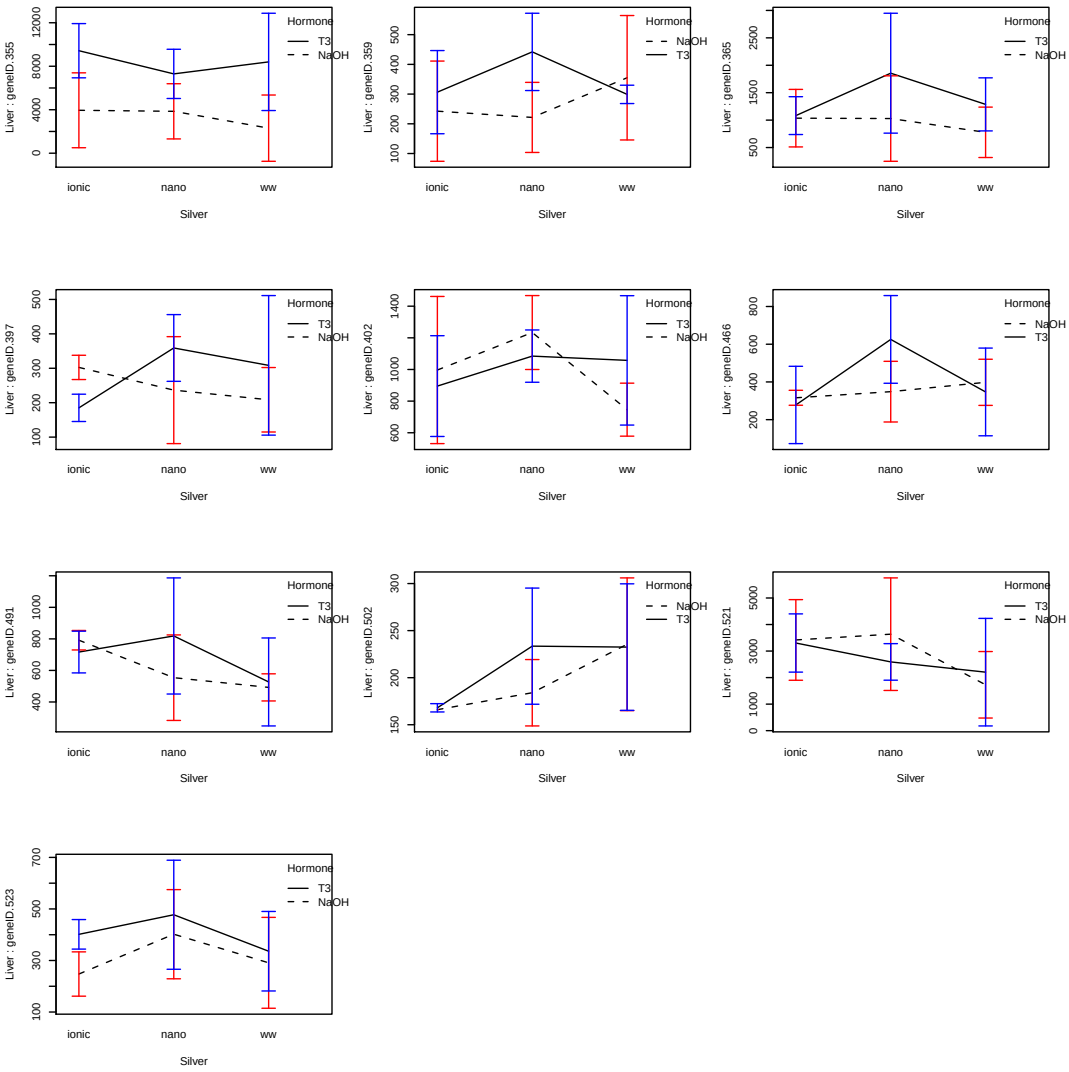

Figure B.4: (b) Interaction plots of genes in the liver microarray set that are picked by Kruskal-Wallis tests.

