Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yukiko Yano Date

A Metabolomics Study of Plasma Biological Changes Related to Polybrominated Biphenyl (PBB 153) Exposure in the Michigan PBB Cohort

By

Yukiko Yano

Master of Public Health

Environmental Health

_____

Dana Boyd Barr

Committee Chair

_____

Paige E. Tolbert

Committee Member

A Metabolomics Study of Plasma Biological Changes Related to Polybrominated Biphenyl (PBB 153) Exposure in the Michigan PBB Cohort

By

Yukiko Yano

B.A., International Christian University, 2012

Thesis Committee Chair: Dana Boyd Barr, PhD

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health
in Environmental Health

2014

Abstract

A Metabolomics Study of Plasma Biological Changes Related to
Polybrominated Biphenyl (PBB 153) Exposure in the Michigan PBB Cohort

By Yukiko Yano

The accidental contamination of the food chain with polybrominated biphenyl (PBB), a brominated flame retardant, in Michigan (1973-1974) has resulted in the exposure of more than 4,000 individuals. PBBs are suspected to disrupt endocrine function, but the long-term health effects are unknown. In this study, newly collected blood plasma samples obtained from participants of the Michigan PBB registry were analyzed to investigate the long-term effects of PBB exposure on human metabolic pathways. Metabolic profiles were compared between different classes of PBB exposure, ranging from low to high body burdens of PBB, in order to identify biomarkers associated with the exposure or potential disease risk.

Metabolomic analysis was performed using liquid chromatography with Fourier-transform mass spectrometry (LCMS) on 179 plasma samples collected from subjects in the PBB registry. Metabolomic data were analyzed using various statistical and computational methods with R software package (MetBN) and the web-based tool, MetaboAnalyst. Samples were categorized into two or three classes of PBB exposure levels based on the measured plasma PBB 153 concentrations. Metabolic features (mass to charge ratios of ions and their retention time) that discriminated between different classes of PBB exposure were identified using False Discovery Rate (FDR) to account for multiple testing. The separation of the data based on these discriminatory features was visualized and assessed using principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), and hierarchical clustering with heatmaps. Selected discriminatory features were tentatively identified using MetaboAnalyst's metabolite set enrichment analysis (MSEA) and metabolic pathway analysis (MetPA).

Of the features strongly correlated with PBB concentrations, 44 features were significantly different between low and high PBB exposure classes, and 119 discriminatory features were identified from the 11,202 total features detected by LCMS. Although the true identity of these features remains uncertain, it was suggested that these discriminatory features contained metabolites involved with lipid metabolism. Results of the study demonstrate that metabolomics can be used to discover biomarkers along the exposure-disease continuum and provides a powerful means to explore the human exposome, which has been coined as a cumulative measure of environmental exposures throughout the lifespan.

A Metabolomics Study of Plasma Biological Changes Related to Polybrominated Biphenyl (PBB 153) Exposure in the Michigan PBB Cohort

By

Yukiko Yano

B.A., International Christian University, 2012

Thesis Committee Chair: Dana Boyd Barr, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health
in Environmental Health

2014

Acknowledgments

I would like to thank my advisor, Dr. Dana Boyd Barr for giving me the opportunity to take part in this research and guiding me through every step of the way. This work would not have been possible without the strong support from Dr. Dean P. Jones, Douglas Walker, Vilinh Tran, and Dr. Karan Uppal from the Clinical Biomarkers Laboratory at the Emory University School of Medicine, Division of Pulmonary, Allergy, and Critical Care Medicine. I would also like to thank Dr. Michele Marcus and Elizabeth Marder, along with other members of the Michigan PBB Registry research team in the Department of Epidemiology at Emory University's Rollins School of Public Health. Lastly, I would like to express my gratitude to all the study participants of the Michigan PBB registry.

Funding for this research was provided by NIEHS (5 R01 ES012014-10).

Table of Contents

1. Introduction

Environmental exposures play an overwhelmingly important role in many of the chronic diseases that constitute the majority of the disease burden in the world today.[1,2] Although the development of chronic diseases can be attributed to both genetic and environmental risk factors, current research shows that 70 – 90% of cancer as well as autoimmune, cardiovascular, pulmonary, and other chronic diseases are attributable to environmental factors.[3-5] Despite this, previous research has predominantly focused on genetic factors, and the complex interplay between environmental exposures and genetic factors is only now beginning to be addressed. With the success of the Human Genome Project, new high-quality technologies have been developed to conduct genome-wide association studies that have revealed genetic associations and networks related to disease etiology. However, it soon became apparent that genetics alone can only account for a fraction of disease risk.[6] Recognizing this limitation, Wild introduced the concept of the "exposome" in 2005, which encompasses the totality of life-course environmental exposures (e.g. chemical exposures, human behavior, diet, and lifestyle choices such as smoking and alcohol intake) from conception onwards, as a complement to the genome.[1] More recently, the exposome definition was refined as "the cumulative measure of environmental influences and associated biological responses throughout the lifespan, including exposures from the environment, diet, behavior, and endogenous processes."[6]

By accounting for the temporal and multifactorial aspects of disease etiology, the exposome has the potential to establish important risk factors for the prevention and treatment of chronic diseases in the future. Unlike the genome, the exposome is a dynamic entity that evolves throughout the course of an individual's lifetime. The exposome focuses on individual longitudinal exposure profiles, specifically on the temporal accumulation of exposure, thereby considering individual exposures as dynamic and continually changing over time.[7] There is strong evidence that chronic diseases are influenced by exposures early in life, and that there are critical

periods in life stages during which an exposure has a stronger effect or induces long-term effects that are carried on into later life.[8] Temporal variation in exposure is a critical feature of the exposome, and the exposome can provide an integrated view of these critical periods, including in utero and early childhood, puberty, adulthood and old age.[7,9] In addition, the exposome provides a way to shift away from the conventional reductionist approach which focuses on one-to-one relations between a specific exposure and disease, towards a more systematic approach capturing the complex association between various risk factors and disease. Chronic diseases are multifactorial in etiology, meaning that a given disease can be caused by more than one causal mechanism and with each causal mechanism involving the combined effect of a multitude of component causes.[10] Most current epidemiologic studies are limited in their ability to capture this multifactorial nature of chronic disease etiology since they tend to concentrate on particular category of exposures separately. Rather than focusing on a specific exposure category such as air and water pollution, occupation, diet and obesity, stress and behavior, or types of infection, the exposome proposes an approach that investigates these exposures holistically, thereby providing a more cohesive view of environmental exposures.[3]

Characterizing the human exposome may seem overwhelming, but recent advances in "omics" technologies and methods offer an unprecedented opportunity to investigate complex human exposure profiles by identifying important biomarkers. Omics refers to the quantitative measurement of global sets of molecules in biological samples using high-throughput techniques, in combination with advanced biostatistics and bioinformatics tools.[11] Omics technologies permit the rapid, parallel analysis of hundreds or thousands of these sets of molecules including profiles of gene transcripts (transcriptomics), proteins (proteomics), and metabolites (metabolomics).[9] Metabolomics entails the measurement of biochemical metabolites (e.g., lipids, sugars, amino acids, etc.) in a biological sample, such as urine or blood, using high-throughput techniques to simultaneously detect all metabolic products in a sample extract, thereby providing a metabolic profile. Since metabolites are the most downstream expression of biochemical pathways,

metabolic profiles represent an integration of the interplay between various upstream factors such as underlying genetics, diet, and the environment.[7,9] By comparing metabolic profiles of different biological states, particularly those represented by diseased and healthy populations, it is possible to identify features (*m/z* or mass to charge ratios of ions and their retention time), or biomarkers, that consistently discriminate according to disease status.[12] A prospective search for intermediate biomarkers that are elevated or attenuated in subjects who eventually develop disease, or a retrospective search for links of intermediate biomarkers to past environmental exposures can be conducted with metabolomics.[8] Such intermediate biomarkers reflect intersecting regions of markers of exposure and disease outcome along the exposure-disease continuum, and thus relate to the effect of exposure (biomarker of exposure) or impact disease risk (biomarker of disease).[12,13] The discovery of biomarkers is essential for characterizing health-impairing exposures and investigating disease causality, and further to offer new approaches for the prevention, diagnosis, and treatment of diseases.

The strength of omics techniques is in its ability to capture a wide range of exposures in a single measurement using only small volumes of biological samples. Today, there are more than 84,000 chemical substances used in the market, and more than 1,000 new chemicals are introduced into the environment every year.[4,14] Humans are exposed to these chemicals every day, and the United States Centers for Disease Control and Prevention has measured ~250 environmental chemicals in human biological fluids so far using a targeted chemical approach.[15] However, such targeted investigations that focus on selected hazardous exposures known a priori are time-consuming and impractical for measuring all existing chemicals in the environment. These targeted studies are driven by a bottom-up, reductionist approach which start with a knowledge base of specific exogenous chemicals measured in air, water, or the diet, thus employing a knowledge-driven investigation. Although this approach has the benefit of being able to unequivocally identify and quantify these known toxicants, it fails to explore unrecognized exposures and also neglects the complex interactions between environmental and

biological systems that are inherently in a constant state of flux.[3] In contrast, untargeted, top-down studies initially measure all chemicals in a biological sample without any preconceived expectations and then narrow them down to potentially important exposures through a data-driven investigation.[5] Although the top-down approach cannot unequivocally identify most chemicals measured, this approach allows for a thorough investigation of all exposures (some known and some unknown) represented in the sample extract and compares their relative importance as risk factors for disease. With exposomics, it is possible to conduct both targeted and untargeted analyses (or a combination of the two in a hybrid analysis) and explore both known and unknown exposures. Untargeted metabolomics analyses are used to compare metabolic profiles widely and identify discriminating features to be chosen as candidate biomarkers, whereas targeted metabolomics analyses are used to confirm and validate these biomarkers by matching with authentic standards or to determine exposure categories for known toxicants. In this way, metabolomics allows for a complete analysis of environmental exposures by integrating both known and unknown components of the exposure profile, thereby making it an optimal tool for exploring the human exposome.

In the present study, a metabolomics analysis is conducted to study serum biological changes related to polybrominated biphenyl (PBB) exposure in an exposed Michigan PBB Cohort. PBB belongs to the largest class of flame retardants used commercially[16], and the chemical was used in the United States as a flame retardant in products such as plastics, textiles, and electronics in the 1970s. However, the US manufacture of PBB was discontinued in 1976 in the aftermath of a large-scale contamination accident that occurred in Michigan. The Michigan Chemical Company made two products at the same plant in the 1970s: FireMaster, a flame retardant mixture of PBBs and NutriMaster, a feed-grade magnesium oxide supplement for cattle. In 1973, the company accidentally shipped PBB instead of the magnesium oxide supplement to the Farm Bureau Services where it was mixed into livestock feed that was shipped to feed mills across the state. The PBB-contaminated feed was ingested by cattle, pigs, chickens, and ultimately residents

in Michigan through consumption of contaminated milk, meat, eggs, and other animal products between 1973 and 1974.[17] The first health problems were observed in dairy cattle in the fall of 1973, but the accidental PBB contamination of animal feed was not identified as the cause of the problem until the spring of 1974.[18] As a result, most Michigan residents who have been tested have detectable levels of PBB in their serum[17], many exceeding the US reference range value for PBB. High PBB concentrations (i.e., > 95th percentile level of the National Health and Nutrition Examination Survey [NHANES] which is an estimated 16.3 pg/mL or 0.016 parts per billion assuming an average serum lipid concentration of 600 mg/dL)[15] have also been detected among some families residing on farms that had received the contaminated feed and in neighboring families who purchased food from these farms. In order to assess potential long-term effects of PBB exposure on human health, the Michigan Department of Community Health and the US Public Health Service established a registry of exposed Michigan residents in 1976, named the Michigan Long-Term PBB Study.[19,20] The registry initially consisted of approximately 4,000 individuals with a wide range of PBB exposure levels and now also includes the offspring of those who were exposed.[21] This registry provides an opportunity to prospectively assess not only the direct effects of human exposure to PBB but also its potential transgenerational effects.

PBBs are classes of structurally similar brominated hydrocarbons in which 2 to 10 bromine atoms are attached to a biphenyl. Although there are 209 different molecular combinations, or congeners, that are possible for PBBs, only a subset of these exists in commercial mixtures. The isomer 2,2',4,4',5,5'- hexabromobiphenyl (PBB 153) was the dominant constituent of FireMaster, which was the source of the contamination in Michigan.[17] Since there are no known natural sources of PBBs in the environment and PBBs are no longer produced in the United States, the general population exposure to PBBs will only be from past releases typically bioaccumulated in the food chain. However, the long-term health effects of PBB exposure continue to be a concern, particularly among Michigan residents who were exposed, because of its long environmental persistence and potential toxic effects.[20] Although

various symptoms (e.g., nausea, abdominal pain, loss of appetite, joint pain, fatigue, and weakness) have been reported since the accident, little convincing evidence exists to definitively link PBB exposure to adverse health effects among Michigan residents. However, animal studies have shown evidence of body weight loss, skin disorders, and nervous system effects, along with damages to the liver, kidneys, thyroid gland, and immune system. Reproductive, developmental, and carcinogenic effects have also been observed from animal studies.[18] In addition, there is evidence that PBBs have endocrine disrupting properties (i.e., mimic or antagonize endogenous hormonal activity) in humans such as its influence on the secondary sex ratio, which is the ratio of males to females at birth.[17] Exposure to high levels of PBB has also been found to be associated with transgenerational endocrine effects such as menarche at an earlier age among girls[22] and congenital urinary tract problems among boys (e.g., hernia, hydroceles, cryptorchidism, hypospadias, varicocele)[21].

Long-term exposure to PBBs is of particular concern since they tend to build up in the human body over many years. Once PBBs enter the human body, they are partially metabolized and may be excreted from the body within a few days, mainly in feces and in very small amounts in urine. However, highly brominated congeners tend to undergo little or no metabolic transformation. No major metabolites were identified in the feces or urine of animals treated with PBB 153, suggesting that this congener is stable and persistent.[18] PBBs may remain in the body for many years since they are mainly stored in body fat and tend to concentrate in breast milk fat. The estimated half-life of PBBs in human blood sera and fat are 10.8 years[23] and 7.8 years[24], respectively, meaning that PBBs may remain in tissues throughout the lifetime of those exposed. Moreover, it is possible for children to be exposed to high doses of PBBs through breast feeding, and children can also be exposed in utero through the placenta.[18] This raises concerns for the potential transgenerational effects of PBB exposure on the development and growth of the children of those who were exposed in Michigan.

Biomarkers are useful tools to identify subjects who have been exposed to PBBs and detecting early signs of disease development among those exposed. Several studies have shown that both serum and adipose PBB levels are reliable biomarkers of exposure, but there are no specific biomarkers of effects for PBBs.[18] Levels of cytochrome P450 enzymes and thyroid hormones have been suggested as potential biomarkers of effect, but these are not specific for PBB exposure. It is well documented that PBBs induce hepatic Phase I xenobiotic metabolizing enzymes (i.e., cytochrome P450-dependent monooxygenases).[18] Subjects exposed to PBB in Michigan were found with elevated levels of two cytochrome P450-dependent enzymes, relative to controls. In addition, the thyroid is a sensitive target for PBBs and characteristic changes include reduced serum levels of T4 and other thyroid hormones. However, these biomarkers are likely to be common to the general class of polyhalogenated aromatic hydrocarbons, which also include structurally similar chemicals such as polychlorinated biphenyls, furans, and dioxins that cause similar endocrine disrupting effects and effects on the thyroid gland.[18]

The present study applies a metabolomics analysis to newly collected blood plasma samples obtained from participants of the Michigan PBB Registry who were (1) exposed to PBB as a result of accidental contamination of animal feed with PBB; or (2) were children or grandchildren of those exposed. The purpose of this study was twofold: first, to identify intermediate biomarkers that both mediate the metabolic effect of PBB exposure and its potential impact on disease risk; second, to investigate differences in the metabolic profiles of subjects divided into PBB exposure categories. We hypothesized that changes in metabolic processes resulting in measurable biomarkers were associated with current body burdens of PBB. Since the long-term health effects of PBB exposure are unknown, this study may lead to a better understanding of the health effects of PBB exposure and its underlying biological mechanism. The results of this study may also provide insight on improving biomarker discovery and applying metabolomics methods to investigate the human exposome.

2. Methods

2.1 Materials

Materials needed for plasma sample preparations included acetonitrile (HPLC grade, Sigma-Aldrich), an internal standard mix, and human reference plasma. The internal standard mix consisted of 14 stable isotopic chemicals: [$^{13}C_6$]-D-glucose, [^{15}N]-indole, [1,2-$^{13}C_2$]-palmitic acid, [^{15}N,$^{13}C_5$]-L-methionine, [2-^{15}N]-L-lysine dihydrochloride, [^{15}N]-choline chloride, [$^{13}C_5$]-L-glutamic acid, [$^{13}C_7$]-benzoic acid, [^{15}N]-L-tyrosine, [$^{15}N_2$]-uracil, [3,4-$^{13}C_2$]-cholesterol, [3,3-$^{13}C_2$]-cystine, [trimethyl-$^{13}C_3$]-caffeine, [U-$^{13}C_5$, U-$^{15}N_2$]-L-glutamine. Two types of pooled human reference plasma were used. One reference sample consisted of pooled plasma donated from the National Institute of Standards and Technology (NIST). This reference plasma was pooled from an equal number of healthy men and women aged 40 – 50 years old. The second reference sample, referred to as the Q standard (Qstd), consisted of pooled plasma from Equitech-Bio, Inc. (Kerrville, TX). This was pooled from an unknown number of males and females without demographic information.

2. 2 Samples

The study used a current cross-sectional sampling of a single time point in the Michigan Long-Term PBB Study. All protocols were reviewed and approved by the Institutional Review Board at Emory University (Appendix A). Blood was collected via venipuncture from 179 participants enrolled from September 2012 through December 2013. Samples used in the study were collected in October 2012 and June 2013. Study participants included members and descendants of farm families who were affected by the Michigan PBB accident in the 1970s. In addition, only plasma samples that were processed during field sampling were included in the present study in order to minimize differences from handling and processing of the samples. Blood samples were shipped to Emory and were frozen at -20 ℃ until analyzed. Samples were

analyzed as a part of the PBB cohort study to determine the PBB153 levels using a solvent extraction, solid phase extraction clean up procedure and analysis using gas-chromatography tandem mass spectrometry with isotope dilution quantification.

2.3 Metabolomic analysis

2.3.1 Sample preparation

The 179 frozen unknown plasma samples were separated into a total of ten batches to process one batch of samples per run. Each batch consisted of a maximum of 20 unknown samples and 2 Qstds to be measured at the beginning and end of each batch. NIST plasma samples were included only in the beginning of the first batch and the end of the last batch.

For each run, frozen plasma samples were thawed and a stock solution consisting of acetonitrile and the internal standard mix (130 µL acetonitrile and 3.25 µL internal standard mix per sample) was prepared. Aliquots of unknown plasma samples (65 µL), Qstds (50 µL), and NIST (50 µL) were all prepared in the same way. The stock solution (130 µL) was added to each Eppendorf tube containing aliquots of the samples. The samples were mixed and allowed to sit on ice for 30 min while letting acetonitrile crash proteins (i.e., causes proteins in the aliquots to denature and precipitate out). The samples were centrifuged at 13.2 rpm for 10 min at 4 ℃ to pellet the solids and precipitate out the proteins. Supernatants (~ 130 µL) were pipetted into autosampler vials for injection.

2.3.2 Chromatography and mass spectrometry

The metabolites in the sample extracts were temporally separated using high-performance C18 liquid chromatography. Reverse phase analysis was performed with a C18 column (Higgins Analytical, 100×2.1 mm, 5µm) and an acetonitrile gradient. The flow rate was 0.35 mL/min for the first 2 min and then changed to 0.40 mL/min for the remaining 8 min. The first 2-min period consisted of 5 % A (2 % formic acid in water), 35 % acetonitrile, 60 % water, followed by a 5-

min linear gradient to 5 % A, 95 % acetonitrile, 0 % water. The final 3-min period was maintained at 5 % A, 95 % acetonitrile, 0 % water. The injection volume was 10μL for each run. Since mass spectrometers detect chemicals as ions in the gas phase, samples were ionized with electrospray ionization in the positive ion mode. Ions were then detected using a quadrupole for precursor ion selection followed in tandem by a high-resolution Fourier transform ion trap (Q Exactive Hybrid Quadrupole-Orbitrap) mass spectrometer.

Data were collected continuously over the 10 min chromatographic separation, and these raw files were converted using Xcalibur file converter software (Thermo Fisher, San Diego, CA) to cdf files for further data processing. Spectral data were processed to extract peaks and quantify ion intensities using an adaptive processing R software package (apLCMS)[25], which is designed for use with liquid-chromatography mass spectrometry (LCMS) data with high mass accuracy. This software provided tables containing unique mass-to-charge ratio (*m/z*) values, retention time, and integrated ion intensity for each feature. In an untargeted metabolomics analysis, unknown chemicals are identified as features in which the retention time and *m/z* are used as identifiers in the absence of knowledge of chemical identity. The advantage of high-resolution and high mass accuracy instruments is that features can be associated with disease status even without chemical identity, and elemental composition can often be predicted from the accurate *m/z* values.[26] All samples were analyzed in triplicate, and the resulting spectra were averaged.

2.4 Data analysis

Two approaches were taken to achieve the aims of the present study: (1) a biomarker discovery approach to identify metabolites that are associated with PBB exposure, and (2) a metabolome-wide approach to identify metabolites that discriminate between classes of PBB exposure, ranging in different body burdens of PBB. Part 1 aims to identify discriminatory features among the set of features that are correlated with PBB exposure. Part 2 aims to identify features that discriminate between different PBB exposure classes among all features that are

detected in the samples (i.e., the metabolome). Samples were categorized into two or three different classes of PBB exposure based on the measured plasma PBB concentrations in order to make comparisons of metabolic profiles and identify discriminatory features among different classes of PBB exposure.