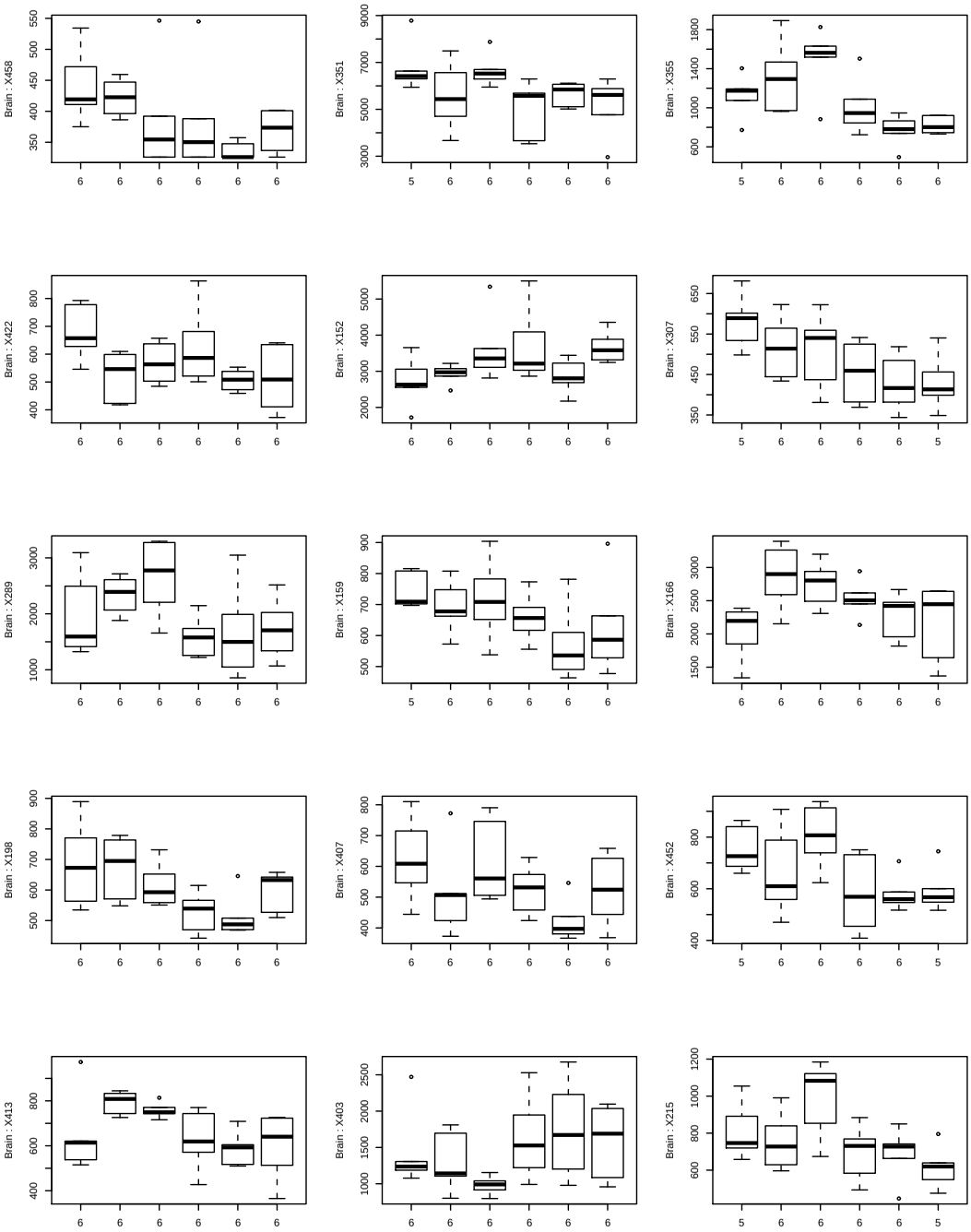

Figure B.5: (a) Boxplots of genes in the brain microarray set that are picked by Kruskal-Wallis tests.