2.4.1 Correlation analysis

Correlation coefficients were derived using the statistical software SAS for every feature with the known concentrations of PBB 153 for each corresponding sample in order to identify features that were strongly associated with PBB exposure. Measured PBB concentrations were used as the independent variable, and the ion intensity profile of each feature was used as the dependent variable. Features were listed in rank order of correlation coefficient values, and features with high correlation with PBB 153 concentrations were selected as candidate biomarkers for further analyses. Features for which the absolute value of the correlation coefficient was greater than or equal to 0.3 were determined to be strongly correlated with PBB concentrations. Features with correlation coefficient values less than -0.5 were considered to be insignificant due to measurement error introduced by the instrument.

2.4.2 Two-class comparisons

All two-class differential expression analyses, which compare the metabolic profiles of two different groups of samples, were performed using the R software package MetBN.[27] MetBN performs two tasks: biomarker discovery using a False Discovery Rate (FDR) analysis and a metabolome-wide correlation analysis of the differentially expressed features. The feature table obtained from the LCMS was formatted into a table including *m/z*, retention time, and the measured ion intensity in each sample for each feature. The first two columns were the *m/z* and retention time, and the remaining columns consisted of the ion intensity profile for each sample. A separate file consisting of the class label information for each sample was also uploaded.

Samples were listed in the same order for both the feature table and the class label file. The data matrix was filtered by eliminating features that did not have intensities for at least 70 % of the samples in either of the two groups. The data matrix was also log2 transformed and quantile normalized.

Features that significantly discriminated between the two classes were selected based on the FDR analysis using the R software package Limma.[28] The Benjamini and Hochberg FDR was used to control for type I error due to multiple hypotheses testing at a significance level of q=0.05. The predictive accuracy of the FDR significant features was evaluated using k-fold cross-validation (CV) with a Support Vector Machine classifier. For CV, the dataset was first divided into 8 subsets, with 7 subsets used for training and the remaining 1 subset for testing. The process was repeated 8 times, and the CV accuracy was determined as the mean of the classification accuracy of the 8 iterations. The best set of significant features was selected using an optimization score that minimizes the number of false positives and increases the classification accuracy.

MetBN also provides several visualization tools to examine differences in the metabolic profiles of samples in the two classes. Principal component analysis (PCA) was used to visually inspect the separation of metabolomics data in relation to the dichotomous PBB exposure classes. PCA is an unsupervised method used to find the directions that best explain the variance in the dataset without referring to the class labels. PCA reduces high-dimensional metabolomics data into lower dimensions (e.g., two-dimensional or three-dimensional) by projecting the data into a new coordinate system using an optimal linear transformation for the data points such that most of the data variance lies in the first few principal axes (i.e., principal components [PCs]).[29] MetBN also performs hierarchical clustering and produces heatmaps to gain a simple overview of metabolomics data and find groupings or patterns in the data. Furthermore, MetBN performs a metabolome-wide network analysis of the selected FDR significant features using Pearson correlation coefficients to visualize the relationship between features.

2.4.3 Three-class comparisons

Analyses comparing the metabolic profiles of three classes of PBB exposure were performed using MetaboAnalyst (http://www.metaboanalyst.ca), which is an open-access, Web-based tool designed for processing, analyzing, and interpreting metabolomics data.[29] It is a tool designed to perform various types of common metabolomics data analysis tasks including data processing, data normalization, statistical analysis, and high-level functional interpretation.[30] MetaboAnalyst provides both univariate and multivariate statistical methods and multiple methods to produce data visualizations that facilitate the identification of significant features and patterns in metabolomics data.

To upload the dataset into MetaboAnalyst, the feature table obtained from the LCMS was formatted into a table with samples in columns and ion intensity profiles of each feature in rows. Once the dataset is uploaded, MetaboAnalyst performs data filtering to remove noise or non-informative features in order to decrease the computational burden of downstream data analyses. Data filtering was performed by removing baseline noise features characterized by ion intensity values close to zero based on the median ion intensity. Moreover, non-informative features characterized by constant ion intensities across samples were removed based on the interquartile range of the ion intensity. Data normalization was performed on features since the distributions of ion intensities were skewed for many features. Features were normalized using generalized log transformations to reduce the impact of very large feature values and to make all features normally distributed.

Both univariate and multivariate statistical analyses were used to identify features that significantly discriminated the three PBB exposure classes. One-way Analysis of Variance (ANOVA) was performed to select features that were significantly differentially expressed among the three classes based on the FDR threshold of $q = 0.05$. Two multivariate methods were used in the present analysis: PCA and partial least squares discriminant analysis (PLS-DA). Whereas PCA is an unsupervised method, PLS-DA is a supervised clustering method that uses

class labels to maximize the difference between classes. Therefore, prior knowledge of PBB exposure classifications can be incorporated into the analysis with PLS-DA. It uses multiple linear regression techniques to extract information in the dataset, via linear combination of features that best predict the class membership.[30] One concern with PLS-DA is that it tends to over-fit the data. In order to address this issue, the model needs to be validated to see whether the separation is statistically significant by using permutation tests. In each permutation, a PLS-DA model is built between the data and the permuted class labels using the optimal number of components.[30] The test statistic that was used as the performance measure for the permutation test was the separation distance. The separation distance is defined as the ratio of the between-group sum of squares and the within-group sum of squares (B/W-ratio). The empirical p-value is calculated as the percentage of times that the B/W of the permutated data achieves the same or better results than the B/W calculated from the original data. If the observed B/W-ratio is part of the distribution based on the permutated class assignments, the class discrimination is not statistically significant.[29] Similar to MetBN, MetaboAnalyst also performs hierarchical clustering and generates heatmaps to visually identify clusters or patterns in the data.

2.4.4 Feature annotation and pathway mapping

Identification of significant features is often only the beginning of metabolomics analysis. In order to make implications on the effect of PBB exposure on human health, these selected features need to be identified as metabolites and mapped in the context of metabolic pathways. Discriminatory features were annotated and mapped using several metabolomic databases. First, the experimental *m/z* values of the discriminatory features were searched for matches using Metlin (http://metlin.scripps.edu/), which is a pre-compiled database of known metabolites and their mass, chemical formula, and structures. Matches were searched with mass accuracy within 10 ppm for ion forms of $M+H^+$, $M+Na^+$, $M+K^+$, $M+H^+(-H_2O)$, and $M+H^+(2H_2O)$. Metlin also provides the Kyoto Encyclopedia of Genes and Genomes identification numbers (KEGG IDs) of

the matched metabolites, and these IDs were used to map identified metabolites in human metabolic pathways. Batches of metabolites were searched by their corresponding KEGG IDs using the pathway search in KEGG (http://www.genome.jp/kegg/). KEGG visually displays and highlights metabolites that are matched in the database. Metlin and KEGG consist of an extensive library of metabolites, and both databases provide a long list of matched metabolites or pathways. However, since there are frequently multiple matches to each feature that is searched, it is often difficult to tease out the matches that are most meaningful or appropriate for the purposes of the study.

MetaboAnalyst provides two functions to facilitate this process: metabolite set enrichment analysis (MSEA) and metabolic pathway analysis (MetPA). Both MSEA and MetPA perform statistical tests to rank matches of metabolite sets or metabolic pathways to facilitate interpretation of the results. MSEA is a method that has been developed as a metabolomic counterpart of the gene set enrichment analysis, which has been used widely in gene expression analysis. The purpose of MSEA is to investigate the enrichment of predefined groups of metabolites (or metabolite sets) instead of looking at metabolites individually.[30] Instead of identifying metabolites independently, MSEA matches groups of features to metabolite sets which consist of metabolites that are functionally related to each other. This allows for higher-level functional interpretation since metabolites are often interrelated and multiple metabolites can derive from the same metabolic pathway. MSEA was performed using over-representation analysis as the algorithm for enrichment analysis. Over-representation analysis uses a hypergeometric test to evaluate whether a particular metabolite set is represented more than expected by chance within the given list of metabolites. A reference metabolome is used to calculate a background distribution to determine if the matched metabolite set is more enriched for certain metabolites compared to random chance. MSEA uses all the metabolites of the metabolite set library as the reference metabolome. The pathway associated metabolite set library was used for the present analysis. However, interpretation of results from over-representation

analyses should be done with caution since the results may be intrinsically biased toward metabolite sets containing compounds that are more abundant or more easily detected because of limitations in current analytical technologies.[30]

MetPA is a tool that helps to identify metabolic pathways that are most likely to be associated with the given list of metabolites. MetPA combines pathway enrichment analysis and pathway topology analysis to identify the pathway that is most significantly associated. Pathway enrichment analysis performs an over-representation analysis using hypergeometric tests. Pathway topology analysis measures the importance of a metabolite in the given metabolic network by using a centrality measure. The degree centrality, which is defined as the number of links occurred upon a node (or a metabolite), was used as the centrality measure for analyses in the present study. This analysis takes into consideration that changes in more important positions of a pathway network will trigger more severe impacts on the pathway. The metabolic pathways used in MetPA are obtained from the KEGG database and presented as networks of chemical compounds, with metabolites represented as nodes and reactions as edges.[30]

3. Results

3.1 Descriptive analysis of PBB exposure among the sample cohort

The distribution of the measured plasma PBB concentrations from the 179 exposed subjects was positively skewed with a median concentration of 0.26 ng/mL and a standard deviation of 14.83 ng/mL. There were 6 samples with no detectable levels of PBB, and the maximum PBB concentration was 164.52 ng/mL.

The 179 PBB exposed plasma samples were classified into different PBB exposure classes using three classification methods (Table 3.1) for further data analyses. Classification Methods 1 and 2 use binary classes (i.e., low and high PBB exposure) with the threshold based on the median PBB concentration (i.e., 0.26 ng/mL) and 1.00 ng/mL, respectively. Classification

Method 3 categorized samples into three classes based on the PBB concentration: less than 0.10 ng/mL (low), 0.10 to 1.00 ng/mL (medium), greater than 1.00 ng/mL (high).

Table 3.1 Descriptive statistics of plasma PBB concentrations of the 179 exposed subjects

Classification Method	Class	PBB concentration range (ng/mL)	Frequency	Percent (%)
1	1 (Low) 2 (High)	< 0.26 (median) ≥ 0.26	89 90	49.72 50.28
2	1 (Low) 2 (High)	< 1.00 ≥ 1.00	138 41	77.09 22.91
3	1 (Low) 2 (Medium) 3 (High)	< 0.10 0.10 – 1.00 > 1.00	60 78 41	33.52 43.58 22.91

3.2 Metabolomic data

Extraction of mass spectral data obtained from the LCMS yielded 11,202 features defined by their unique *m/z*, retention time, and the measured ion intensity for each sample. The *m/z* values ranged from 85.020 to 1,274.922. The averaged ion intensity values of the triplicates were used for further data analyses. The quality of the data was checked to confirm there were no significant systematic variations among samples measured in different batches (i.e., batch effects) by using PCA prior to analyses.

3.3 Part 1. Biomarker discovery: Identification of discriminatory features associated with PBB exposure

Correlation analysis was used to identify features that were strongly associated with PBB exposure. Spearman's rank correlation coefficient was determined for every feature with the measured PBB 153 concentrations of the 179 exposed samples. Spearman's rank correlation was used since the distribution of PBB concentrations was skewed and the ion intensity profiles of the features were not normally distributed. Of the 11,202 detected features, a total of 73 features

(0.65 %) were found to be strongly correlated with PBB concentrations based on Spearman's rank correlation. Correlation coefficients (r_s) of the 11,202 features were normally distributed with a mean of -0.0041 and a standard deviation of 0.102. Of the 73 significant features, approximately half of the features were negatively correlated with PBB exposure (34 features, 46.58 %) and the other half was positively correlated with PBB exposure (39 features, 53.42 %). The feature with the strongest correlation had a $r_s = 0.477$.

Two-class differential expression analysis was performed on the 73 features strongly correlated with PBB exposure to determine if any of these features discriminated between binary classes of PBB exposure. None of the features were filtered out for the following analyses since all 73 features had ion intensities for more than 70 % of the samples in either of the two exposure classes.

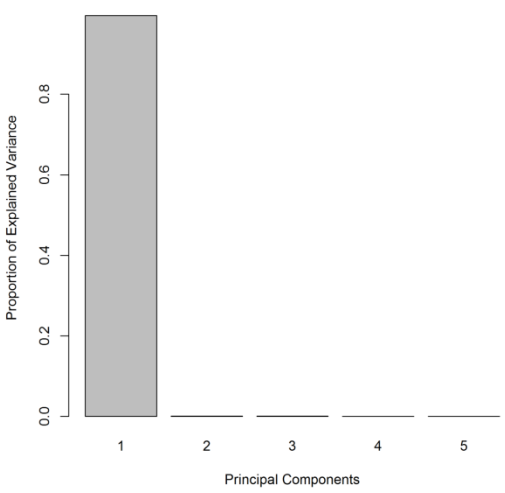
3.3.1 Feature selection based on Classification Method 1

First, analysis was performed with Classification Method 1, where exposed samples were classified based on the median PBB concentration. The 73 features that were significantly correlated with PBB exposure were first examined globally to get an overview of the data. Global PCA was performed to visually examine if there were sample groupings without referring to the assigned PBB exposure classes. The global scree plot shows the amount of variance explained by the top five PCs (Fig 3.1.a). Most of the variance in the data was contained in the first PC (approximately 90 %). Global pair-wise score plots between the first three PCs are shown in Fig 3.1. The scores in these plots are the values of the samples in the PC coordinate system. Low PBB exposure samples are represented in red, and high PBB exposure samples are represented in green. Based on the 73 features, the pair-wise score plot of PC1 and PC2 (Fig 3.1.b) shows that the two classes were somewhat separated but not completely. Low PBB samples are clustered in the top half and high PBB samples are clustered in the bottom half of the plot. There was more overlap between the two classes and the separation was not clear in the other two score plots (Fig

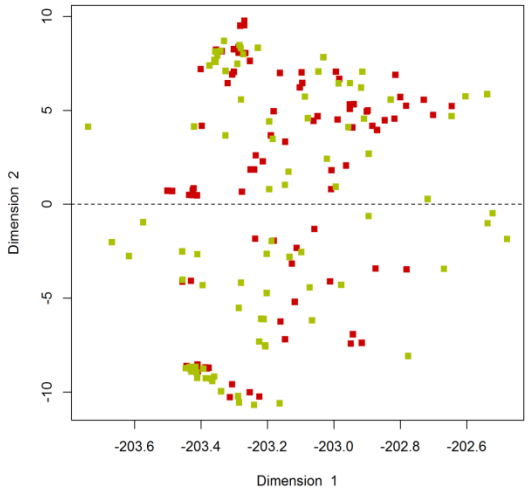
3.1.c and Fig 3.1.d). This indicated that there was no natural separation (i.e., without referring to PBB exposure classification) of samples according to the level of PBB exposure based on the 73 features that were associated with PBB concentrations.

The global Manhattan plot shows the FDR analysis of the 73 features comparing 89 low exposure subjects with 90 high exposure subjects (Fig 3.2). The negative log p-value is plotted against the *m/z* values of the features on the x-axis, ordered in increasing value from 102.128 to 1,178.440. The horizontal dashed line represents the FDR significance threshold (q = 0.05). Green dots above the threshold line represent features that were significantly differentially expressed between the two classes. With Classification Method 1, there were 37 FDR significant features with a CV accuracy of 80.45 %.

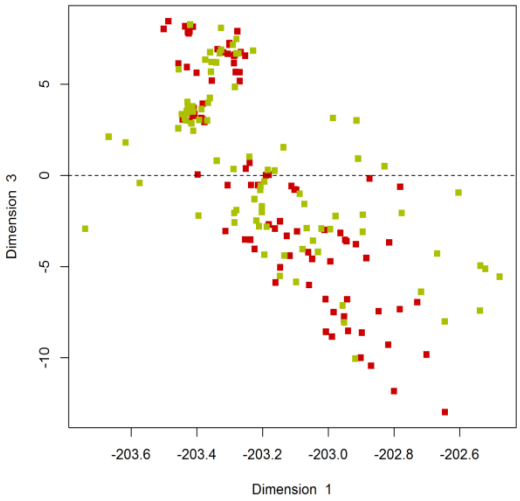
Figure 3.1 Classification Method 1: Global PCA of the 73 features correlated with PBB exposure (red: low PBB, green: high PBB)



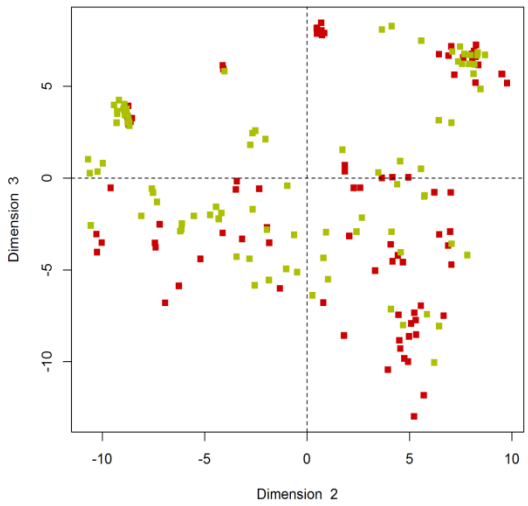
a) Scree plot



b) Score plot: PC1 vs. PC2

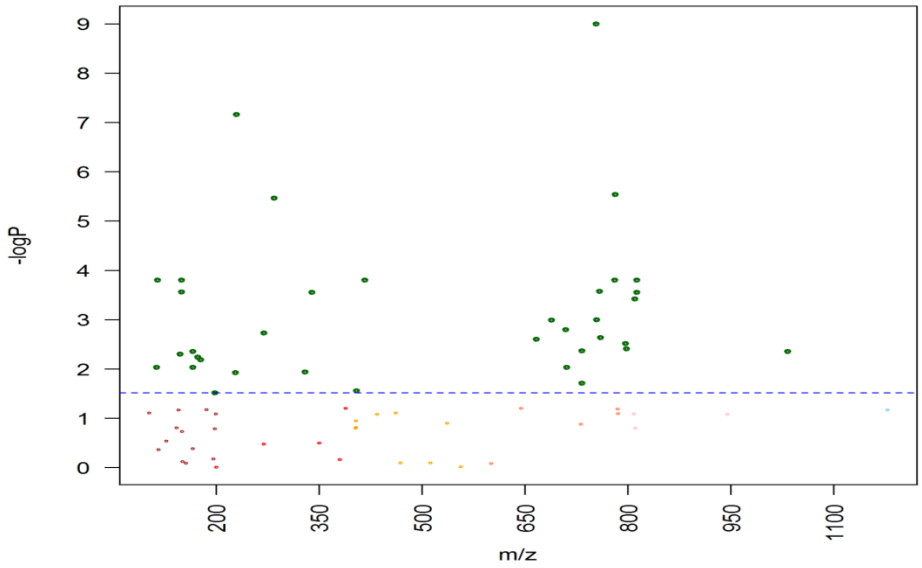


c) Score plot: PC1 vs. PC3



d) Score plot: PC2 vs. PC3

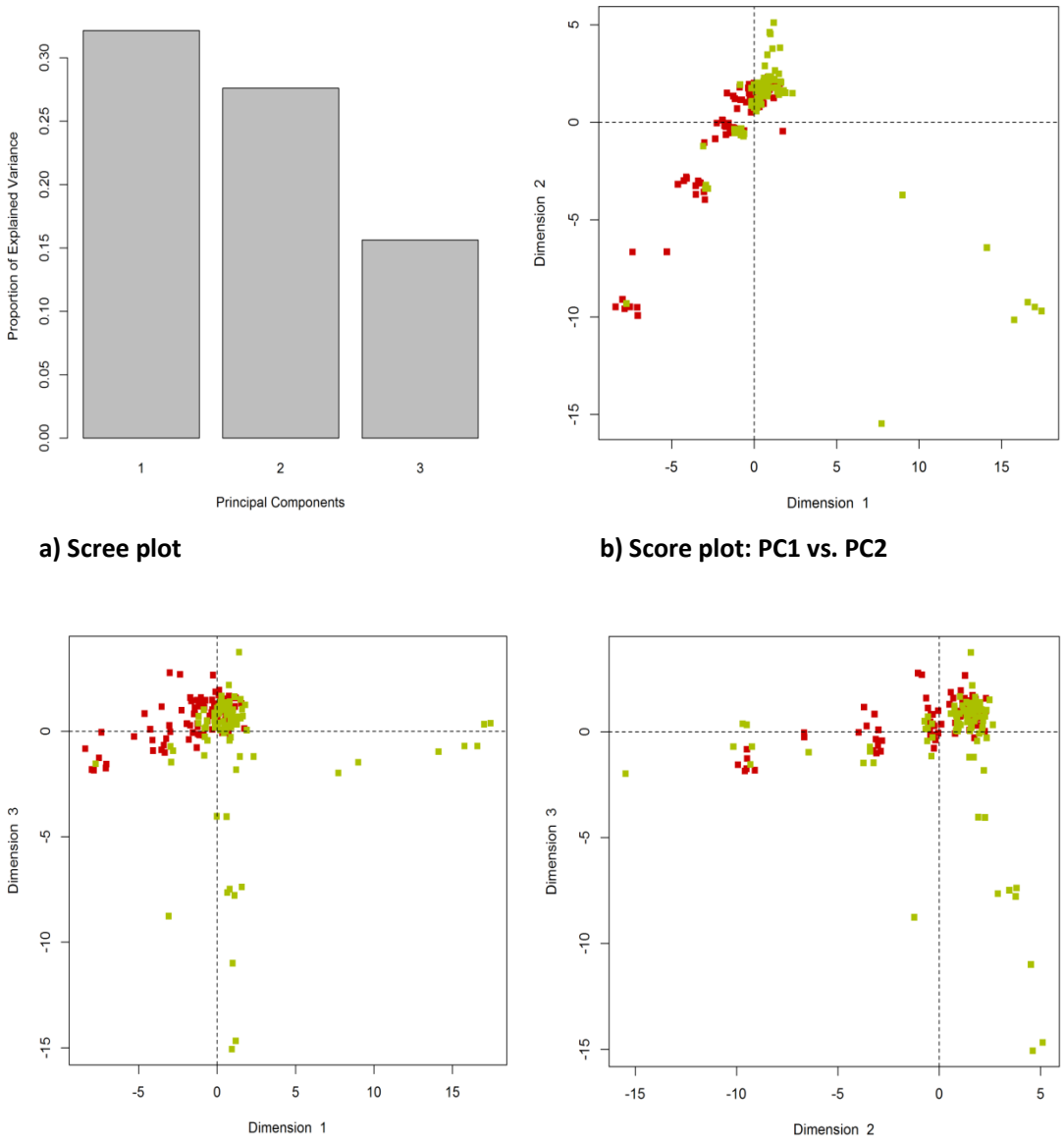
Figure 3.2 Classification Method 1: Global Manhattan plot



A focused PCA was performed to examine whether there were sample groupings based on the 37 FDR significant features. The scree plot of the 37 FDR significant features shows that most of the variance was explained by the first three components (PC1: approximately 33 %, PC2: approximately 28 %, PC3: 16 %) (Fig 3.3.a). A comparison of PC1 and PC2 (Fig 3.3.b) shows a good separation between the two classes where low PBB samples were clustered to the left and high PBB samples to the right side of the plot. The remaining two score plots also show signs of separation with little overlap (Fig 3.3.c and Fig 3.3.d). By comparing the global PCA with the focused PCA, it can be seen that the 37 FDR significant features selected from the initial set of 73 features led to a noticeable separation of samples based on PBB exposure levels. Fig 3.4 shows a heatmap overview of the standardized z-score of the ion intensities with hierarchical clustering performed on both samples (horizontal axis) and the 37 FDR significant features (vertical axis). Low PBB samples are represented in red, and high PBB samples are represented in green. The dendrogram on the top showed that several small clusters of samples were

identified. There were no obvious patterns of change in the metabolic profiles of the samples that could be discerned from the heatmap. Box-and-whisker plots comparing the mean and standard error of features for low PBB exposure and high PBB exposure samples were also examined for each of the 37 selected features (Appendix B: Fig S1).