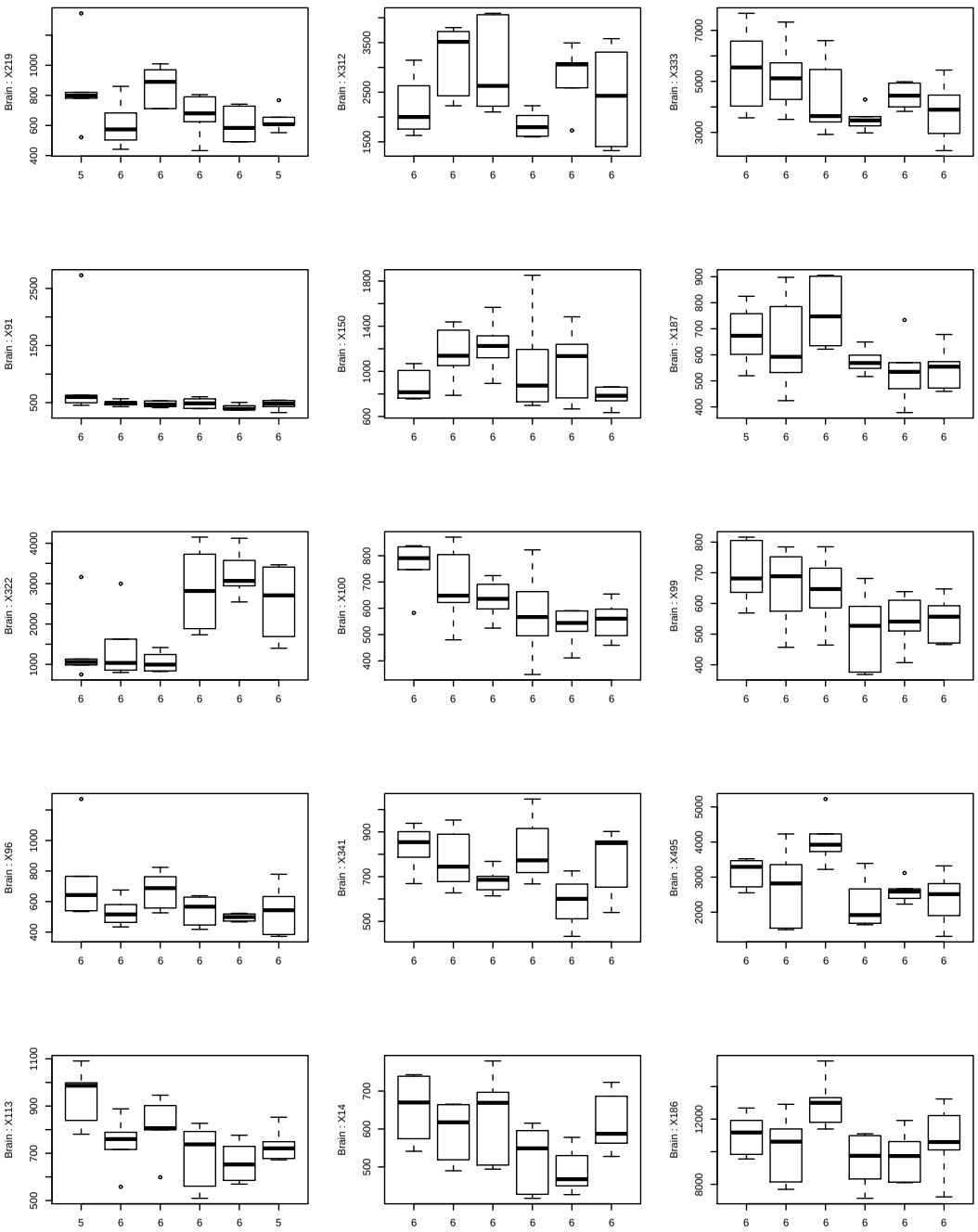

Figure B.6: (b) Boxplots of genes in the brain microarray set that are picked by Kruskal-Wallis tests.

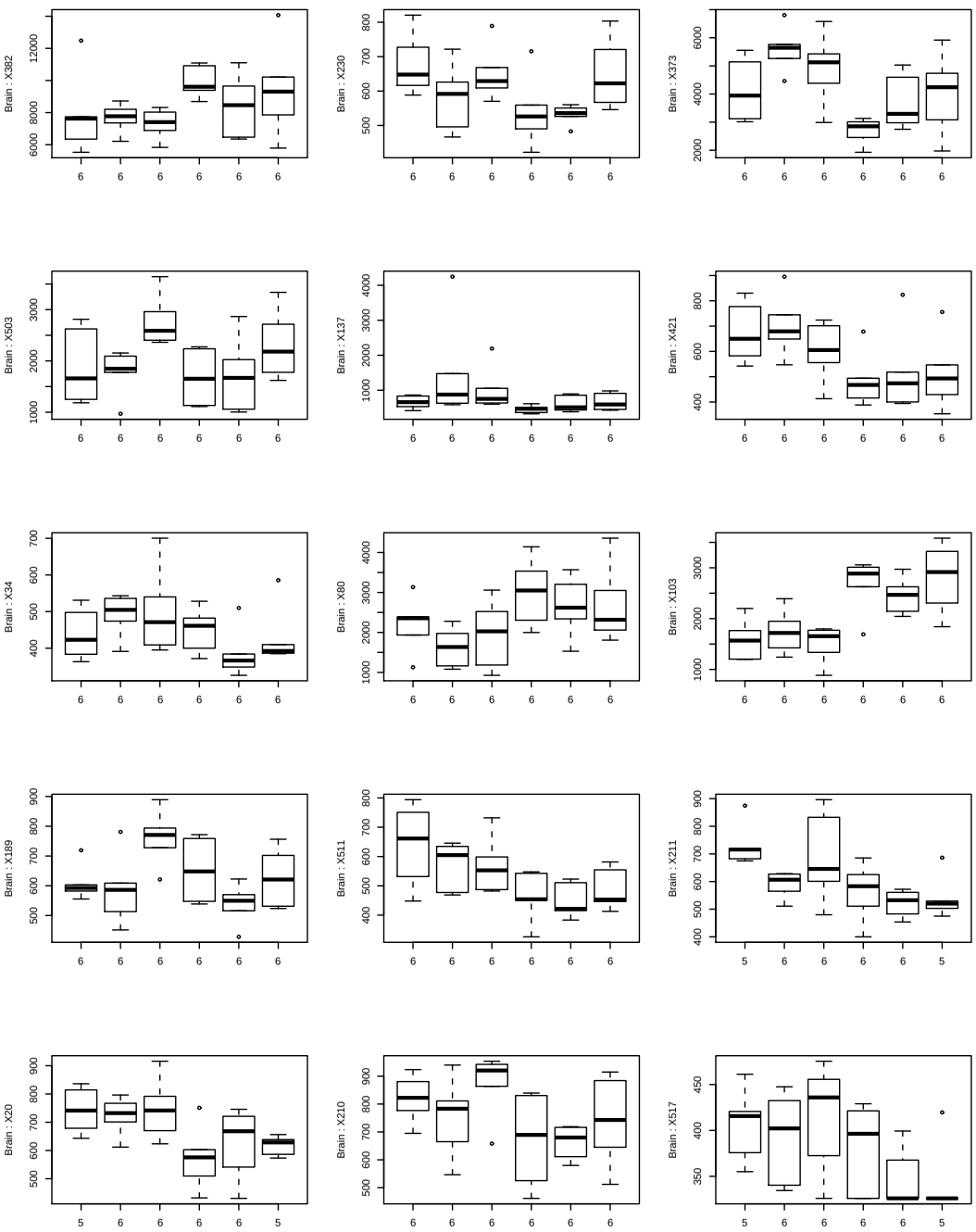

Figure B.7: (c) Boxplots of genes in the brain microarray set that are picked by Kruskal-Wallis tests.