Figure 3.3 Classification Method 1: Focused PCA of the 37 FDR significant features correlated with PBB exposure (red: low PBB, green: high PBB)



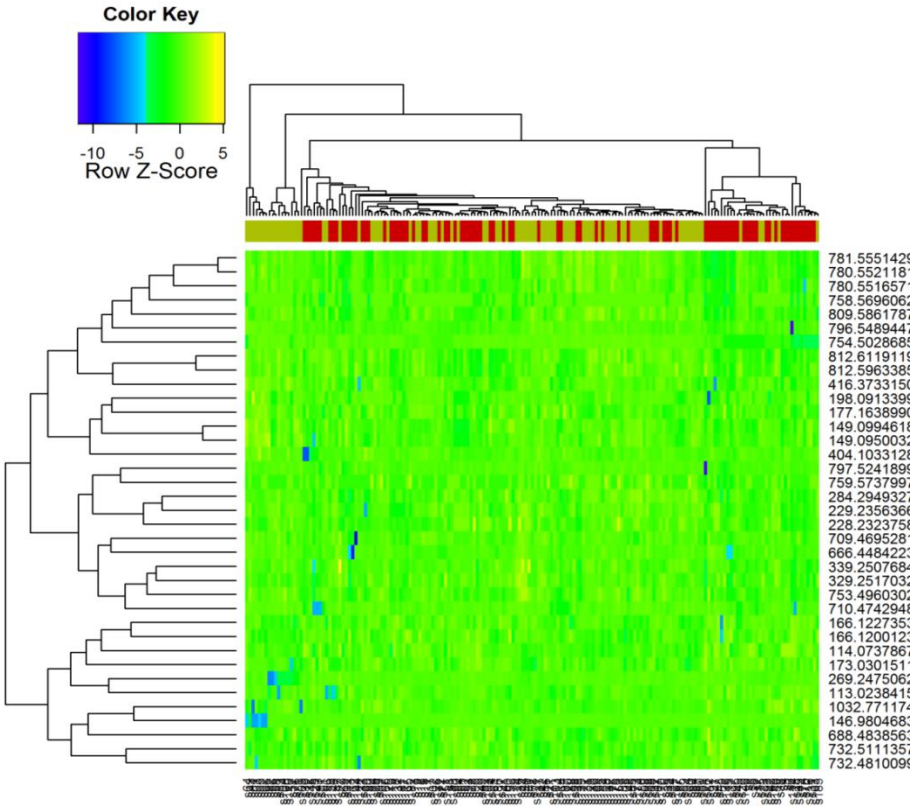
a) Scree plot

b) Score plot: PC1 vs. PC2

c) Score plot: PC1 vs. PC3

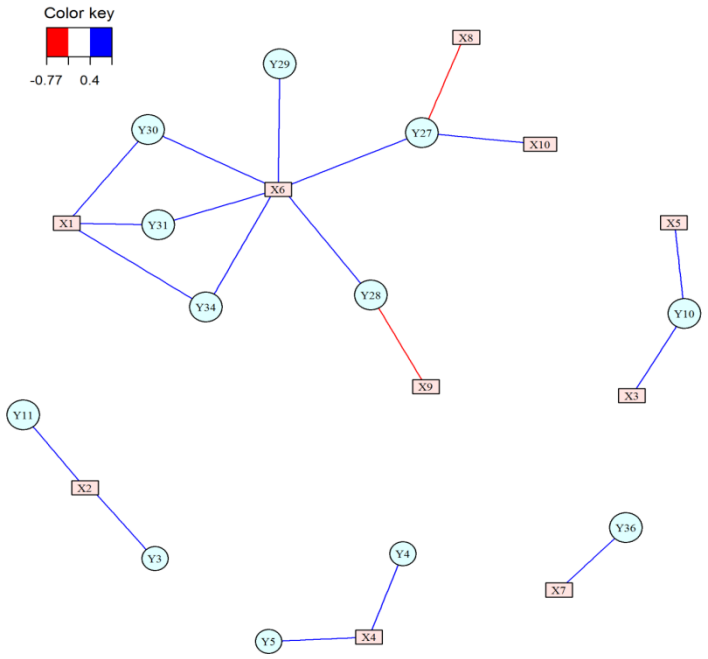
d) Score plot: PC2 vs. PC3

Figure 3.4 Classification Method 1: Hierarchical clustering and heatmap of the 37 FDR significant features (red: low PBB, green: high PBB)



The correlation network based on the metabolome-wide correlation analysis of the 37 differentially expressed features is shown in Fig 3.5. The correlations between features were assessed based on Pearson correlation coefficients. The 37 FDR significant features are represented as circles, and all other features from the initial set of 73 features are represented as rectangles. Only features with a significant correlation based on the FDR threshold of q = 0.05 are shown in the figure. Negative correlations between features are represented with red lines, whereas positive correlations are represented with blue lines. There was one large cluster and 4 smaller clusters of features that were significantly correlated with one another. Most features were positively correlated with each other.

Figure 3.5 Classification Method 1: Correlation network based on the 37 FDR significant features

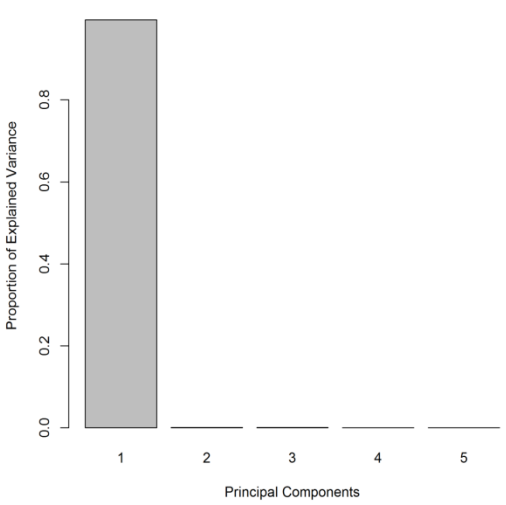


3.3.2 Feature selection based on Classification Method 2

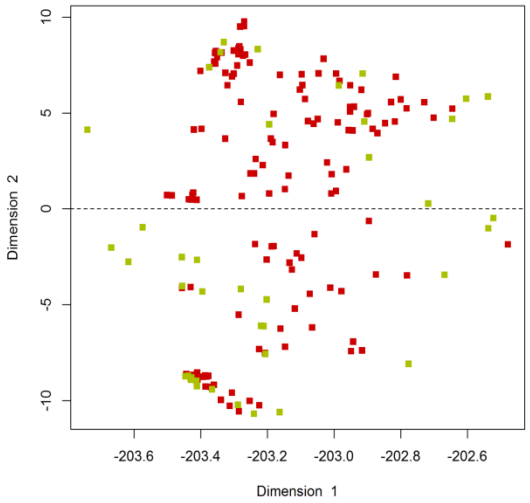
Next, two-class differential expression analysis was performed with Classification Method 2, where samples were classified based on the threshold PBB concentration of 1 ng/mL. Global PCA showed that there was no clear separation of the two classes based on the 73 features using Classification Method 2 (Fig 3.6), thus indicating that there was no natural separation of samples according to PBB exposure levels. Most of the variance was explained by PC1 (Fig 3.6.a), and all pair-wise score plots showed overlap between the two PBB exposure classes (Fig 3.6.b, Fig 3.6.c, Fig 3.6.d). The global Manhattan plot shows the FDR analysis of the 73 features comparing 138 low exposure subjects with 41 high exposure subjects (Fig 3.7). With Classification Method 2, 39 FDR significant features were identified above the threshold, and the CV accuracy was 84.92 %. Compared to Classification Method 1, more features were selected using Classification Method 2. The CV accuracy was also higher with Classification Method 2,

indicating that more samples were correctly classified into their assigned PBB exposure class based on the discriminatory features selected using Classification Method 2.

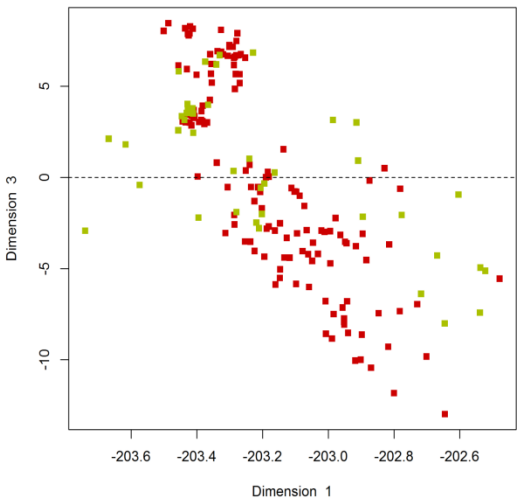
Figure 3.6 Classification Method 2: Global PCA of the 73 features correlated with PBB exposure (red: low PBB, green: high PBB)



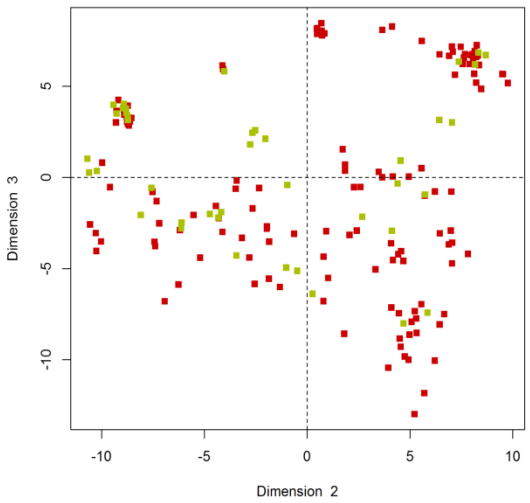
a) Scree plot



b) Score plot: PC1 vs. PC2

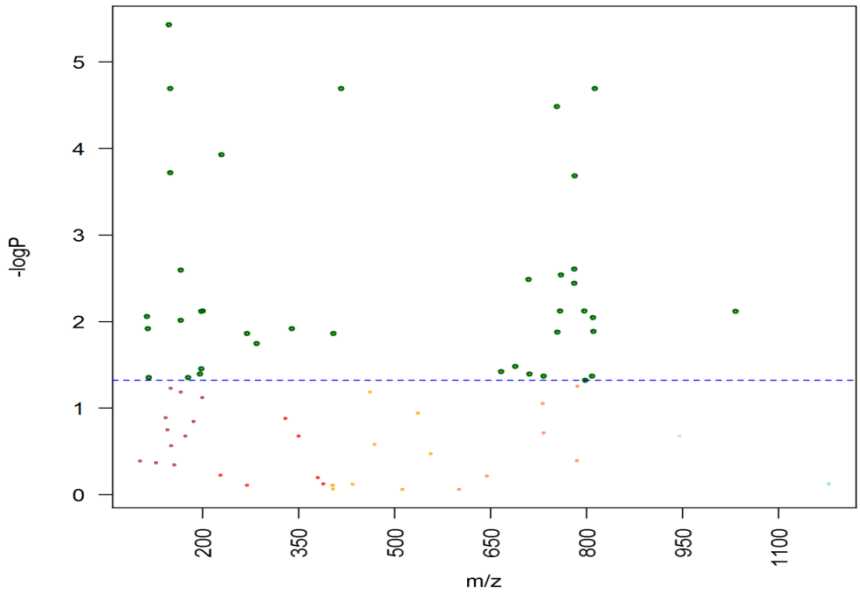


c) Score plot: PC1 vs. PC3



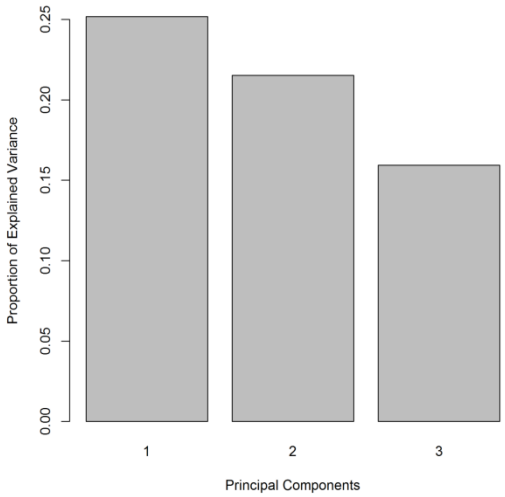
d) Score plot: PC2 vs. PC3

Figure 3.7 Classification Method 2: Global Manhattan plot

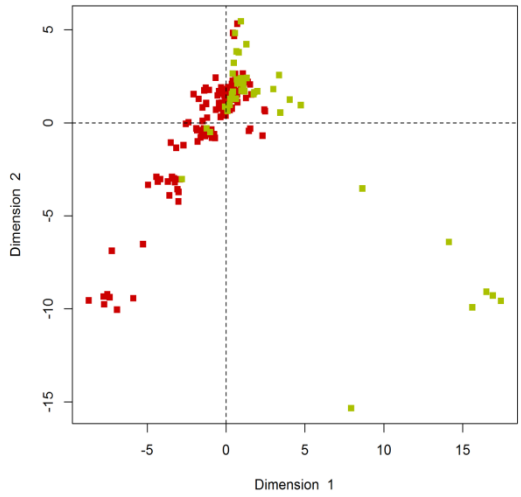


The PCA scree plot of the 39 FDR significant features show that PC1, PC2, and PC3 explained approximately 25 %, 22 %, and 17 % of the variance in the data, respectively (Fig 3.8.a). Pair-wise score plots of PC1 vs. PC2 (Fig 3.8.b) and PC1 vs. PC3 (Fig 3.8.c) show good separation of the two classes, meaning that PC1 contributed in the separation of the classes the most. With the 39 FDR significant features, there appears to be a separation of samples according to differences in the level of PBB exposure. Although there were no obvious differences in the ion intensity profiles between the two classes based on the heatmap, multiple clusters of low PBB exposure samples were found from the dendrogram on the horizontal axis (Fig 3.9). Low PBB samples tended to cluster more since a larger portion of the sample set was classified as low PBB exposure based on Classification Method 2. Box-and-whisker plots of 39 FDR significant features can be found in Appendix B (Fig S2).

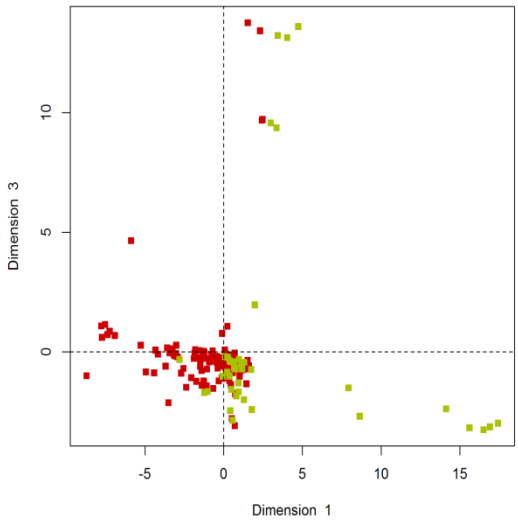
Figure 3.8 Classification Method 2: Focused PCA of the 39 FDR significant features correlated with PBB exposure (red: low PBB, green: high PBB)



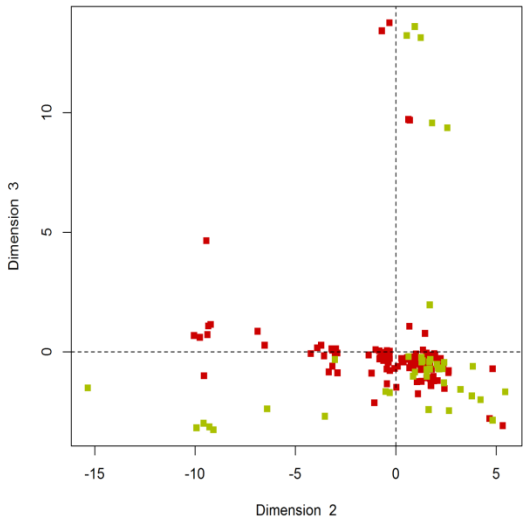
a) Scree plot



b) Score plot: PC1 vs. PC2

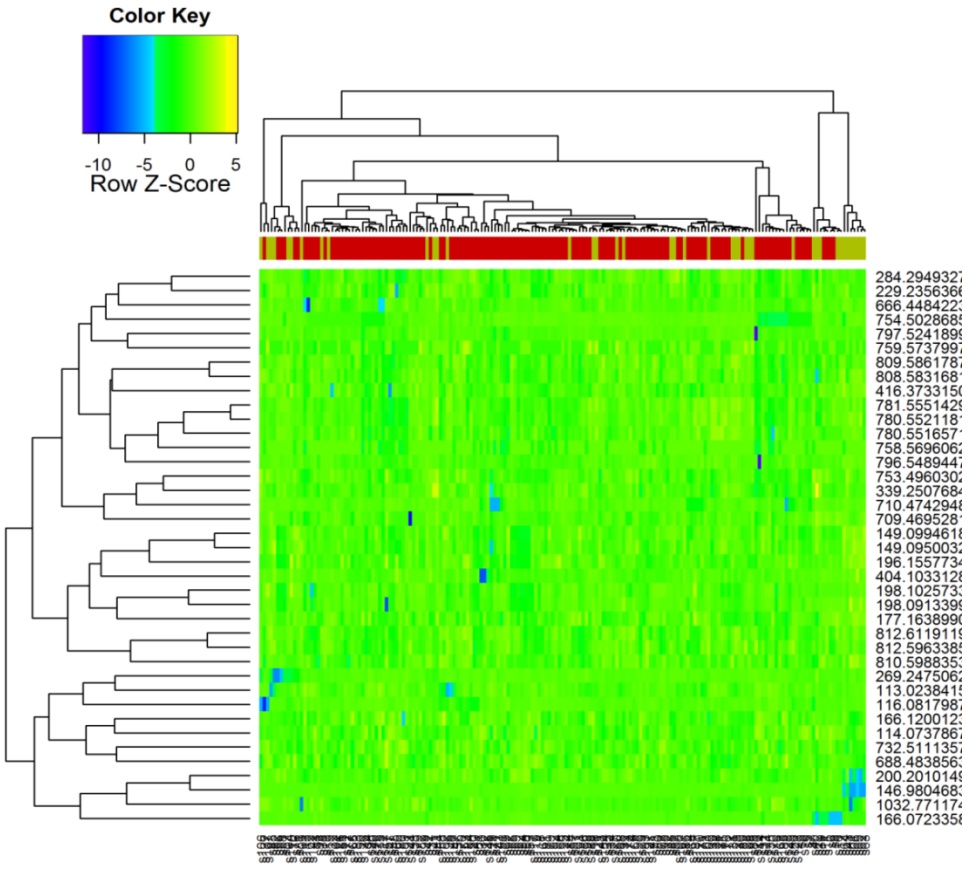


c) Score plot: PC1 vs. PC3



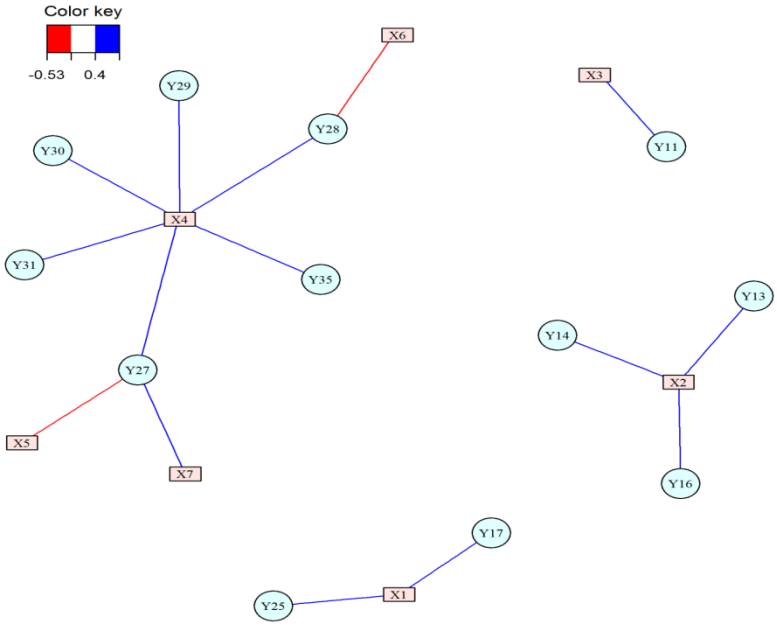
d) Score plot: PC2 vs. PC3

Figure3.9 Classification Method 2: Hierarchical clustering and heatmap of the 39 FDR significant features (red: low PBB, green: high PBB)



The correlation network in Fig 3.10 shows the relationships between selected features with significant correlations. There was one large cluster and three smaller clusters of features that were significantly correlated with one another. Similar to the results from Classification Method 1, most of the features were positively correlated with each other. There were fewer clusters identified with Classification Method 2 compared to Method 1.