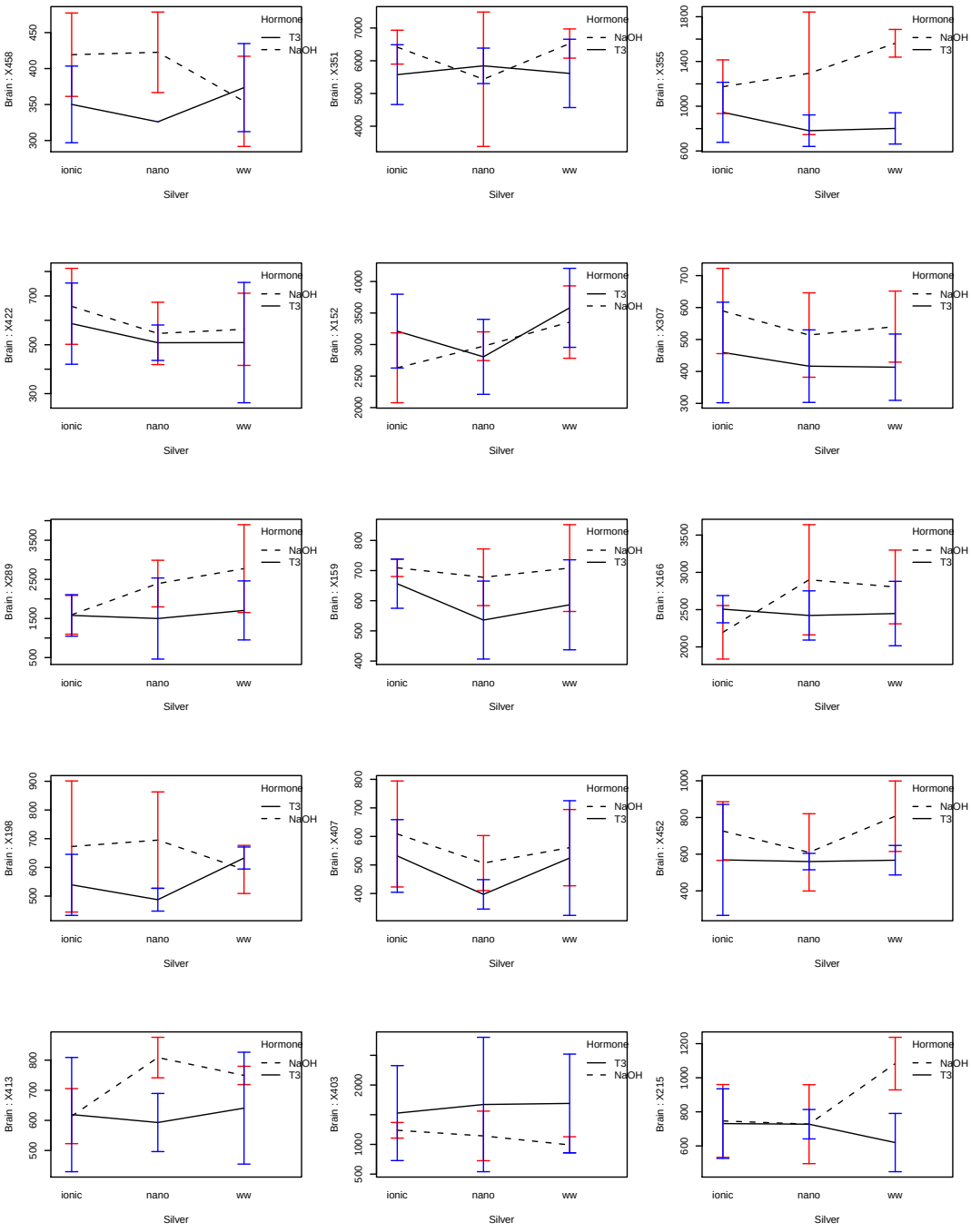

Figure B.8: (a) Interaction plots of genes in the brain microarray set that are picked by Kruskal-Wallis tests.