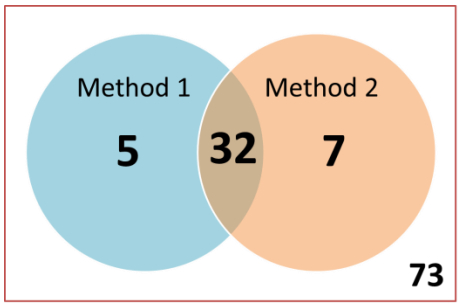
Figure 3.10 Classification Method 2: Correlation network based on the 39 FDR significant features



3.3.3 Feature annotation and mapping selected features to metabolic pathways

Metabolomic database searches were used to tentatively identify the significant discriminatory features associated with PBB exposure that were selected based on Classification Method 1 and 2. Classification Method 1 and 2 identified 37 and 39 significant features, respectively. Of these significant features, 32 features were selected by both methods (Fig 3.11). With the two classification methods combined, a total of 44 features associated with PBB exposure were found to discriminate between low and high PBB exposures. With the 32 features mutually selected by both Classification Method 1 and 2, 815 metabolites were listed as potential matches using Metlin. Multiple features were assigned several matched metabolites, and some of the metabolites were similar forms of one another. Of the 815 metabolites suggested by Metlin, 59 were identified with KEGG IDs. The 59 metabolites with KEGG IDs were then analyzed using MetaboAnalyst's MSEA and MetPA.

Figure 3.11 Comparison of discriminatory features correlated with PBB exposure selected by Classification Method 1 and 2



Of the 59 metabolites with KEGG IDs, 38 were exactly matched with compounds contained in MetaboAnalyst's metabolite set library (Appendix C: Table S1). Over-representation analysis of the matched metabolites was performed using the hypergeometric test to evaluate whether a particular metabolite set was represented more than expected by chance within the given list of metabolites. Fig 3.12 shows the summary plot of MSEA performed on the 32 discriminatory features that were mutually selected by both Classification Method 1 and 2. Table 3.2 shows the results of the over-representation analysis providing the raw and adjusted p-values of the matched metabolite sets. The metabolite set that had the most number of matches was the arachidonic acid metabolite set, but this was not statistically significant at the FDR significance level of q = 0.05 (FDR adjusted p-value = 1.000). A total of 37 compounds are included in the arachidonic acid metabolism set, and 4 metabolites matched the compounds in this set.

Figure 3.12 Summary plot of the Metabolite Set Enrichment Analysis (MSEA) of the 32 discriminatory features mutually selected by both Classification Method 1 and 2

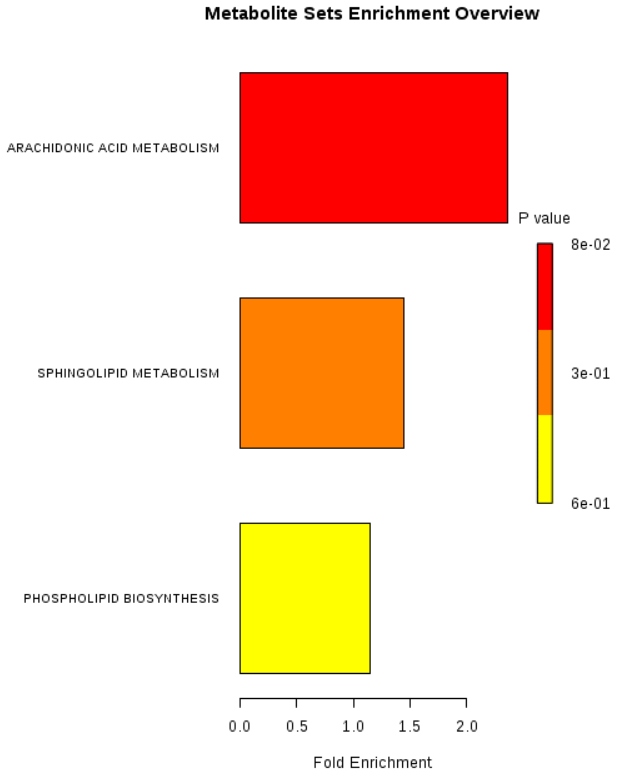


Table 3.2 MSEA: Results of over-representation analysis of the 32 discriminatory features mutually selected by both Classification Method 1 and 2

Metabolite set	Total*	Expected*	Hits*	Raw p*	Holm p*	FDR p*	Match*
ARACHIDONIC ACID METABOLISM	37	1.700	4	0.085	1.000	1.000	14,15-DHET; 8,9-DHET; 11,12-DHET; 5,6-DHET [+]
SPHINGOLIPID METABOLISM	15	0.691	1	0.510	1.000	1.000	Sphinganine
PHOSPHOLIPID BIOSYNTHESIS	19	0.875	1	0.596	1.000	1.000	PC(16:0/16:0)

* Total number of compounds in the metabolite set (Total), number of matched compounds (Hits), the original p-value calculated from the enrichment analysis (Raw p), the adjusted p-value using the Holm-Bonferroni method (Holm p), the adjusted p-value using the FDR (FDR p), matched metabolites (Match)

[+] DHET: dihydroxyeicosatrienoic acid

The 59 metabolites with KEGG IDs were then analyzed with MetPA. Again, the 59 metabolites were matched with 38 compounds in MetaboAnalyst's library and these were used for further analysis. Over-representation analysis using the hypergeometric test was performed to test if compounds involved in a particular pathway were enriched compared to random chance. The pathway topology analysis was performed using degree centrality to measure the amount of impact the matched compounds present within the particular pathway. Fig 3.13 shows the summary plot and Table 3.3 shows the full results of MetPA performed on the 32 discriminatory features that were mutually selected by both Classification Method 1 and 2. In Fig 3.13, all the matched pathways are displayed as circles, with red and orange circles indicating lower p-values. The size of each circle is based on the pathway impact value. The x-axis represents the pathway impact values from the topology analysis, and the y-axis represents the negative log of the p-values from the enrichment analysis. Pathways lying near the top part of the diagonal line ($y = x$) indicate higher number of matches in compounds and that these compounds are more likely to have larger impacts on the pathway. From the summary plot, it can be seen that the valine, leucine, and isoleucine biosynthesis pathway lies closest to the diagonal. However, this pathway did not significantly match the set of 32 features that were analyzed since only 3 compounds matched the metabolites involved in this pathway (FDR adjusted p-value = 0.218). Although the arachidonic acid metabolism pathway had the highest number of matched compounds, this was still statistically not significant at the FDR threshold of $q = 0.05$ (FDR adjusted p-value = 0.198). Similar results were obtained when MSEA and MetPA were performed with the total of 44 discriminatory features that were selected using Classification Method 1 and 2.

Figure 3.13 Summary plot of the Metabolic Pathway Analysis (MetPA) of the 32 discriminatory features mutually selected by both Classification Method 1 and 2

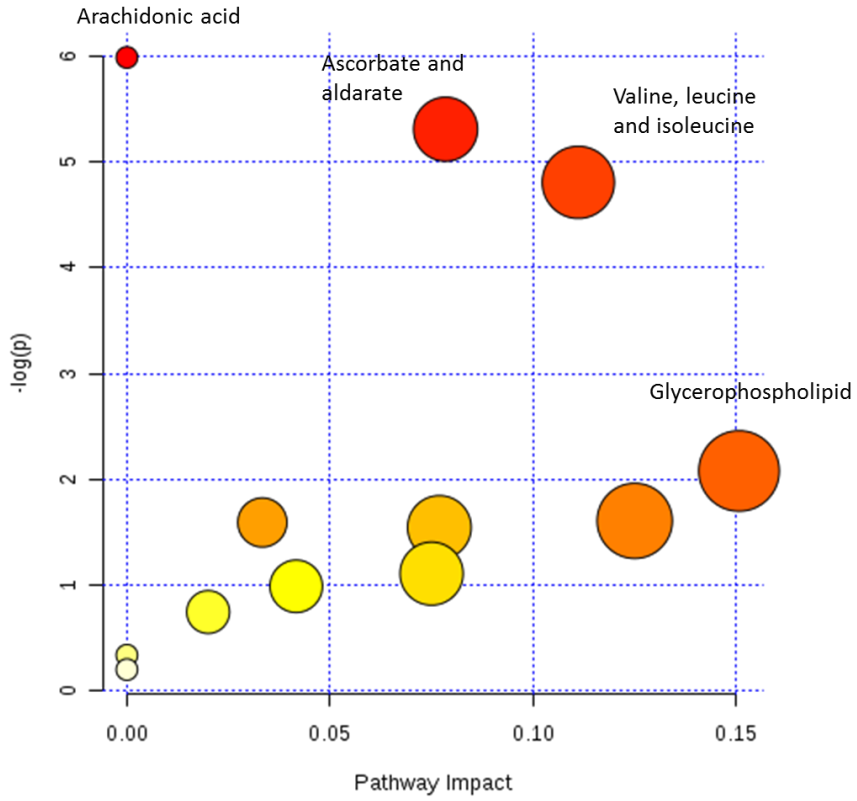


Table 3.3 MetPA: Results of the enrichment analysis and topology analysis of the 32 discriminatory features mutually selected by both Classification Method 1 and 2

Pathway	Total*	Expected*	Hits*+	Raw p*	Holm p*	FDR p*	Impact*
Arachidonic acid metabolism	62	0.979	5	0.003	0.201	0.198	0.000
Ascorbate and aldarate metabolism	45	0.710	4	0.005	0.391	0.198	0.078
Valine, leucine and isoleucine biosynthesis	27	0.426	3	0.008	0.638	0.218	0.111
Glycerophospholipid metabolism	39	0.616	2	0.125	1.000	1.000	0.151
Pentose and glucuronate interconversions	53	0.837	2	0.203	1.000	1.000	0.033

* Total number of compounds in the metabolite set (Total), number of matched compounds (Hits), the original p-value calculated from the enrichment analysis (Raw p), the adjusted p-value using the Holm-Bonferroni method (Holm p), the adjusted p-value using the FDR (FDR p); the pathway impact value (Impact)
+ Only pathways with more than one matched compound are shown in the table

The results of MSEA and MetPA suggest that some of the selected discriminatory features associated with PBB exposure are likely to be metabolites in the arachidonic acid metabolic pathway. The arachidonic acid metabolic pathway entails the production and subsequent metabolism of arachidonic acid, which is an omega-6 fatty acid. The five matched compounds in this pathway consisted of 14,15- DHET; 8,9- DHET; 11,12- DHET; 5,6-DHET (DHET: dihydroxyeicosatrienoic acid); and phosphatidylcholine. These DHET metabolites are generated as a result of cytochrome p450 (CYP) enzymes converting arachidonic acid into epoxyeicosatrienoic acids, which are then converted into vicinal diols with epoxide hydrolases.[31] However, when checking the input *m/z* values that these compounds were matched on, it was found that all four DHET metabolites were matched based on the same single feature (*m/z* = 339.251). Phosphatidylcholine was matched on 2 features (*m/z* = 758.570, 780.552) (Table S1). Therefore, only 3 of the 32 selected discriminatory features were contained in the arachidonic acid metabolic pathway. Since the ascorbate and aldarate metabolic pathway was the second best match, the *m/z* values of the matched compounds in this pathway were also checked. However, all 4 of the matched compounds in the pathway were based on the same 1 feature (*m/z* = 113.024) (Table S1).

3.4 Part 2. Metabolome-wide identification of discriminatory features between PBB exposure classes

A metabolome-wide FDR analysis was used to identify features that significantly discriminated between different classes of PBB exposure among the initial set of 11,202 features that were detected in the metabolome. Two-class differential expression analysis was applied for binary classification methods (i.e., Classification Methods 1 and 2), and MetaboAnalyst was used for Classification Method 3, where PBB exposure was classified into three groups (i.e., low, medium, high).

3.4.1 Feature selection based on Classification Method 1

A global PCA was performed to look at the variation in the metabolic profiles of the samples for the 11,202 detected features. The scree plot showed that approximately 90 % of the variance in the ion intensity profiles was explained by the first PC (Fig S3.a). Pair-wise comparisons of the global score plots showed that there was no clear natural separation of the 11,202 features between the low PBB exposure high PBB exposure samples using Classification Method 1 (Score plot of PC1 vs. PC2 shown in Fig 3.14; Score plots of other PCs shown in Appendix B: Fig S3.b and Fig S3.c). Fig 3.15 shows the global Manhattan plot of the FDR analysis of 11,202 features comparing low PBB exposure with high PBB exposure samples. The *m/z* values of features on the x-axis ranged from 85.020 to 1,274.922. There were 47 FDR significant features, and the CV accuracy was 88.27 %.

Figure 3.14 Classification Method 1: Global PCA score plot (PC1 vs. PC2) of the 11,202 features of the metabolome (red: low PBB, green: high PBB)

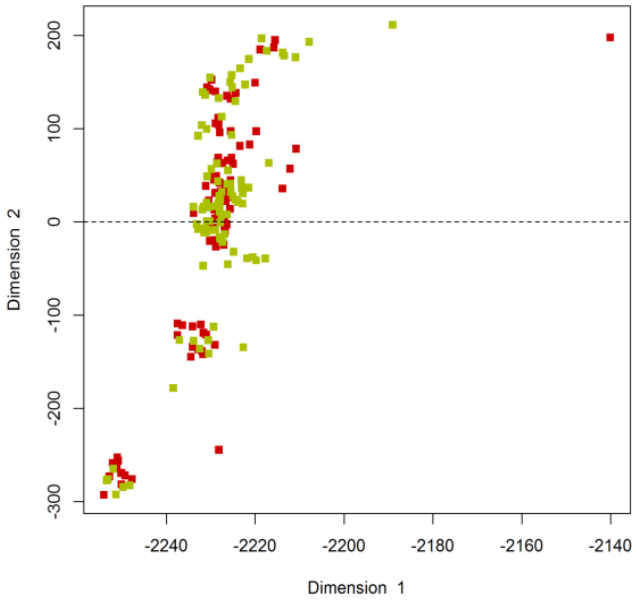
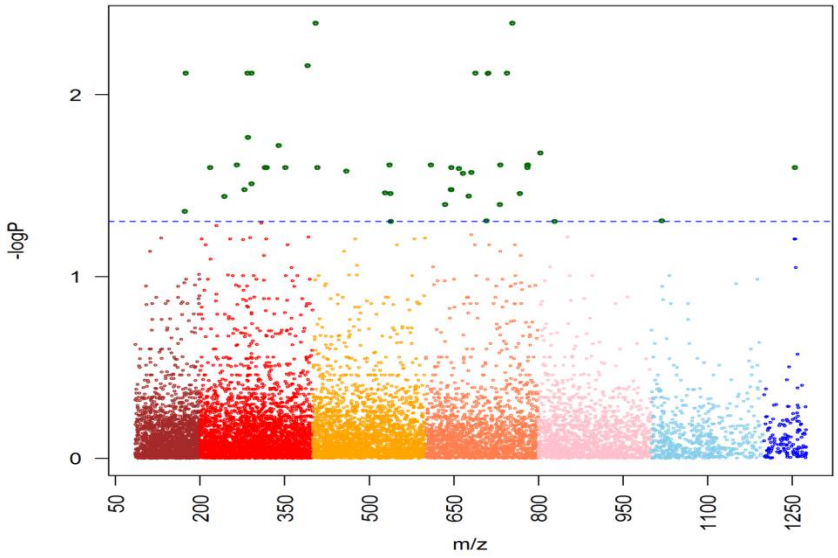
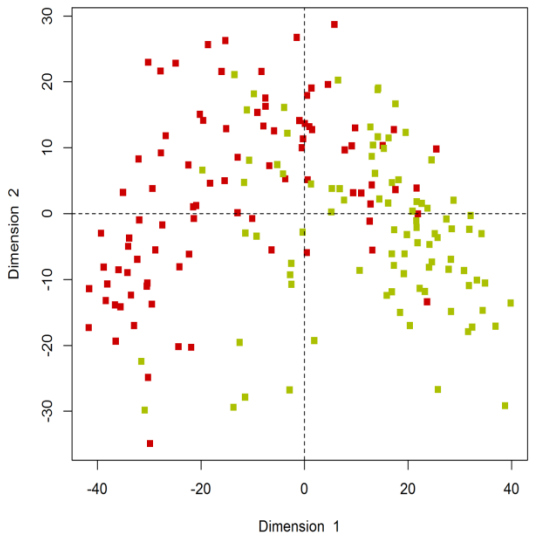


Figure 3.15 Classification Method 1: Global Manhattan plot of the 11,202 features of the metabolome

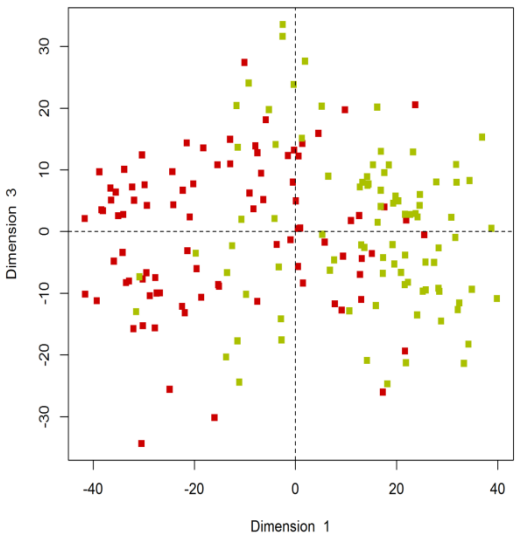


The scree plot of the 47 FDR significant features showed that PC1, PC2, and PC3 explained approximately 21 %, 7%, and 6% of the variation, respectively (Fig S4.a). Pair-wise score plots of PC1 vs. PC2 (Fig 3.16.a), and also PC1 vs. PC3 (Fig 3.16.b) show signs of separation between the two classes, where low PBB exposure samples are clustered to the left and high PBB exposure samples are clustered to the right of the plot. There was more overlap between the two classes in the score plot between PC2 and PC3 (Fig S4.b). The heatmap overview of the ion intensity profiles of the samples shows that low PBB exposure samples cluster more to the right of the dendrogram on the horizontal axis and high PBB exposure samples to the right (Fig 3.17). The box-and-whisker plots of the 47 FDR significant features can be seen in Appendix B (Fig S5).

Figure 3.16 Classification Method 1: Focused PCA of the 47 FDR significant features of the metabolome (red: low PBB, green: high PBB)

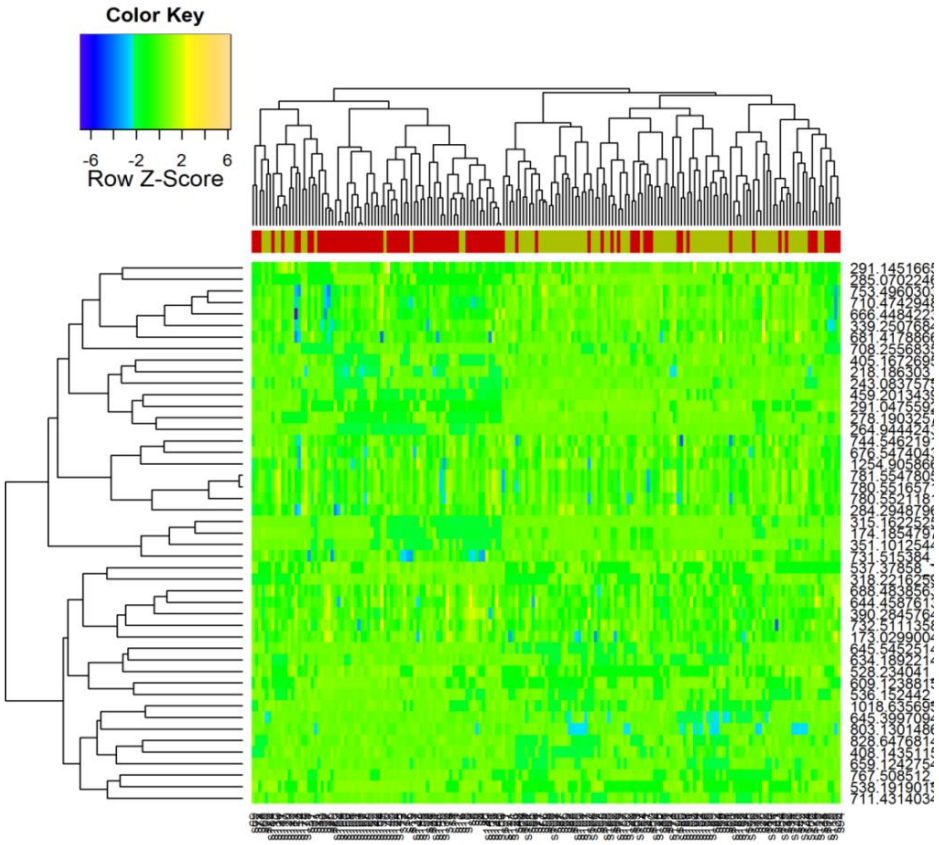


a) Score plot: PC1 vs. PC2



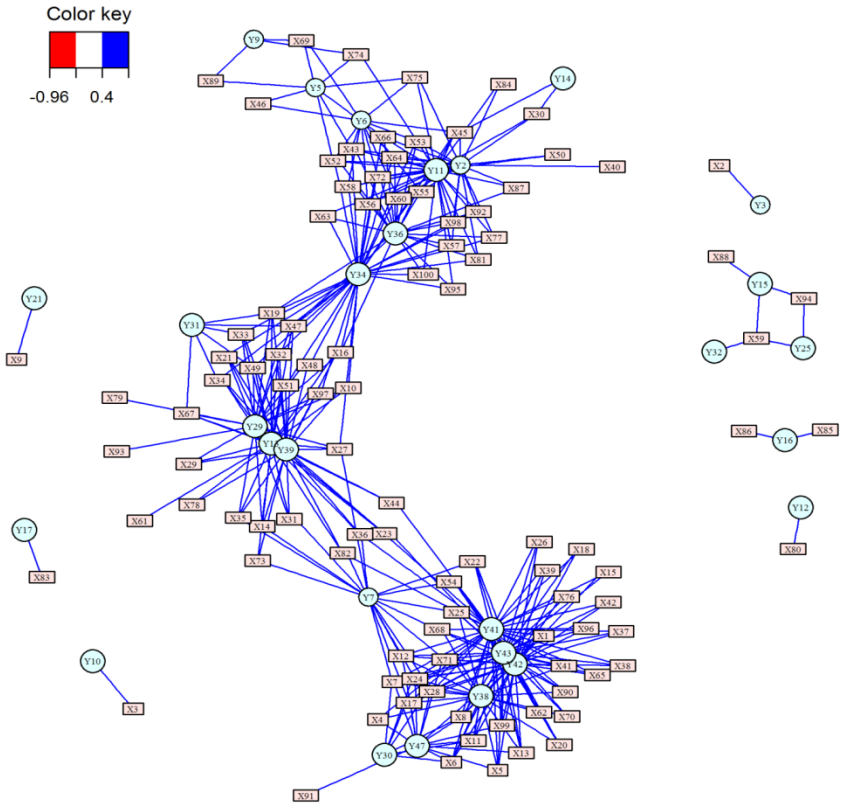
b) Score plot: PC1 vs. PC3

Figure 3.17 Classification Method 1: Hierarchical clustering and heatmap of the 47 FDR significant features (red: low PBB, green: high PBB)



The correlation network in Fig 3.18 shows the relationships between the 47 FDR significant features and others from the initial set of 11,202 features. Features with significant correlations are displayed. It can be seen that there was one large complex cluster with three central hubs and seven smaller clusters of features that were significantly correlated with one another. In the large cluster, the hubs consisted of FDR significant features, and these hubs were surrounded by features that did not significantly discriminate between the PBB exposure classes. All features were positively correlated with each other.

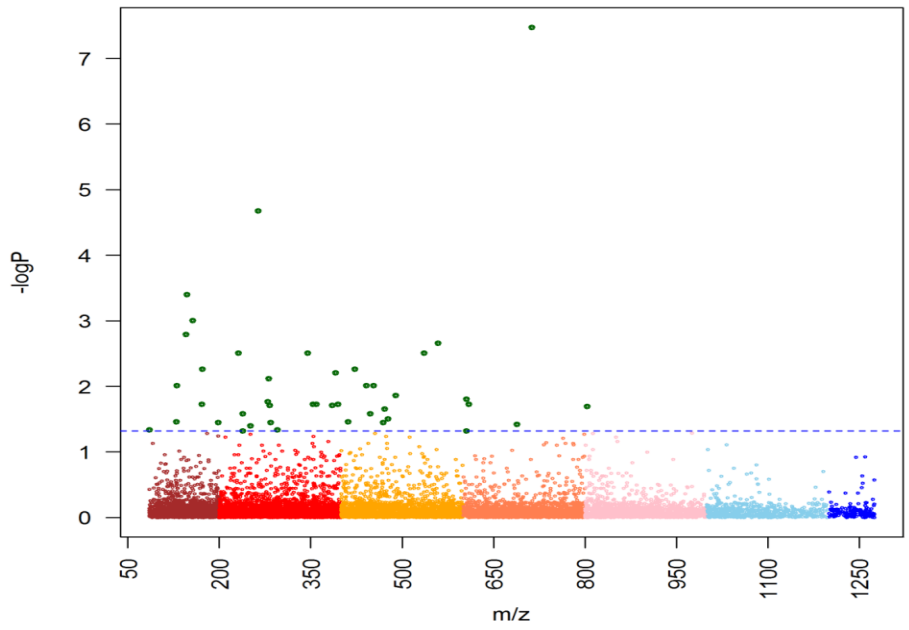
Figure 3.18 Classification Method 1: Correlation network based on the 47 FDR significant features



3.4.2 Feature selection based on Classification Method 2

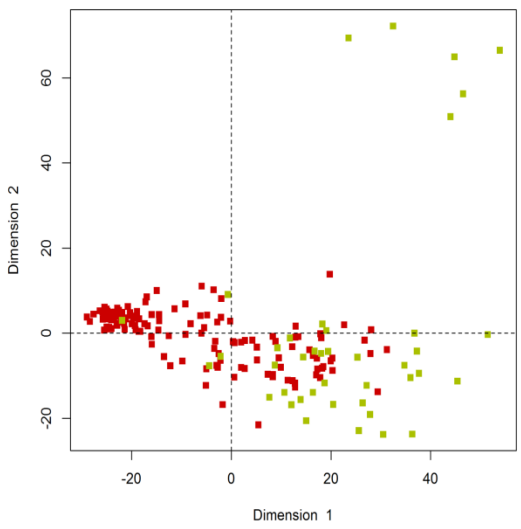
Next, metabolome-wide feature selection was performed using Classification Method 2. The global PCA scree plot showed that approximately 90 % of the variance in the 11,202 features was explained by PC1 with Classification Method 2 (Fig S6.a). Similar to the results from Classification Method 1, pair-wise comparisons of the score plots of the first three PCs showed no clear separation between the two classes based on Classification Method 2 (Fig S6.b, S6.c, S6.d). The global Manhattan plot of the FDR analysis based on Classification Method 2 is shown in Fig 3.19. There were 44 FDR significant features, and the CV accuracy was 92.18 %. Compared to Classification Method 1, there were fewer significant features, but the CV accuracy improved with discriminatory features selected using Classification Method 2.

Figure 3.19 Classification Method 2: Global Manhattan plot of the 11,202 features of the metabolome

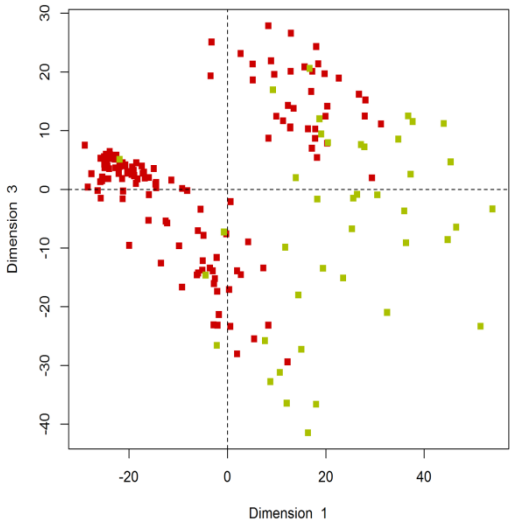


The focused PCA scree plot of the 44 FDR significant features showed that PC1, PC2, and PC3 explained approximately 25 %, 10 %, and 9 % of the variance in the data, respectively (Fig S7.a). The pair-wise comparison of the score plot between PC1 and PC2 shows a separation of samples, where low PBB exposure samples are clustered to the left and high PBB exposure samples are clustered to the right (Fig 3.20.a). The score plot of PC1 and PC3 also shows relatively good separation (Fig 3.20.b), whereas overlap was present in the score plot of PC2 and PC3 (Fig S7.b). The heatmap in Fig 3.21 shows small clusters of high PBB exposure samples on the right side of the horizontal dendrogram and larger clusters of low PBB exposure samples on the left side. In the left bottom corner of the heatmap, there was a cluster of 14 features with low ion intensities among high PBB exposure samples. The box-and-whisker plots of the 44 FDR significant features can be found in Appendix B (Fig S8).

Figure 3.20 Classification Method 2: Focused PCA of the 44 FDR significant features of the metabolome (red: low PBB, green: high PBB)

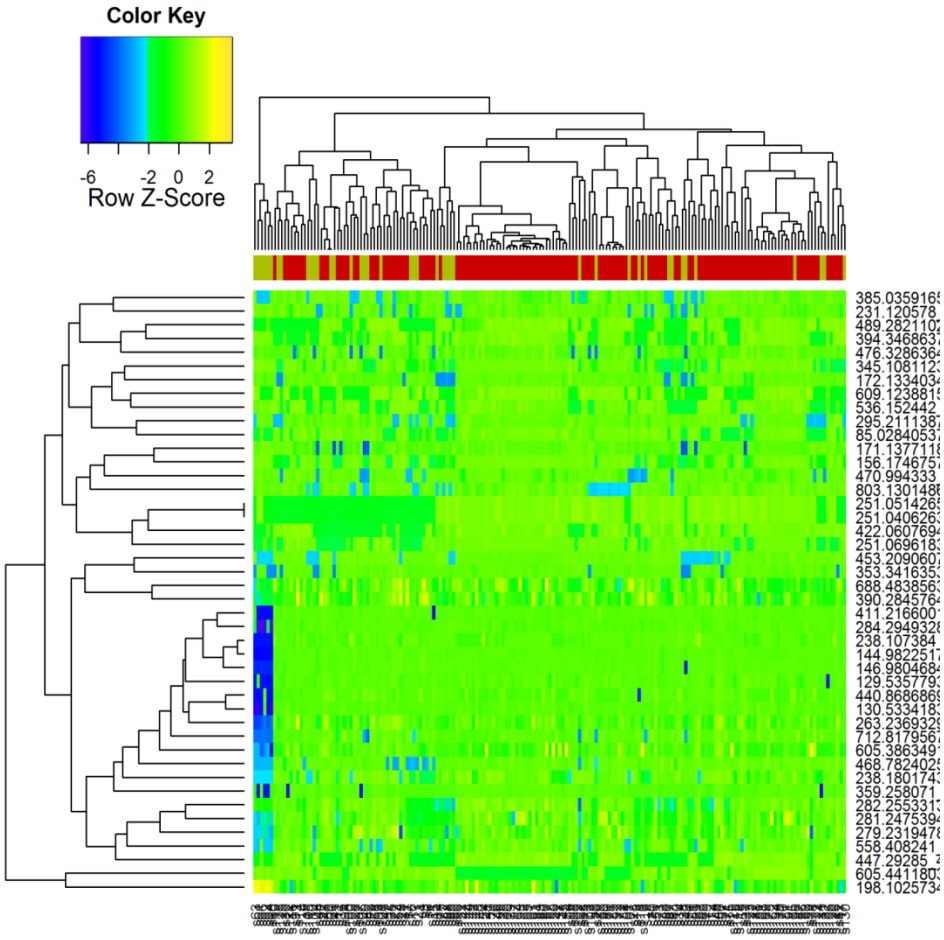


a) Score plot: PC1 vs. PC2



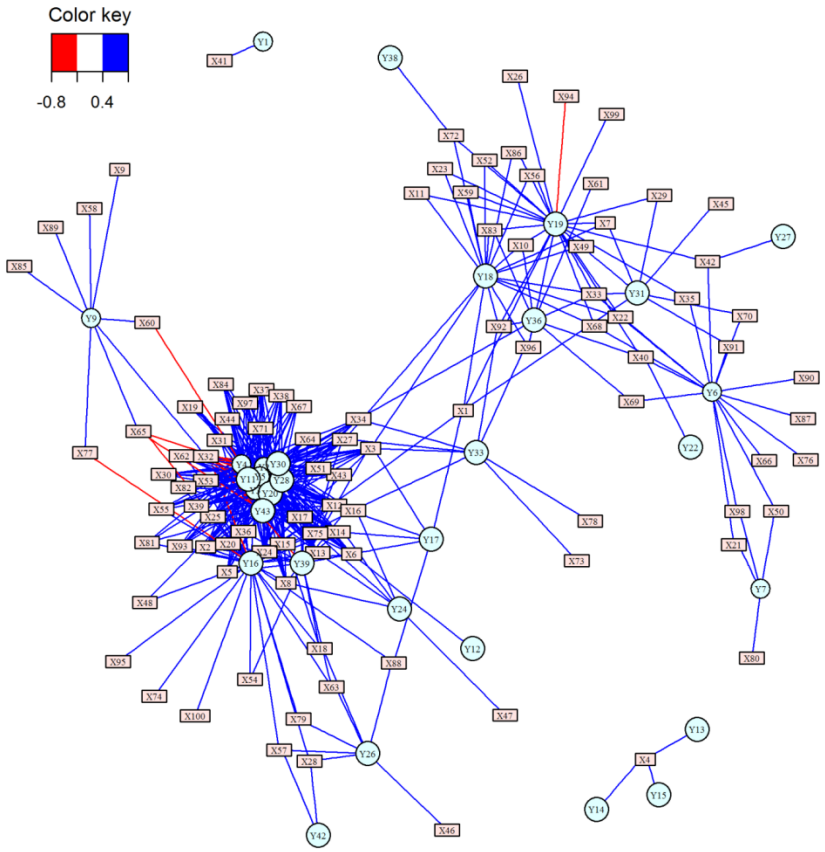
b) Score plot: PC1 vs. PC3

Figure 3.21 Classification Method 2: Hierarchical clustering and heatmap of the 44 FDR significant features (red: low PBB, green: high PBB)



One large and complex cluster of features was identified from the metabolome-wide correlation analysis based on the 44 discriminatory features selected using Classification Method 2 (Fig 3.22). There was one main hub within this cluster. In contrast to the correlation network from Classification Method 1, there were a few positive correlations among features.

Figure 3.22 Classification Method 2: Correlation network based on the 44 FDR significant features



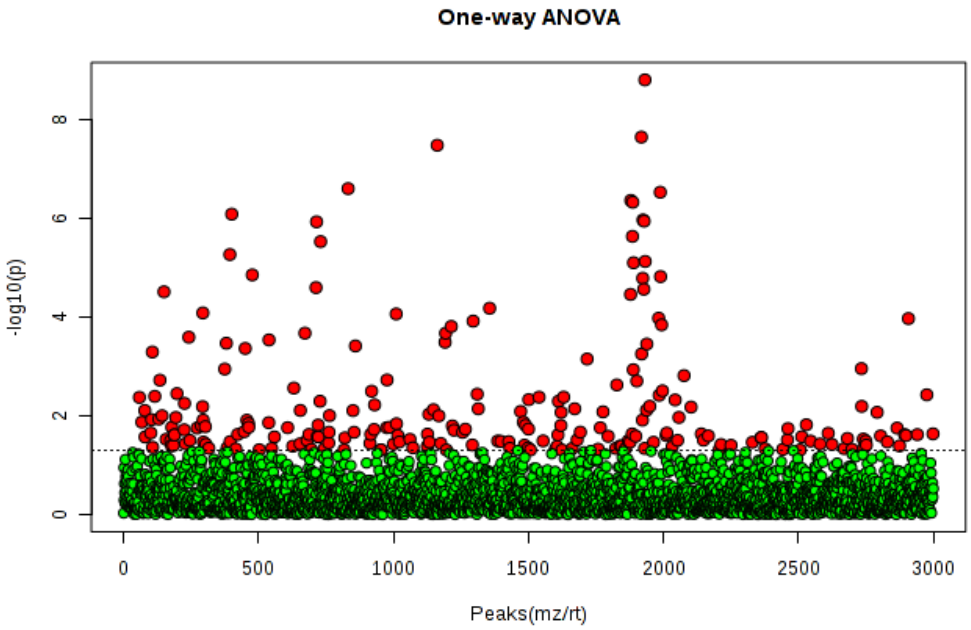
3.4.3 Feature selection based on Classification Method 3

MetaboAnalyst was used to identify features that significantly discriminated between the three different classes of PBB exposure (i.e., low, medium, high) based on Classification Method 3. Data filtering resulted in the reduction of the dataset to 2,999 features with *m/z* values ranging from 85.040 to 1,273.585. Features were normalized using generalized log transformations since the distributions of ion intensity values were skewed for many features (Fig S9).

Univariate statistical analysis was used to perform an analysis of differences in the ion intensity profiles of features between the three classes. One-way ANOVA was used to compare the mean ion intensities of the three exposure classes for each feature. There were 235 features that were identified to be significantly different among the three classes based on ANOVA with a

p-value threshold of p = 0.05 (Fig 3.23). With a FDR significant threshold of q = 0.05, there were 43 significant features. The full list of *m/z* values of the 43 FDR significant features and the results of the post-hoc analysis, Tukey's Honestly Significant Difference, is provided in Appendix C (Table S2).

Figure 3.23 Selection of discriminatory features using one-way ANOVA*

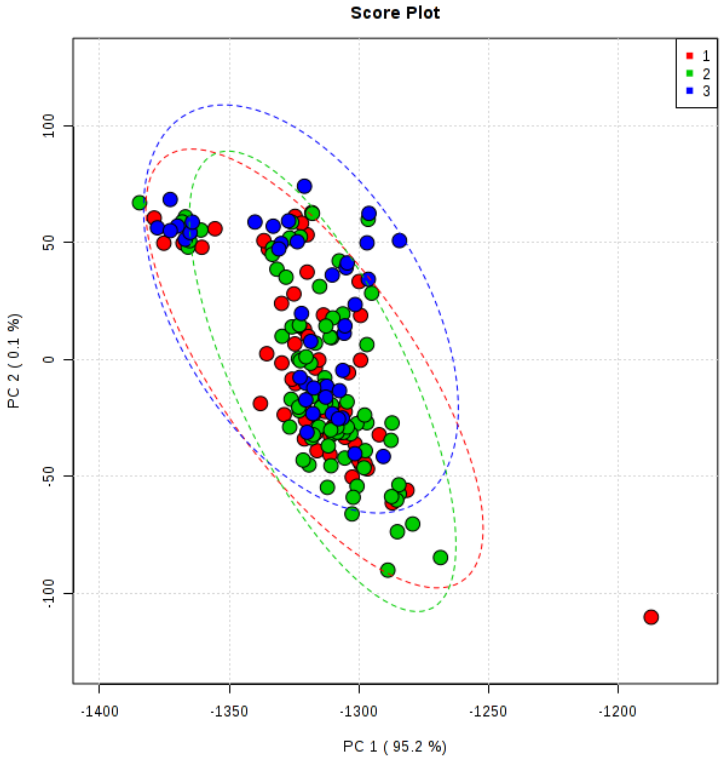


*Significant features above the p-value threshold (p=0.05) are represented in red

PCA was used to detect natural groupings of the dataset without referring to the three-class PBB exposure classes among the 2,999 features included in the analysis. PC1 explained most of the variance by itself (95.2 %, Fig S10.a), and the pair-wise score plot of PC1 and PC2 is shown in Fig 3.24. Samples are represented by dots and corresponding colors of PBB exposure class (i.e., red = low PBB; green = medium PBB; blue = high PBB). The dotted lines represent the 95 % confidence limit for each exposure class. It can be seen from Fig 3.24 that there was no clear separation among the three PBB exposure classes based on the 2,999 features. Similar to the

score plot of PC1 and PC2, pair-wise score plots of PC1 vs. PC3 (Fig S10.b) and PC2 vs. PC3 (Fig S10.c) also showed overlap between the three PBB exposure classes.

Figure 3.24 Classification Method 3: Global PCA score plot (PC1 vs. PC2) of the 2,999 features included in the analysis (red: low PBB, green: medium PBB, blue: high PBB)



PLS-DA was used to see if the data separated well based on the 2,999 features using the prior knowledge of PBB exposure classifications. Pair-wise score plots between the first five PCs and the explained variance of each PC is shown in the corresponding diagonal cell in Fig 3.25. It can be seen that only approximately 1 % of the variance was explained by all of the five PCs. Three clusters of low (red), medium (green), and high (blue) PBB exposure samples can be seen in the score plot between PC1 and PC2 (Fig 3.26.a). There was more overlap between the three classes in the remaining two score plots between PC1 and PC3 (Fig 3.26.b) and between PC2 and PC3 (Fig S11). A permutation test was performed to assess whether there was a statistically significant separation between the three classes (Fig S12). In each permutation, a PLS-DA model

was built between the data and the permuted PBB exposure classes using the first three PCs. The separation distance based on the ratio of the B/W ratio was used as the test statistic for measuring the significance of the class discrimination. Based on 1,000 permutations, there was no statistically significant separation between the three PBB exposure classes based on the 2,999 features included in the analysis at the significance level of 0.05 (p-value = 0.511).