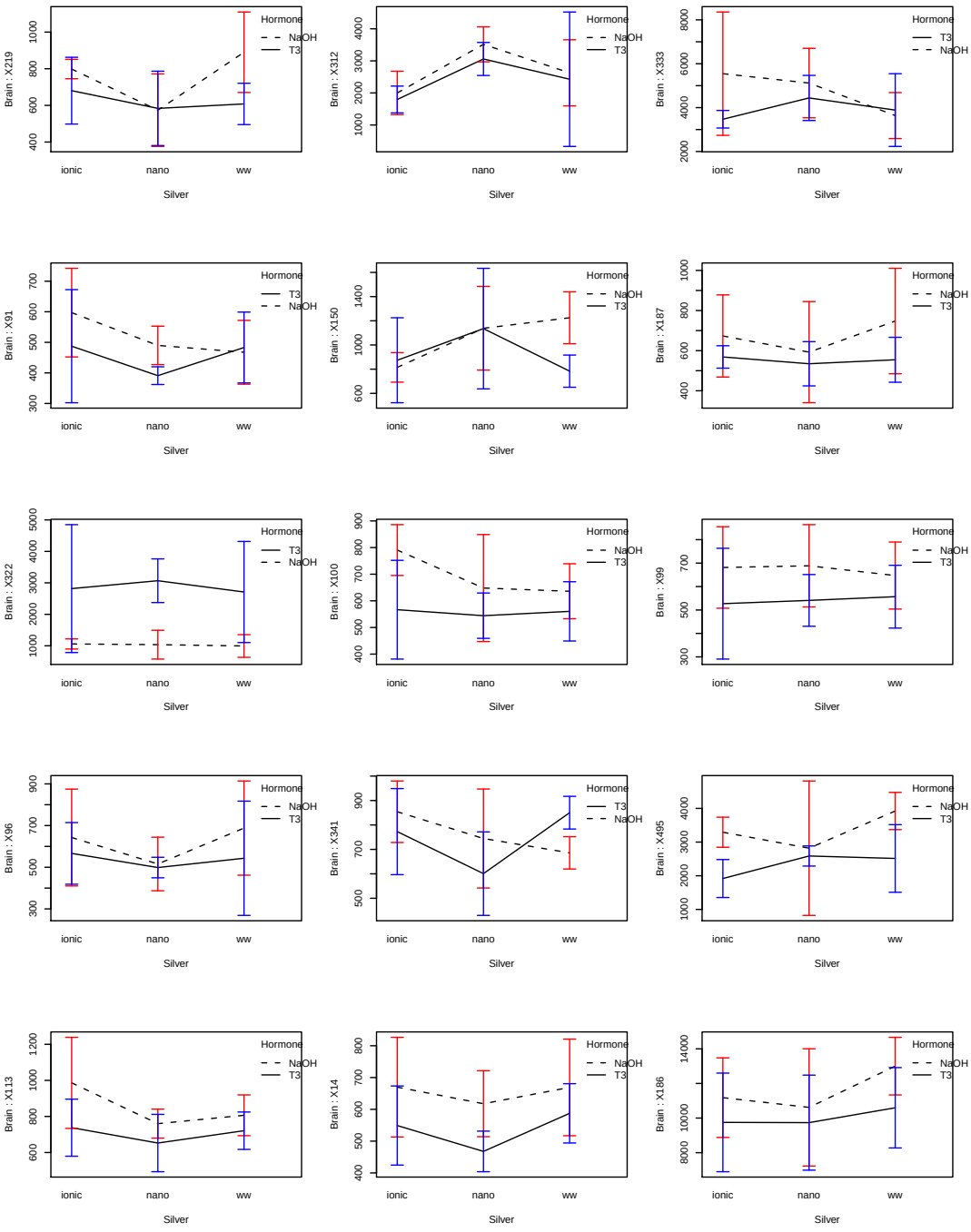

Figure B.9: (b) Interaction plots of genes in the brain microarray set that are picked by Kruskal-Wallis tests.