Figure 3.25 Classification Method 3: PLS-DA score plot matrix of the 2,999 features: Amount of variance explained by each PC (red: low PBB, green: medium PBB, blue: high PBB)

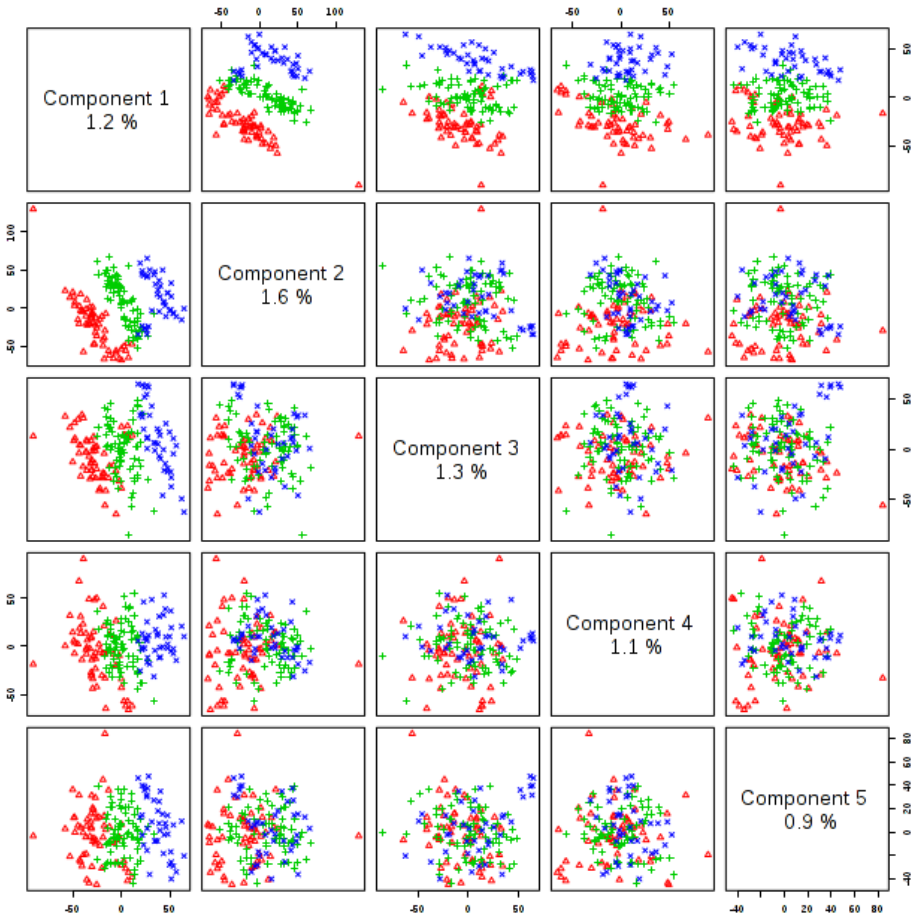
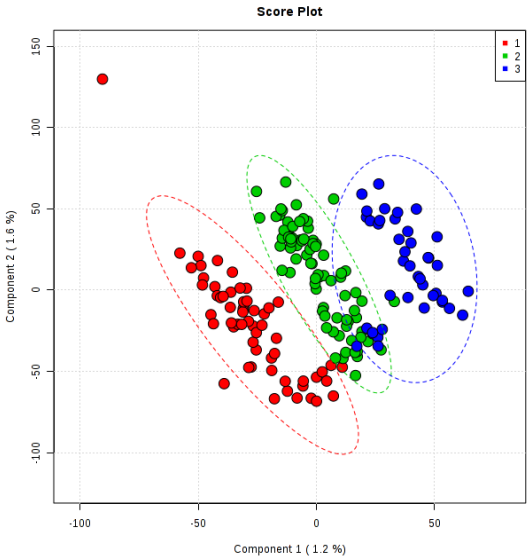
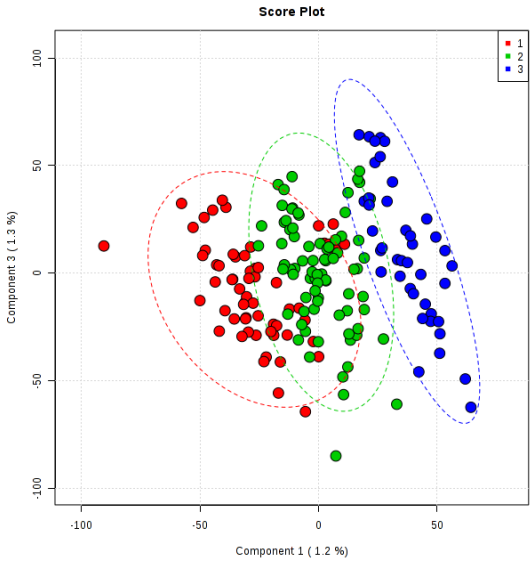


Figure 3.26 Classification Method 3: PLS-DA score plots (red: low PBB, green: medium PBB, blue: high PBB)



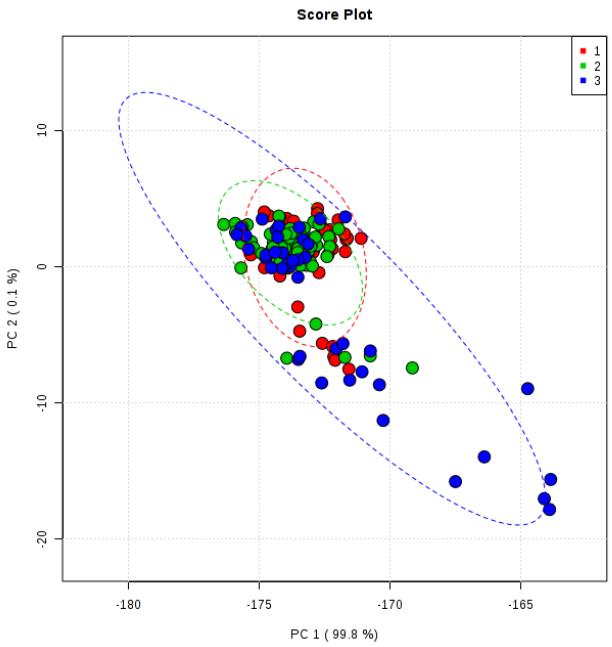
a) Score plot: PC1 vs. PC2



b) Score plot: PC1 vs. PC3

Next, PCA and PLS-DA were performed on the 43 features selected by one-way ANOVA to assess how well these features discriminated between the three PBB exposure classes. PCA showed that most of the variance in the data was explained by the first PC (99.8 %, Fig S13.a). The pair-wise score plot of PC1 vs. PC2 showed overlap between the three PBB exposure classes, but the high exposure class showed a larger, scattered cluster compared to the other two classes (Fig 3.27). Similar patterns were observed in the pair-wise score plots of PC1 vs. PC3 (Fig S13.b) and PC2 vs. PC3 (Fig S13.c) where the high exposure class had a scattered cluster.

Figure 3.27 Classification Method 3: PCA score plot (PC1 vs. PC2) of the 43 selected features (red: low PBB, green: medium PBB, blue: high PBB)



PLS-DA was used to assess whether there was a significant separation of data between the three exposure classes based on the 43 discriminatory features. Pair-wise score plots between the first five PCs and the explained variance of each PC is shown in the corresponding diagonal cell in Fig 3.28. Similar to the results from PCA, pair-wise score plots of PC1 vs. PC2 (Fig 3.29.a) and PC1 vs. PC3 (Fig 3.29.b) from PLS-DA showed a larger, scattered cluster for the high exposure class. In the score plot between PC2 and PC3, the three exposure classes showed equally dispersed, overlapping clusters (Fig S14). The statistical significance of the separation between the three exposure classes based on the 43 discriminatory features was tested with a permutation test. PLS-DA models were built using five PCs, which was determined as the optimal number of components based on the cross-validated sum of squares captured by the model (Fig S15). Based on 1,000 permutations, there was a statistically significant separation between the three PBB exposure classes based on the 43 discriminatory features (p-value < 0.001, Fig 3.30).

Figure 3.28 Classification Method 3: PLS-DA score plot matrix of the 43 selected features: Amount of variance explained by each PC (red: low PBB, green: medium PBB, blue: high PBB

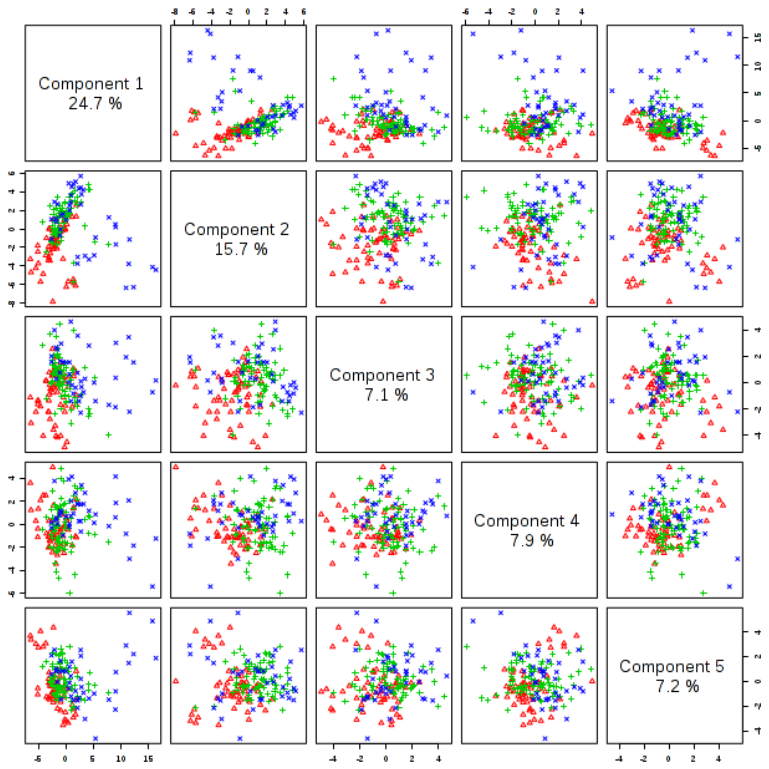
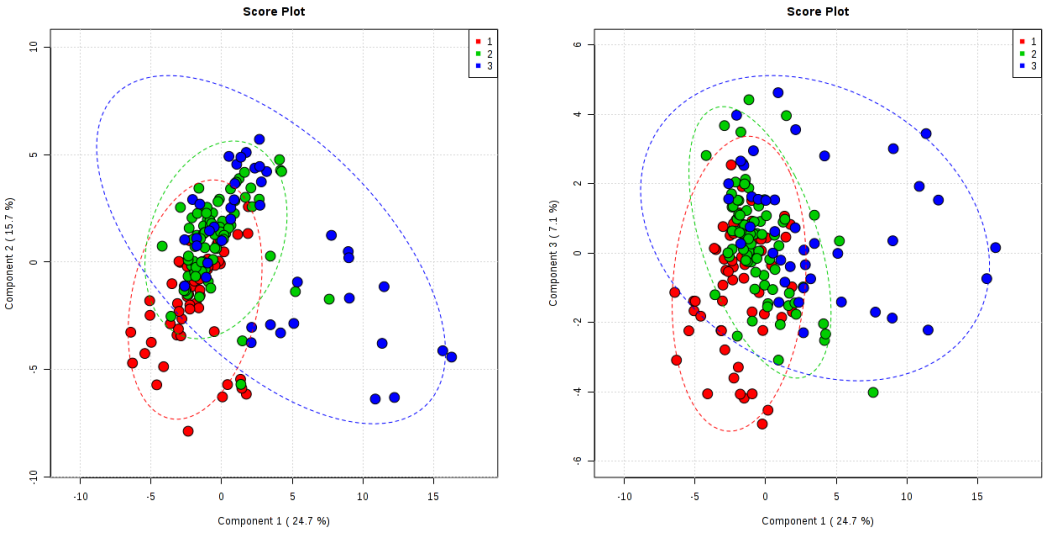


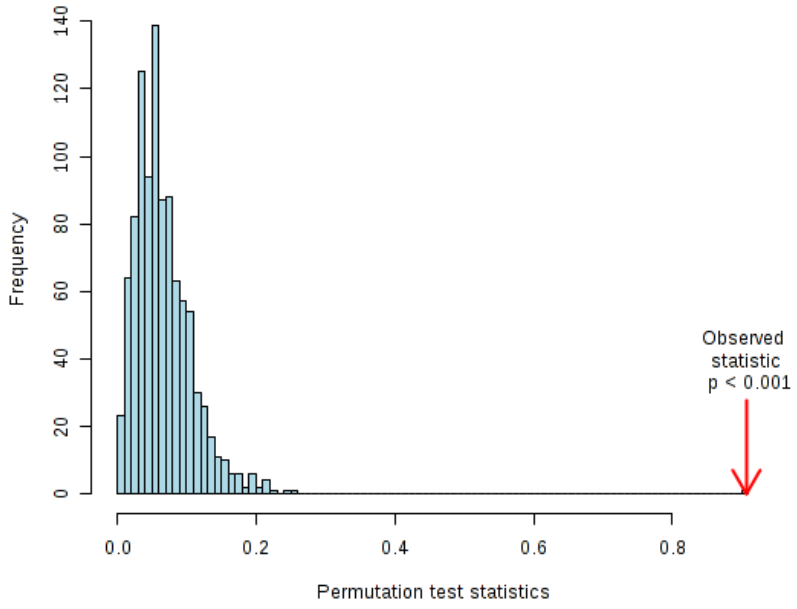
Figure 3.29 Classification Method 3: PLS-DA score plots of the 43 selected features (red: low PBB, green: medium PBB, blue: high PBB)



a) Score plot: PC1 vs. PC2

b) Score plot: PC1 vs. PC3

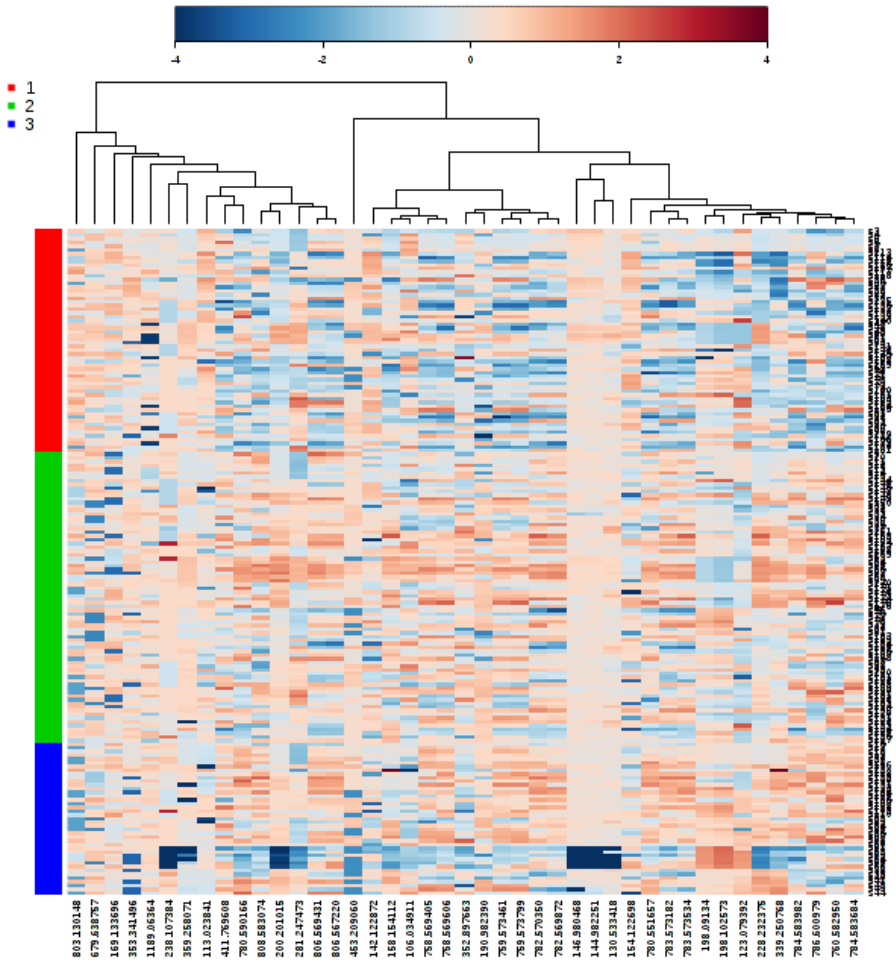
Figure 3.30 Classification Method 3: PLS-DA model validation by permutation tests based on separation distance (based on the 43 selected features)



A heatmap was generated to visually identify patterns or groupings of features that were expressed differently among the three PBB exposure classes (Fig. 3.31). The heatmap was generated using Pearson's correlation coefficients as distance measures and Ward's method as the clustering algorithm. Samples are displayed on the vertical axis from low PBB exposure (red) on the top through medium PBB exposure (blue) to high PBB exposure (green) on the bottom. The *m/z* values of the 43 discriminatory features are displayed on the horizontal axis. Hierarchical clustering was performed on the features, and the dendrogram on the horizontal axis shows the observed clusters of features. The horizontal color key indicates the ion intensities of the features (blue: lowest, red: highest). There were three clusters of features for which the low PBB exposure class generally showed lower ion intensities compared to the medium and high exposure classes. The first cluster consisted of features with *m/z* values of 411.770 to 806.567 from the left of the horizontal axis. The second cluster consisted of features with *m/z* values from 158.154 to 782.570, and the third from 780.552 to 784.584 from the left of the horizontal axis. Therefore, features

included in these three clusters on the heatmap may be particularly important for discriminating the low PBB exposure class.

Figure 3.31 Classification Method 3: Hierarchical clustering and heatmap of the top 50 features selected based on ANOVA (red: low PBB, green: high PBB)

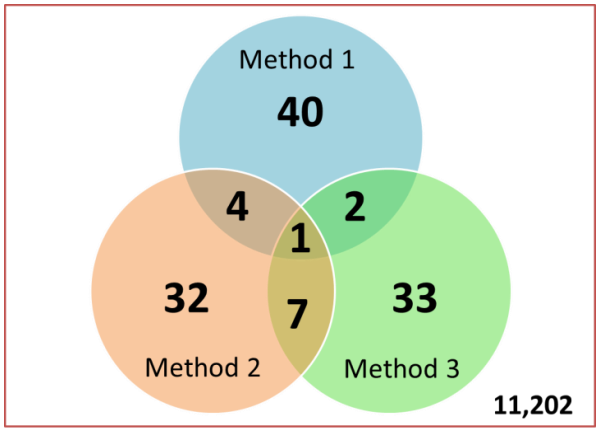


3.4.4 Mapping selected features to metabolites and metabolic pathways

Metabolomic database searches were used to tentatively identify the significant features that discriminated between the different PBB exposure classes. Classification Methods 1, 2, and 3 identified 47, 44, and 43 significant features, respectively. Of these significant features, 1 feature

was selected by all three methods, and 13 features were selected by two of the three methods (Fig 3.32). With the three classification methods combined, a total of 119 features were found to discriminate between PBB exposure classes among the 11,202 features detected in the metabolome. There were 769 potential matches when the 14 mutually selected features were searched using Metlin, and 32 of these were identified with KEGG IDs. The 32 metabolites were then analyzed using MetaboAnalyst's MSEA and MetPA.

Figure 3.32 Comparison of metabolome-wide discriminatory features selected by Classification Method 1, 2, and 3



Of the 32 metabolites with KEGG IDs, 10 were exactly matched with compounds in MetaboAnalyst's metabolite set library (Table S3). Over-representation analysis was performed on the matched metabolites using the hypergeometric test (Table 3.4). Fig 3.33 shows the summary plot of MSEA performed on the 14 discriminatory features that were mutually selected by at least two of the three PBB exposure classification methods. There were two metabolite sets with matches, and the metabolite set that had the most number of matches was the arachidonic acid metabolite set. There were 4 matched metabolites out of the 37 compounds included in the arachidonic acid metabolite set. Although this was identified as a statistically significant match

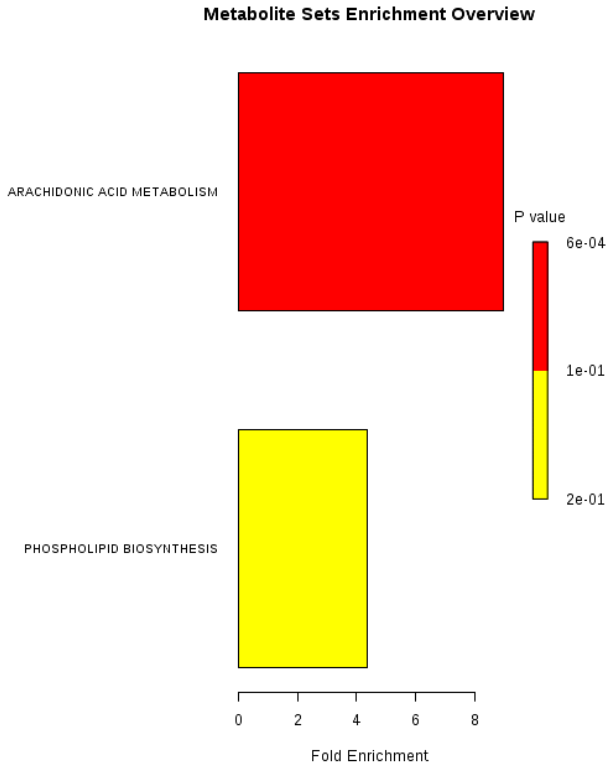
(FDR adjusted p-value = 0.0476), this is thought to have been caused by the small number of metabolites that were analyzed with MSEA.

Table 3.4 MSEA: Results of the over-representation analysis for the 14 discriminatory features mutually selected by at least two of the three classification methods

Metabolite set	Total*	Expected*	Hits*	Raw p*	Holm p*	FDR p*	Match*
ARACHIDONIC ACID METABOLISM	37	0.448	4	< 0.001	0.048	0.048	14,15-DiHETrE; 8,9DiHETrE; 11,12DiHETrE; 5,6-DHET
PHOSPHOLIPID BIOSYNTHESIS	19	0.230	1	0.209	1.000	1.000	PC(16:0/16:0)

* Total number of compounds in the metabolite set (Total), number of matched compounds (Hits), the original p-value calculated from the enrichment analysis (Raw p), the adjusted p-value using the Holm-Bonferroni method (Holm p), the adjusted p-value using the FDR (FDR p), matched metabolites (Match)

Figure 3.33 Summary plot of the Metabolite Set Enrichment Analysis (MSEA) of the 14 discriminatory features mutually selected by at least two of the three classification methods



The 10 metabolites that were matched with MetaboAnalyst's metabolite library were next analyzed with MetPA. Fig 3.34 shows the summary plot and Table 3.5 shows the results of MetPA performed on the 14 discriminatory features that were mutually selected by at least two of the three PBB exposure classification methods. In Figure 3.34, there are no matched pathways along the diagonal y = x line, indicating that there were no matched pathways that had a high number of matched compounds and for which matched compounds had a high impact on the pathway. Although the arachidonic acid metabolic pathway had the highest number of matched compounds, these compounds were identified to have little impact on the pathway itself (impact value = 0.000). There were 5 matched metabolites out of the 62 compounds that are included in the arachidonic acid metabolic pathway (FDR adjusted p-value < 0.001). Although the 2 matched metabolites in the glycerophospholipid metabolic pathway had the highest impact on the pathway (impact value = 0.151), this match was not statistically significant (FDR adjusted p-value = 0.424).