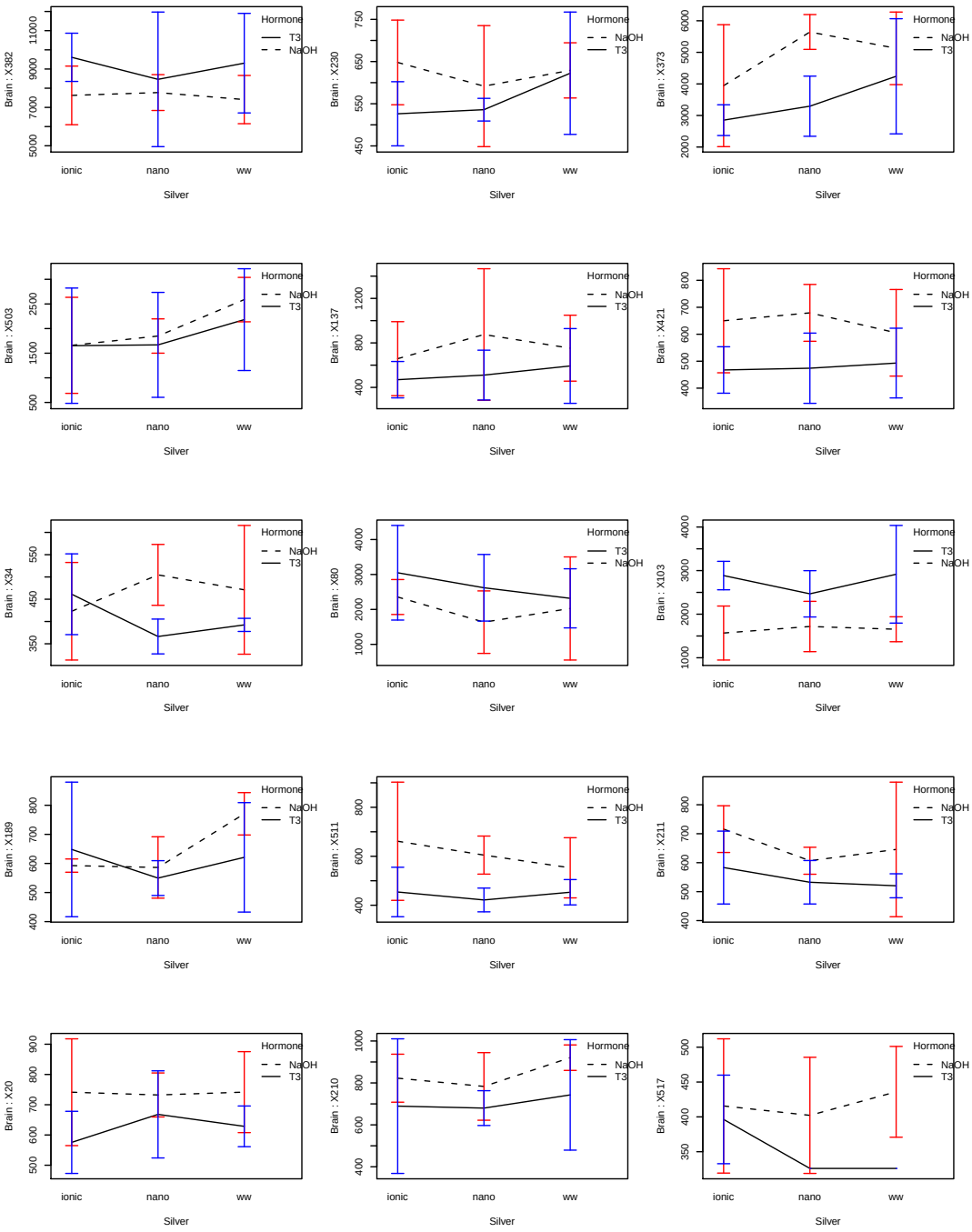

Figure B.10: (c) Interaction plots of genes in the brain microarray set that are picked by Kruskal-Wallis tests.

Appendix C

Plots for iTraq data

C.1 Boxplots and Interaction Plots of Proteins Picked by Kruskal-Wallis Tests in the iTraq Sets.