Figure 3.34 Summary plot of the Metabolic Pathway Analysis (MetPA) of the 14 discriminatory features mutually selected by at least two of the three classification methods

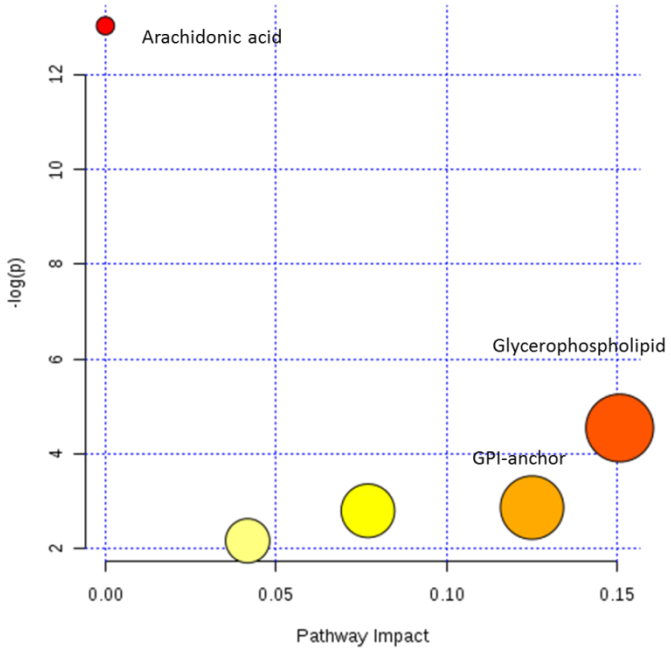


Table 3.5 MetPA: Results of the enrichment analysis and topology analysis of the 14 discriminatory features mutually selected by at least two of the three classification methods

Pathway	Total*	Expected*	Hits*+	Raw p*	Holm p*	FDR p*	Impact*
Arachidonic acid metabolism	62	0.258	5	< 0.001	< 0.001	< 0.001	0.000
Glycerophospholipid metabolism	39	0.162	2	0.011	0.838	0.424	0.151

* Total number of compounds in the metabolite set (Total), number of matched compounds (Hits), the original p-value calculated from the enrichment analysis (Raw p), the adjusted p-value using the Holm-Bonferroni method (Holm p), the adjusted p-value using the FDR (FDR p); the pathway impact value (Impact)
+ Only pathways with more than one matched compound are shown in the table

Analysis with MSEA and MetPA were similarly performed for the 119 features that were collectively selected from the three classification methods. When these features were searched using Metlin, there were 5,205 potential matches and 273 of these were identified with KEGG IDs. The 273 metabolites were analyzed with MSEA and MetPA.

Of the 273 metabolites with KEGG IDs, 126 were matched with compounds in MetaboAnalyst's metabolite set library (Table S4). Fig 3.35 shows the summary plot of MSEA performed on the combined 119 discriminatory features that were selected by the three classification methods. Results of the over-representation analysis are shown in Table 3.6. There were four metabolite sets with at least 2 matches. The matched metabolite set with the lowest p-value was the α-linolenic and linoleic acid metabolite set with 3 matched compounds, but this match was statistically not significant (FDR adjusted p-value = 1.000). The metabolite set with the highest number of matched compounds was the arachidonic acid metabolite set but this was also statistically insignificant (FDR adjusted p-value = 1.000).

Figure 3.35 Summary plot of the Metabolite Set Enrichment Analysis (MSEA) of the 119 discriminatory features selected by the three classification methods

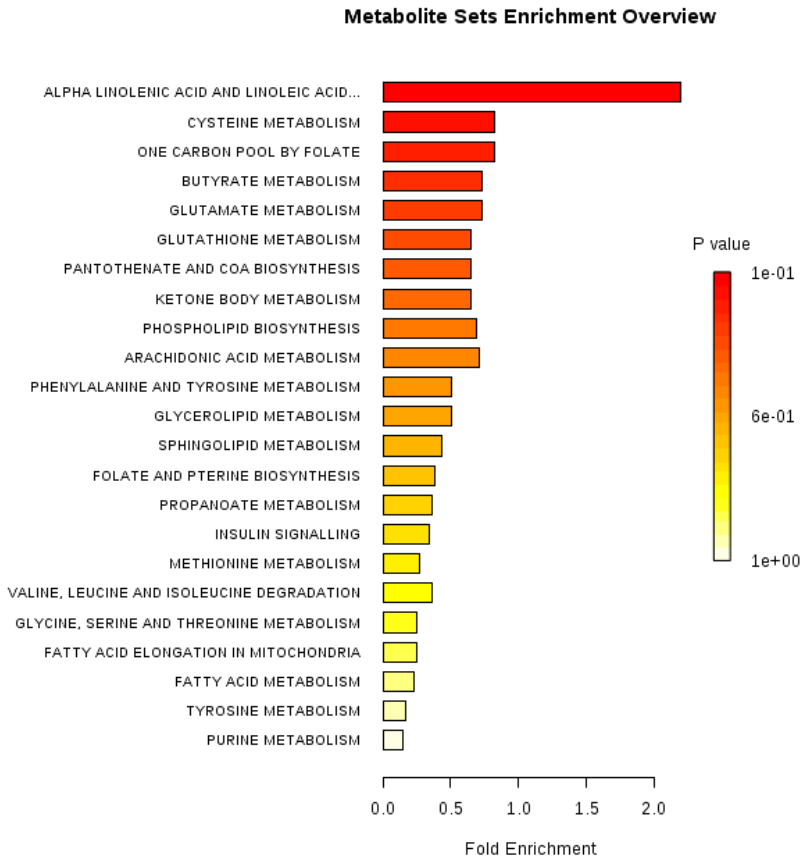


Table 3.6 MSEA: Results of the over-representation analysis for the 119 discriminatory features selected by the three classification methods

Metabolite set	Total*	Expected*	Hits*	Raw p*	Holm p*	FDR p*	Match*
ALPHA LINOLENIC ACID AND LINOLEIC ACID METABOLISM	9	1.370	3	0.146	1.000	1.000	Linoleic acid; Alpha-Linoleic acid; Gamma-Linoleic acid
GLUTAMATE METABOLISM	18	2.750	2	0.788	1.000	1.000	Gamma-Glutamylcysteine; Succinic acid semialdehyde
PHOSPHOLIPID BIOSYNTHESIS	19	2.900	2	0.814	1.000	1.000	PC(16:0/16:0); PS(16:0/16:0)
ARACHIDONIC ACID METABOLISM	37	5.650	4	0.844	1.000	1.000	14,15- DHET; 8,9-DHET; 11,12-DHET; 5,6-DHET
VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	36	5.500	2	0.983	1.000	1.000	Acetoacetic acid; (S)-Methylmalonic acid semialdehyde

* Total number of compounds in the metabolite set (Total), number of matched compounds (Hits), the original p-value calculated from the enrichment analysis (Raw p), the adjusted p-value using the Holm-Bonferroni method (Holm p), the adjusted p-value using the FDR (FDR p), matched metabolites (Match)
+ Only metabolite sets with more than one matched compound are shown in the table

When the 119 features were analyzed with MetPA, there was one metabolic pathway with a significant match. In the summary plot (Fig 3.36), the linoleic acid metabolic pathway lies along the y = x diagonal line in the right top corner, indicating that this pathway had a high number of matching compounds (FDR adjusted p-value < 0.001) and that these matched compounds also have a high impact on the pathway (impact value = 0.923). Table 3.7 provides the full result of MetPA performed on the 119 features that were selected by the three PBB exposure classification methods. The linoleic acid metabolic pathway had 10 matching compounds out of 15 compounds included in the pathway. Although statistically insignificant, the valine, leucine, and isoleucine biosynthesis metabolic pathway had the second highest proportion of matched compounds, followed by the arachidonic acid metabolic pathway. However, it can be seen from the summary plot that the linoleic acid metabolic pathway was the most important

matched pathway (Fig 3.36). The results of MetPA are concordant to the results from MSEA, which identified that the 119 features matched best with the α-linolenic and linoleic acid metabolite set.

Figure 3.36 Summary plot of the Metabolic Pathway Analysis (MetPA) of the 119 discriminatory features selected by the three classification methods

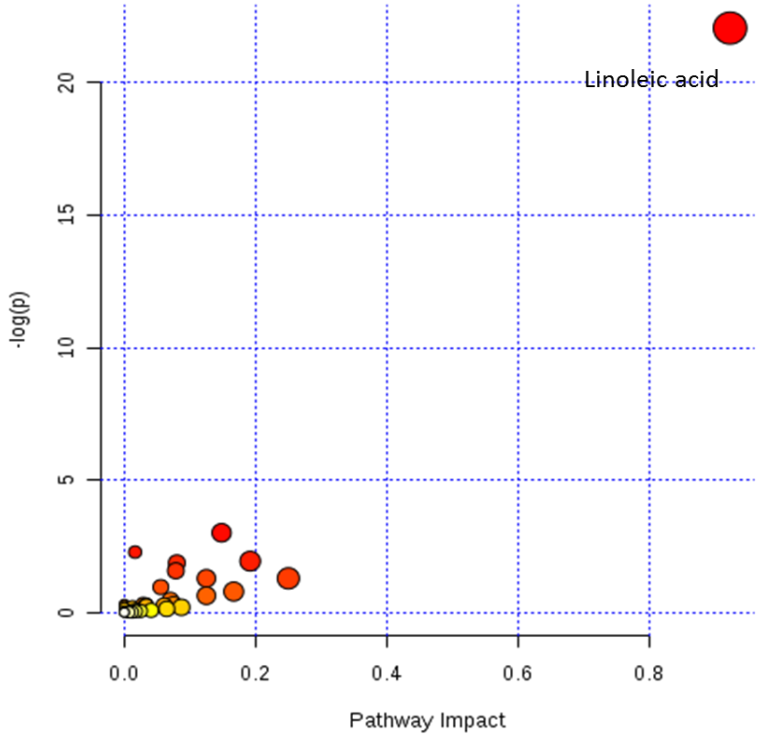


Table 3.7 MetPA: Results of the enrichment analysis and topology analysis of the 119 discriminatory features selected by the three classification methods

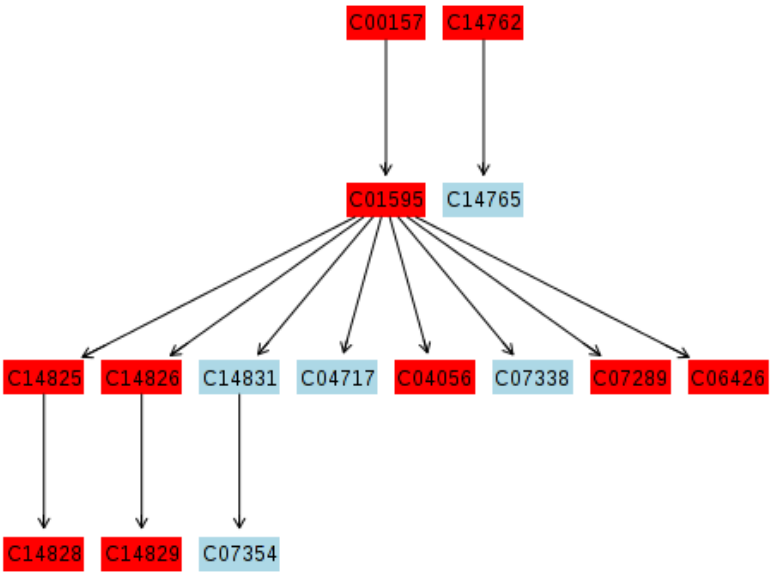
Pathway	Total*	Expected*	Hits*+	Raw p*	Holm p*	FDR p*	Impact*
Linoleic acid metabolism	15	0.785	10	< 0.001	< 0.001	< 0.001	0.923
Valine, leucine and isoleucine biosynthesis	27	1.413	4	0.049	1.000	1.000	0.148
Arachidonic acid metabolism	62	3.246	6	0.102	1.000	1.000	0.016
Glycerophospholipid metabolism	39	2.042	4	0.144	1.000	1.000	0.192
Butanoate metabolism	40	2.094	4	0.154	1.000	1.000	0.080
Ascorbate and aldarate metabolism	45	2.356	4	0.207	1.000	1.000	0.078
Propanoate metabolism	35	1.832	3	0.276	1.000	1.000	0.125
alpha-Linolenic acid metabolism	29	1.518	2	0.454	1.000	1.000	0.167
Valine, leucine and isoleucine degradation	40	2.094	2	0.629	1.000	1.000	0.070
Glycine, serine and threonine metabolism	48	2.513	2	0.727	1.000	1.000	0.000
Fatty acid metabolism	50	2.617	2	0.747	1.000	1.000	0.033
Pentose and glucuronate interconversions	53	2.774	2	0.776	1.000	1.000	0.033
Cysteine and methionine metabolism	56	2.931	2	0.802	1.000	1.000	0.012
Tyrosine metabolism	76	3.978	2	0.916	1.000	1.000	0.011
Drug metabolism - cytochrome P450	86	4.502	2	0.947	1.000	1.000	0.025

* Total number of compounds in the metabolite set (Total), number of matched compounds (Hits), the original p-value calculated from the enrichment analysis (Raw p), the adjusted p-value using the Holm-Bonferroni method (Holm p), the adjusted p-value using the FDR (FDR p); the pathway impact value (Impact)
+ Only pathways with more than one matched compound are shown in the table

Based on the combined 119 discriminatory features that were selected using the three classification methods, results of MSEA and MetPA suggest that these features include metabolites in the linoleic acid metabolic pathway. The 10 metabolites that were matched with compounds in this pathway were: phosphatidylcholine (KEGG ID: C00157); 13S-hydroxyoctadecadienoic acid (C14762); linoleic acid (C01595); 9,10-epoxyoctadecenoic acid (C14825); 9,10-DHOME (C14828); 12,13-EpOME (C14826); 12,13-DHOME (C14829); bovinic

acid (C04056); crepenynate (C07289); and γ-linolenic acid (C06426) (Fig 3.37). Phosphatidylcholine was matched to 13 features with different *m/z* values (Table S4). Of the 10 compounds, 7 were matched to the same 1 feature with the *m/z* value = 279.232 (13S-hydroxyoctadecadienoic acid; 9,10-epoxyoctadecenoic acid; 9,10-DHOME; 12,13-EpOME; 12,13-DHOME; crepenynate; and γ-linolenic acid). Linoleic acid and bovinic acid were matched on the same three features with *m/z* value = 263.237, 281.248, 281.247. Since many of the features were matched to multiple metabolites, there is uncertainty in the true identity of the features.

Figure 3.37 The linoleic acid metabolic pathway (matched metabolites indicated in red, metabolites represented by KEGG IDs)



4. Discussion

The purpose of the present study was to investigate the metabolic effects of PBB exposure to identify biomarkers associated with PBB exposure and its potential impact on disease risk. PBB exposure was classified into two or three categories in order to compare the metabolic profiles and identify features that discriminate between exposure groups. In the first part, a total of 44 discriminatory features were selected from features that were strongly correlated with plasma PBB concentrations. In the second part, a total of 119 discriminatory features were collectively selected from the 11,202 features that were detected in the metabolome. This study demonstrates the ability of current analytical technologies and methods in metabolomics to discover biomarkers along the exposure-disease continuum, and further, to explore the exposome.

In metabolomics studies, it is often expected that the majority of observed changes in features are within normal physiological variations (i.e., background noise), and that only a small proportion of these are associated with the exposure or disease status of interest.[30] The aim of the present study was to identify these 'key features' that discriminate different classes of PBB exposure. This was done by selecting discriminatory features based on the FDR threshold of $q = 0.05$, and confirming the separation of PBB exposure classes using multivariate statistical analyses (i.e., PCA and PLS-DA) and visualization methods (hierarchical clustering and heatmaps). On average, approximately 40 discriminatory features were successfully identified from the initial set of 11,202 features that were detected in the plasma samples for each analysis that was performed in this study. This demonstrates that high-resolution metabolic profiling with LCMS is a powerful means of capturing this small window of key features.

Although discriminatory features were successfully identified between subjects exposed to different levels of PBB in the Michigan cohort, it is possible that there may be more discriminatory features when comparing subjects exposed and not exposed to PBB. PBBs are artificial compounds and do not occur naturally in the environment.[18] Subjects with current body

burdens of PBB were either exposed previously when PBB was still being produced or from bioaccumulation through the food chain. Therefore, unexposed subjects should have little or no detectable level of PBB in their blood. Thus, comparing metabolic profiles with a control group consisting of subjects unexposed to the Michigan PBB accident may have resulted in the discovery of an increased number of important features associated with PBB exposure.

In order to compare metabolic profiles and identify discriminatory features, PBB exposure was classified into groups using three classification methods. The use of these different classification methods gave rise to differences in the selected discriminatory features, although there were also features that were mutually selected by more than one PBB exposure classification method. This shows that the selection of discriminatory features depends on the choice of the threshold by which PBB exposure is classified. The threshold concentrations used in the present analyses were based on the distribution of the measured plasma PBB concentrations. The inclusion of a control group with little or no detectable level of PBB would have been informative to compare the changes in metabolic profiles with increasing PBB exposure.

Discriminatory features were selected based on two different approaches, and this resulted in the identification of two discrete sets of selected features. In the first part, a total of 44 discriminatory features were selected from the 73 features that were strongly correlated with the measured plasma PBB concentrations. The second approach identified a total of 119 discriminatory features from the complete set of 11,202 features that were detected in the metabolome. Of the two sets of discriminatory features identified from the two approaches, only one feature was selected by both approaches (*m/z* value = 710.474), leaving a total of 162 unique discriminatory features. The 162 discriminatory features that were collectively identified from the two approaches were tentatively identified to be metabolites predominantly involved in lipid metabolism and xenobiotic metabolism (Fig S16). Whereas the first approach focused on identifying features directly associated with PBB exposure, the latter was based on a metabolome-wide approach. As can be seen from the metabolome-wide correlation networks (Fig

3.5; 3.10; 3.18; 3.22), metabolites are interrelated with one another in a complex network. Therefore, the metabolome-wide approach allows for the investigation of indirect effects of PBB exposure. With the metabolome-wide approach, potential signs of disease risk can be investigated by exploring multiple metabolic pathways. With the two approaches combined, it is possible to discover biomarkers associated with both exposure and disease risk, thereby encompassing the exposure-disease continuum. Since most of the discriminatory features associated with PBB exposure did not overlap with features identified from the metabolome-wide approach, this indicates that PBB exposure may induce alterations in multiple metabolic pathways both directly and indirectly.

Based on the results of MSEA and MetPA, the discriminatory features associated with PBB exposure were suggested to be metabolites in the arachidonic acid metabolic pathway, and discriminatory features from the metabolome-wide approach were identified to be metabolites in the linoleic acid metabolic pathway. Matched metabolites in the arachidonic acid metabolic pathway included 14,15- DHET; 8,9- DHET; 11,12- DHET; and 5,6-DHET (DHET: dihydroxyeicosatrienoic acid). DHETs are cytochrome P450 (CYP)-mediated metabolites of arachidonic acid, which is a polyunsaturated omega-6 fatty acid found in human fat, liver, brain, and glandular organs.[32] Increased levels of CYP arachidonic acid metabolites have previously been associated with diabetes, pregnancy, hepatorenal syndrome, cyclosporine-induced nephrotoxicity, and alcohol-induced liver disease.[33] It is possible that PBB exposure induces alterations in this pathway since previous literatures have found PBBs to induce hepatic CYP-dependent monooxygenases, which are Phase I xenobiotic metabolizing enzymes.[34] Linoleic acid (an omega-6 fatty acid) and α-linolenic acid (an omega-3 fatty acid) are two essential fatty acids that are the precursors for the synthesis of a variety of other unsaturated fatty acids, including arachidonic acid.[35] In the first step of linoleic acid metabolism, linoleic acid is converted to γ-linolenic acid, which is then converted to dihomo-γ-linolenic acid and finally to arachidonic acid. Therefore, the arachidonic acid and linoleic acid metabolic pathways are both involved in the

regulation of signaling molecules called eicosanoids, which are closely related to various diseases such as cardiovascular disease, stroke, myocardial infarcation, asthma, hyptertension, and cancer when disruptions in levels occur.[31]

Although some of the discriminatory features were tentatively identified as metabolites in the arachidonic acid and linoleic acid metabolic pathways, there is much uncertainty in these findings. The four metabolites in the arachidonic acid metabolic pathway were matched based on one feature. In other words, one *m/z* value was matched to four metabolites. There were also redundancies in the *m/z* values of the metabolites matched in the linoleic acid metabolic pathway. This means that only a few features were matched to multiple metabolites in a given pathway, thereby introducing a false enrichment of metabolites in that pathway. Targeted analyses are necessary to confirm the identity of these discriminatory features.

Feature annotation and pathway mapping were also impeded by the lack of finding matches in metabolomic databases. In both Part 1 and 2 of the present study, less than half of the identified discriminatory features were successfully matched in MetaboAnalyst's MSEA. Of the 32 PBB associated discriminatory features that were mutually selected by the two classification methods in Part 1, 12 features (37.5 %) were matched to compounds in MSEA. In Part 2, 51 (42.9 %) of the 119 metabolome-wide discriminatory features had matches in MSEA. Since more than half of the features were unable to be matched to compounds in the database, this led to loss of information. In addition, this may have introduced bias in the results by matching features to metabolites that are more common or measureable with current technology. Discriminatory features identified in the analyses may have been matched to arachidonic acid and linoleic acid metabolic pathways since these pathways include commonly known metabolites with higher abundance. It is possible that the unidentified metabolites may have greater implications on the health effects of PBB exposure, and that this information was neglected in the process of feature annotation and pathway mapping. Despite the uncertainty in the true identity of features, the presence of discriminatory features between different classes of PBB exposure implies that there

are measureable alterations in metabolic processes related to different levels of exposure to PBB. This gives more reason to investigate the unidentified discriminatory features. Loss of information during feature annotation and pathway mapping raises the next major challenge of analyzing unidentified metabolites and low abundance features for metabolomic studies.[36]

Finally, this study demonstrated that advanced methods and technology are currently available for the investigation of the human exposome. The present study investigated the long-term effect of PBB exposure on human metabolic pathways using metabolic profiling and computational methods. Non-targeted investigations of metabolic profiles were performed with LCMS, which was able to detect over 11,000 features. Only small amounts of samples were required and sample processing was minimal in this study, thereby allowing high-throughput. With recent advances in technology, it is possible to detect metabolites with high sensitivity, and in principle, the coverage can be increased to detect over 10,000 chemicals.[37] This is particularly essential for exposomics, which aims to holistically investigate multiple known and unknown environmental exposures to account for the multifactorial nature of disease etiology. Moreover, quantitative measures of current body burdens of PBB were used in the present study to investigate the long-term effects of PBB exposure, which occurred approximately 40 years ago. This accounts for the temporal lag in disease onset, which is a common characteristic of many of the chronic diseases that are of concern today. Although the data extracted from LCMS is high-dimensional and complex, various bioinformatics and computational methods are currently available to reduce the complexity and facilitate interpretations of the data. In the present analysis, R software packages for metabolomics analysis and MetaboAnalyst were primarily used to reduce the complexity of the metabolomics data by extracting information on features that most contributed to the variation among samples with different levels of PBB exposure. By using these high-resolution metabolic profiling and computational tools, exposomics can be employed to study the complex system of human health, where a single exposure can have multiple effects and multiple exposures can have a single health effect.