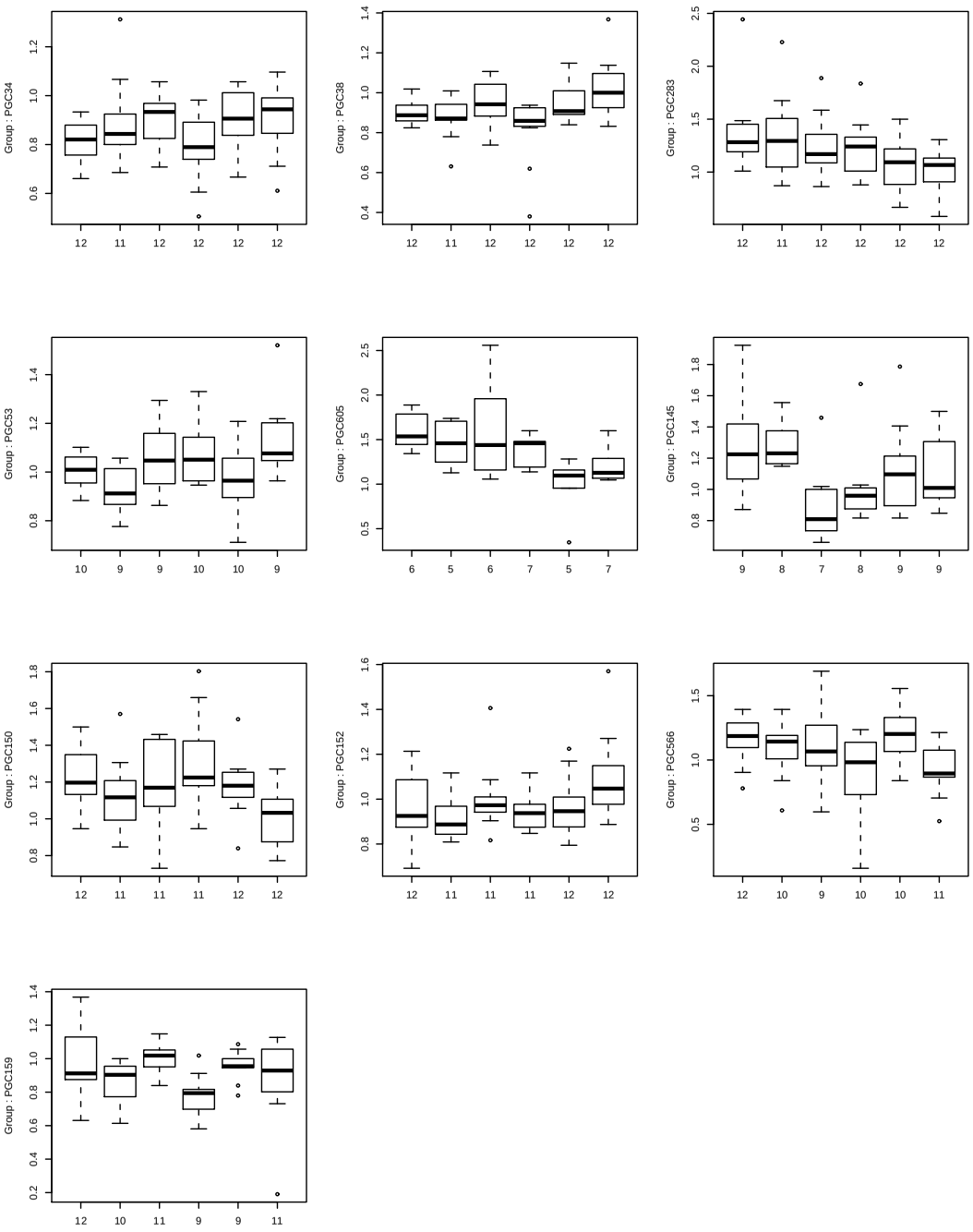

Figure C.1: (a) Boxplots of protein groups in the iTraq data set that are picked by Kruskal-Wallis tests.

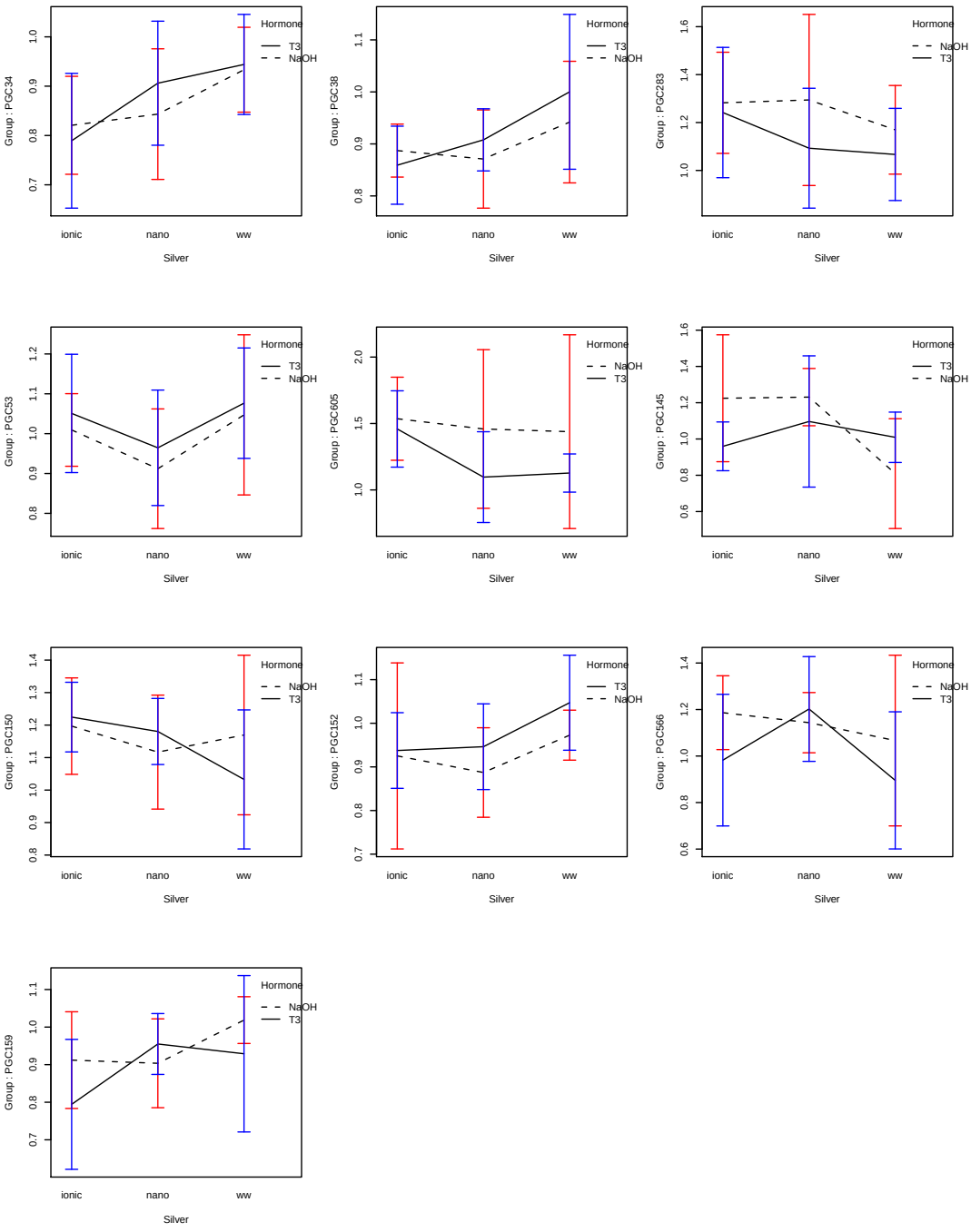

Figure C.2: Interaction plots of protein groups in the iTraq data set that are picked by Kruskal-Wallis tests.

Bibliography

[1] Abdi, H., 2007: *The Bonferroni and Šidák corrections for multiple comparisons.* In N.J. Salkind (ed.). Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage.

[2] Benjamini Y. and Hochberg Y., 1995: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.*, 57, 289-300.

[3] Brown, B.M., 1985: *Median estimates and sign tests, In: Encyclopedia of Statistical Sciences*, Vol. 5, Wiley.

[4] Carroll, R.J. and Ruppert, D., 1988: *Transformation and Weighting in Regression.* Chapman and Hall.

[5] Cho, H. and Lee, J.K., 2005: *HEM: Heterogeneous error model for identification of differentially expressed genes under multiple conditions.* R package version 1.18.0.

[6] Cohen Freue, G.V., McMaster, R., Bergman, A., Hollander, Z., Wilson-McManus, J., Balshaw, R., Keown, P., McManus, B. and Ng, R., 2005: PGCA: A New Algorithm to Link Protein Groups Created from MS/MS Data. *Bioinformatics*, 00(00), 1-7.

[7] Corder, G.W. and Foreman, D.I., 2009: *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. New York: Wiley.

[8] Davidian, M. and Giltinan, D.M., 1995: *Nonlinear Mixed Effects Models for Repeated Measurement Data*. Chapman and Hall.

[9] Denkert, C., Budczies, J., Kind, T., Weichert, W., Tablack, P., Sehouli, J., Niesporek, S., Könsgen, D., Dietel, M. and Fiehn, O., 2006: Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *CANCER RESEARCH*, 66(22): 10795-10804.

[10] Devore, J.L., 1995: *Probability and Statistics for Engineering and the Sciences*. Wadsworth Publishing, fourth edition.

[11] Gentleman, R.C. et al., 2004: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), Article R80.

[12] Good, P., 2000: *Permutation tests: a practical guide to resampling methods for testing hypotheses*, second edition. Springer series in statistics.

[13] Greenacre, M.J., 1984: *Theory and Applications of Correspondence Analysis*. Academic Press, London.

[14] Greenacre, M.J., 2007: *Correspondence Analysis in Practice*, second edition. Chapman & Hall/CRC, Boca Raton.

[15] Han, J., Danell, R.M., Patel, J.R., Gumerov, D.R., Scarlett, C.O., Speir, J.P., Parker, C.E., Rusyn, I., Zeisel, S. and Borchers, C.H., 2008: Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry. *Metabolomics*, 4(2), 128-140.

[16] Helbing, C.C., Maher, S.K., Han, J., Gunderson, M.P., and Borchers, C., 2010: Peering into molecular mechanisms of action with frogSCOPE. *General and Comparative Endocrinology*, 168: 190-198.

[17] Hochberg Y., 1988: A sharper Bonferroni procedure for multiple tests of significance. *Biometika*, 75, 800-802.

[18] Holm S., 1979: A simple sequentially rejective multiple test procedure. *Scand J Stat*, 6, 65-70.

[19] Hosmer, D.W. and Lemeshow, S., 2000: *Applied Logistic Regression*, 2nd edition. Wiley, New York.

[20] Hoyle, D.C., Rattray, M., Jupp, R. and Brass, A., 2002: Making sense of microarray data distributions. *Bioinformatics*, 18(4), 576-584.

[21] Hilbe, J.M., 2009: *Logistic Regression Models*. Chapman & Hall/CRC Press.

[22] Jolliffe, I.T., 2002: *Principal component analysis*, 2nd ed., Springer-Verlag New York, Inc.

[23] Kim, H., Golub, G.H., Park, H., 2005: Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2), 187-198.

[24] Levene, H., 1960: *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Olkin, I., et al. eds., Stanford University Press, 278-292.

[25] Liang, K.Y. and Zeger, S.L., 1986: Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

[26] Maronna, R., Martin, D. and Yohai, V., 2006: *Robust Statistics: Theory and Methods*, John Wiley & Sons.

[27] Mueller, L.N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.Y., Vitek, O., Aebersold, R. and Müller, M., 2007: SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Bioinformatics*, 7(19), 3470-3480.

[28] Myles H. and Douglas A.W., 1973: *Nonparametric Statistical Methods*. John Wiley & Sons, New York, 115-120.

[29] Nenadić, O. and Greenacre, M., 2007: Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *JSS*, 20(3), 1-13.

[30] Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I. and Ishii, S., 2003: A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16), 2088-2096.

[31] Pollard, K.S., Gilbert, H.N., Ge, Y., Taylor, S. and Dudoit, S., 2005: *multtest: Resampling-based multiple hypothesis testing*. R package version 2.2.0.

[32] Smyth, G. K., 2004: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3.

[33] Snedecor, G.W. and Cochran, W.G., 1989: *Statistical Methods*, Eighth Edition, Iowa State University Press.

[34] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B., 2001: Missing value estimation methods for DNA microarrays. *Bioinformatics*. 17(6), 520-525.

[35] Wheeler, B., 2010: lmPerm: Permutation tests for linear models. R package version 1.1-2. http://CRAN.R-project.org/package=lmPerm.

[36] Xiao, Y. and Yang, J.Y.H., 2007. *DEDS: Differential Expression via Distance Summary for Microarray Data*. R package version 1.20.0.