5. Conclusion

In the present study, a metabolomics analysis was applied to newly collected plasma samples collected from subjects in the Michigan PBB cohort who were exposed to PBB in the 1970s or were descendants of those exposed. High-resolution metabolic profiling along with bioinformatics and computational tools were used to study the long-term effect of PBB exposure on human metabolic pathways. Differences in the metabolic profiles of subjects with low to high current body burdens of PBB were detected, and these differences were identified as discriminatory features. Although the identity of these discriminatory metabolites require further investigation, the study demonstrated that metabolomics methods can be applied to discover biomarkers related to the exposure, or biomarkers indicative of potential disease risk. Moreover, the study provides evidence that omics technology and methods, such as those used in the present metabolomics analysis, are a promising means of exploring the human exposome. Although conventional methods have failed to account for the multifactorial etiology and long latency of many human diseases, exposomics holds much promise in providing new insights into disease prevention and treatment by accounting for these factors. Conventional methods have also overlooked the effect of unrecognized exposures, but exposomics provides a way to simultaneously explore thousands of environmental exposures, both known and unknown. Furthermore, exposomics has the potential to elucidate the complex interplay between genes and the environment that affect the complex system of human health.

Appendix A. IRB Approval Letter

TO: Michele Marcus
Principal Investigator
Epidemiology

DATE: October 29, 2013

RE: **Notification of Amendment Approval**
AM20_IRB00045959
IRB00045959
Brominated Flame Retardants: Multigenerational Endocrine Disruption?

Thank you for submitting an amendment request. The Emory IRB reviewed and approved this amendment under the expedited review process on **10/29/2013**. This amendment includes the following:

Personnel Change only: Adding Tamar Wainstock as other Emory study staff.
Important note: If this study is NIH-supported, you may need to obtain NIH prior approval for the change(s) contained in this amendment before implementation. Please review the NIH policy directives found at the following links and contact your NIH Program Officer, NIH Grants Management Officer, or the Emory Office of Sponsored Programs if you have questions.
Policy on changes in active awards: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-129.html
Policy on delayed onset awards: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-130.html
In future correspondence with the IRB about this study, please include the IRB file ID, the name of the Principal Investigator and the study title. Thank you.
Sincerely,
Donna Thomas
Administrative Assistant
This letter has been digitally signed

CC	Fershteyn	Zarina	Psychiatry - Main
	Barr	Dana	Envir & occup Health
	Howards	Penelope	Epidemiology
	Manatunga	Amita	Biostatistics
	Pearson	Melanie	Envir & occup Health
	Spencer	Jessica	REI
	Tsai	Han-Hsuan	Public Health

Appendix B: Supplementary Figures

Figure S1. Part 1 Classification Method 1: Box-and-whisker plots of the 37 FDR significant features (1: low PBB, 2: high PBB)

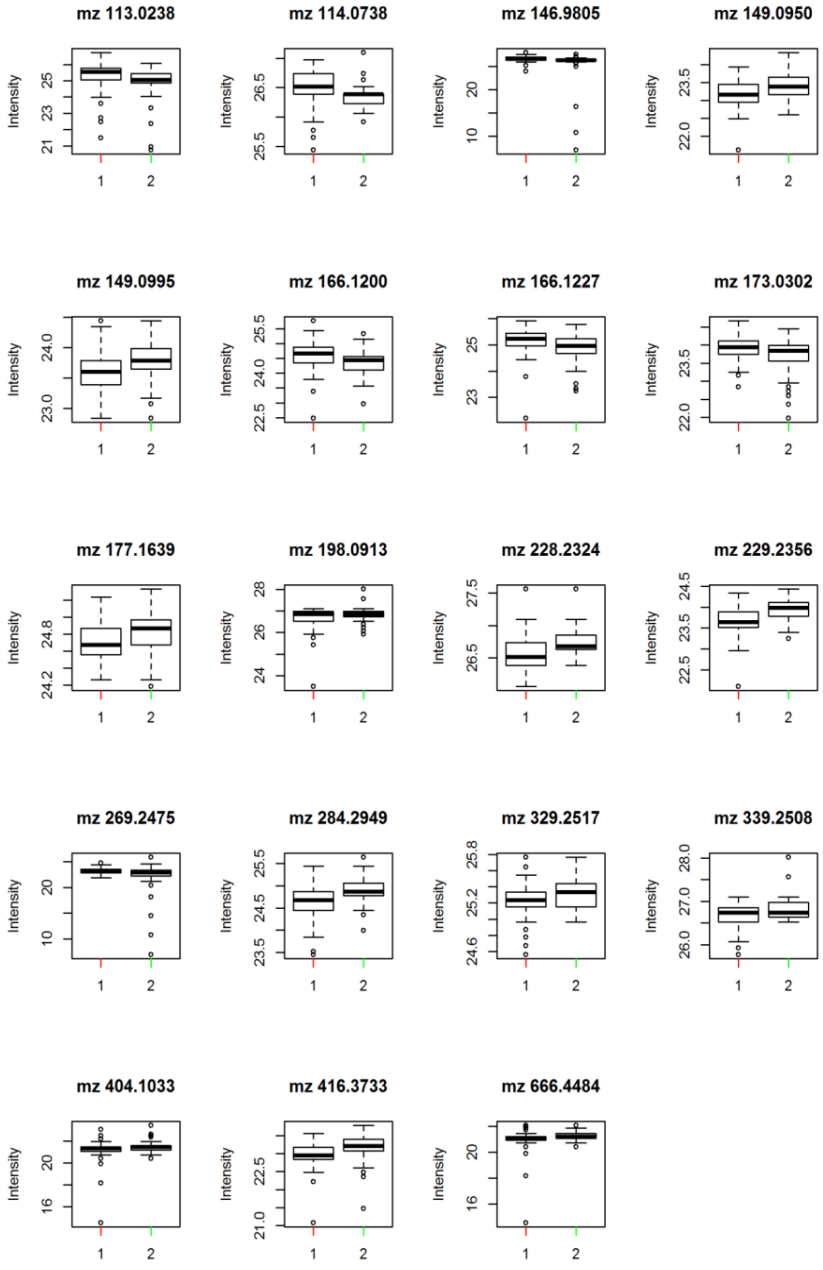


Figure S1. (Continued)

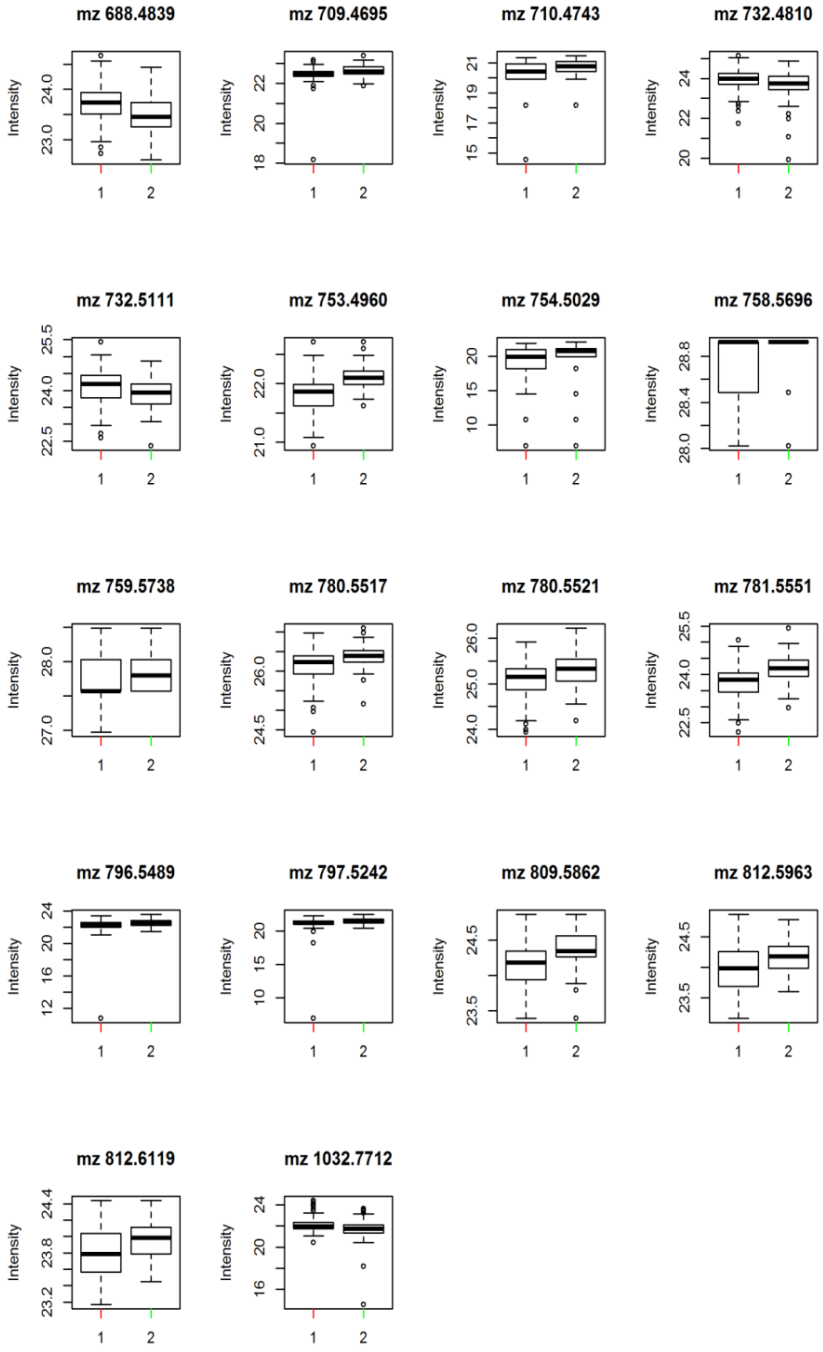


Figure S2. Part 1 Classification Method 2: Box-and-whisker plots of the 39 FDR significant features (1: low PBB, 2: high PBB)

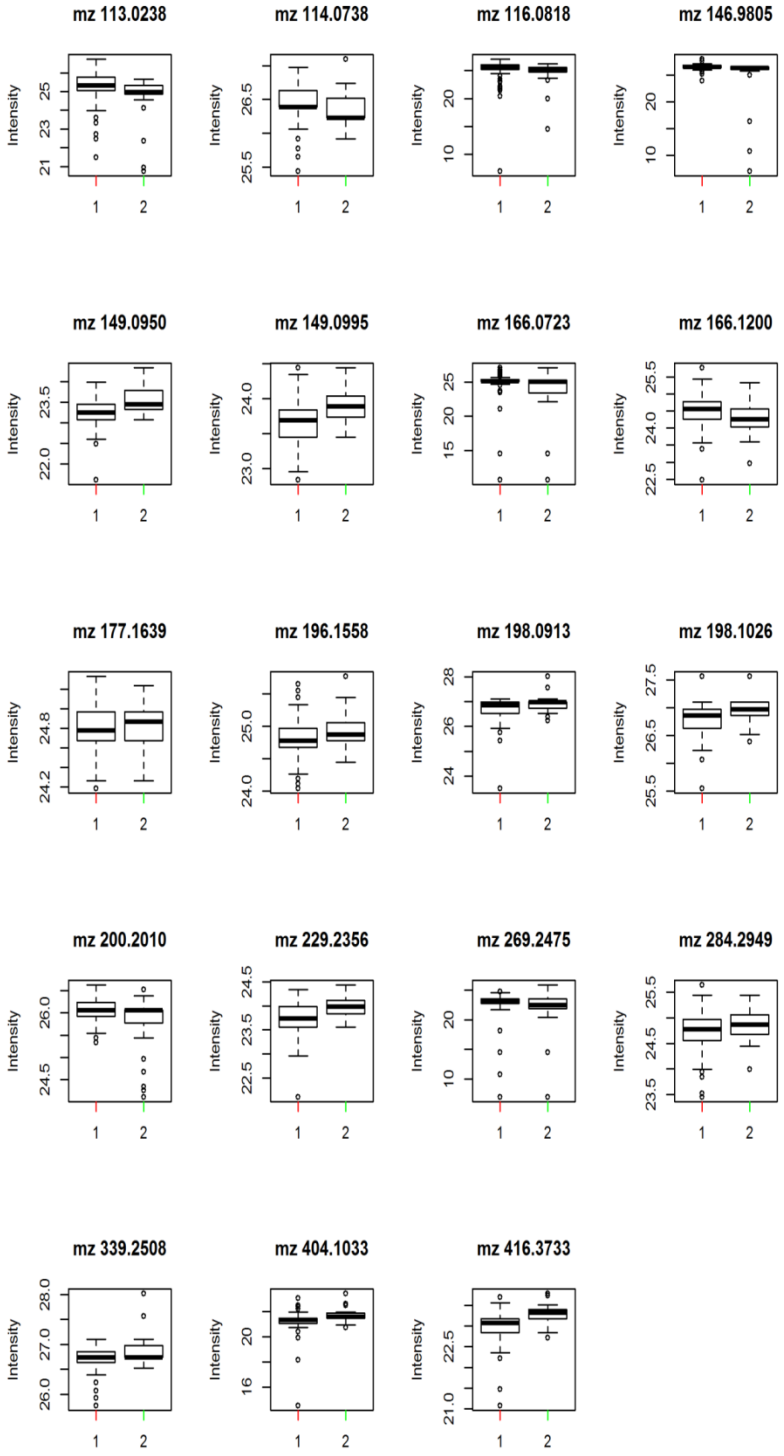


Figure S2. (Continued)

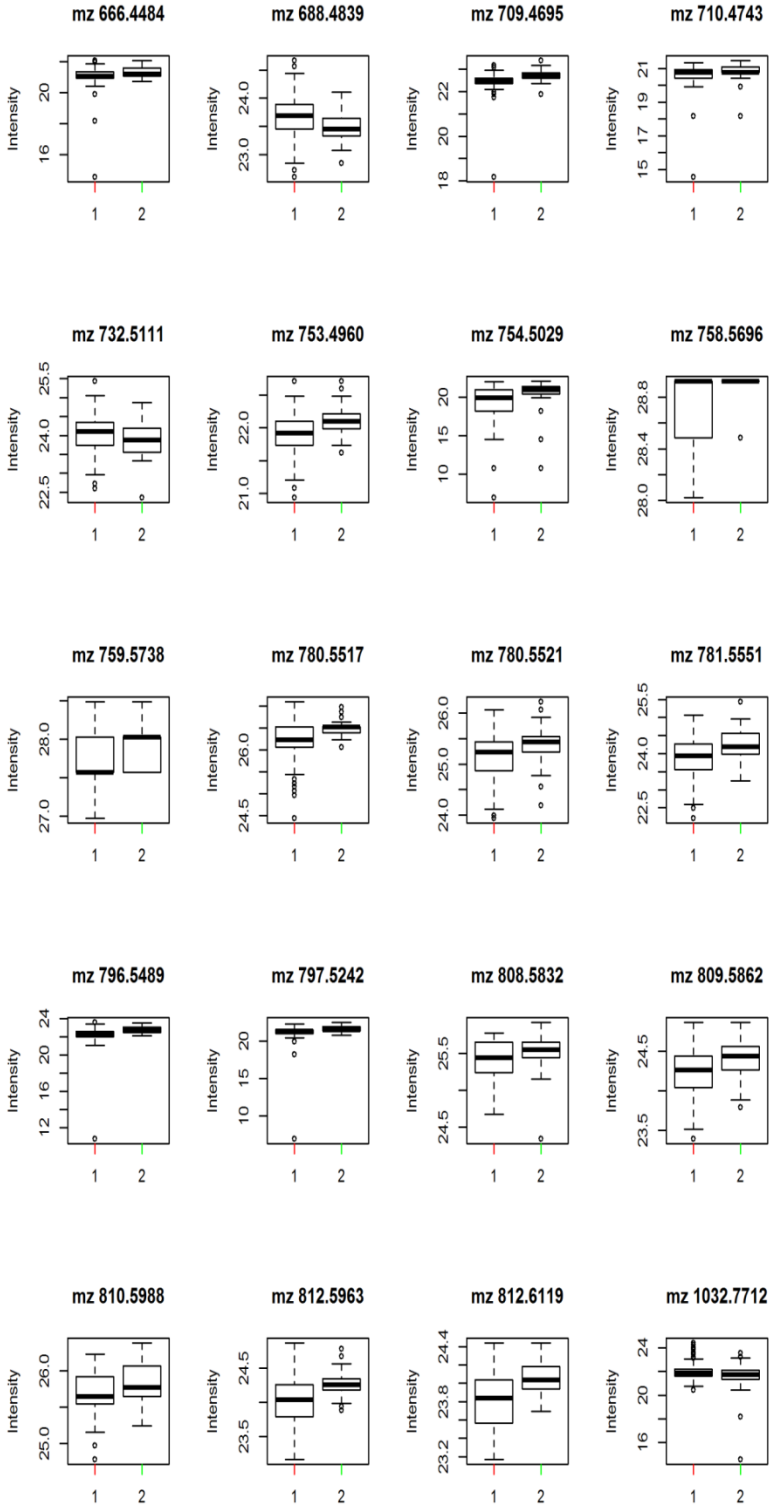
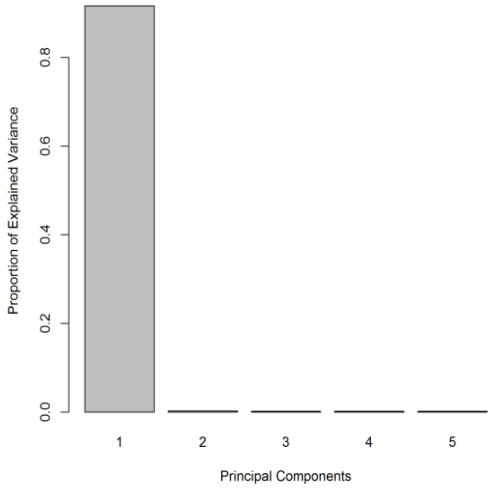
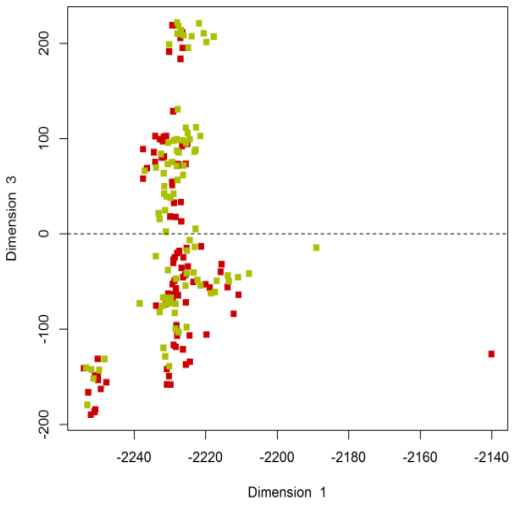


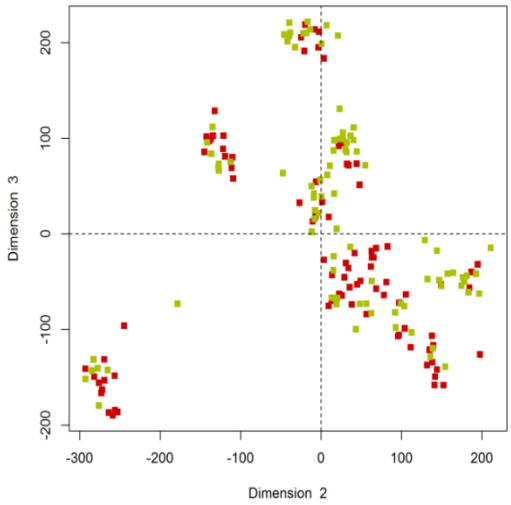
Figure S3. Part 2 Classification Method 1: Global PCA of the 11,202 features of the metabolome (red: low PBB, green: high PBB)



a) Scree plot

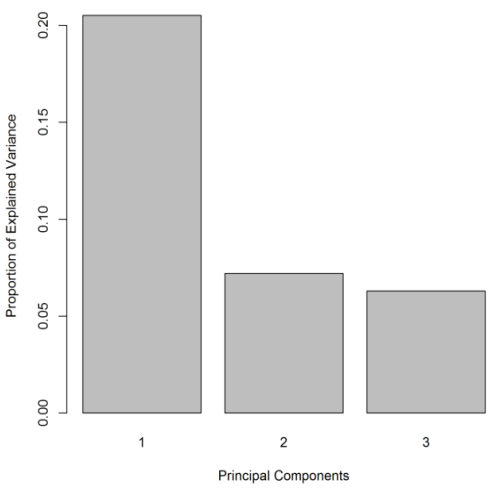


b) Score plot: PC1 vs. PC3

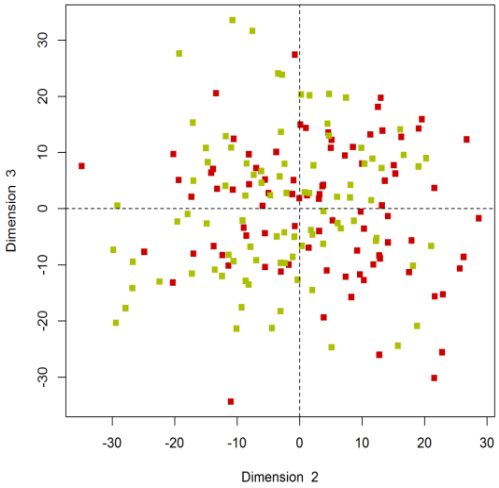


c) Score plot: PC2 vs. PC3

Figure S4. Part 2 Classification Method 1: Focused PCA of the 47 FDR significant features of the metabolome (red: low PBB, green: high PBB)



a) Scree plot



b) Score plot: PC2 vs. PC3

Figure S5. Part 2 Classification Method 1: Box-and-whisker plots of the 47 FDR significant features (1: low PBB, 2: high PBB)

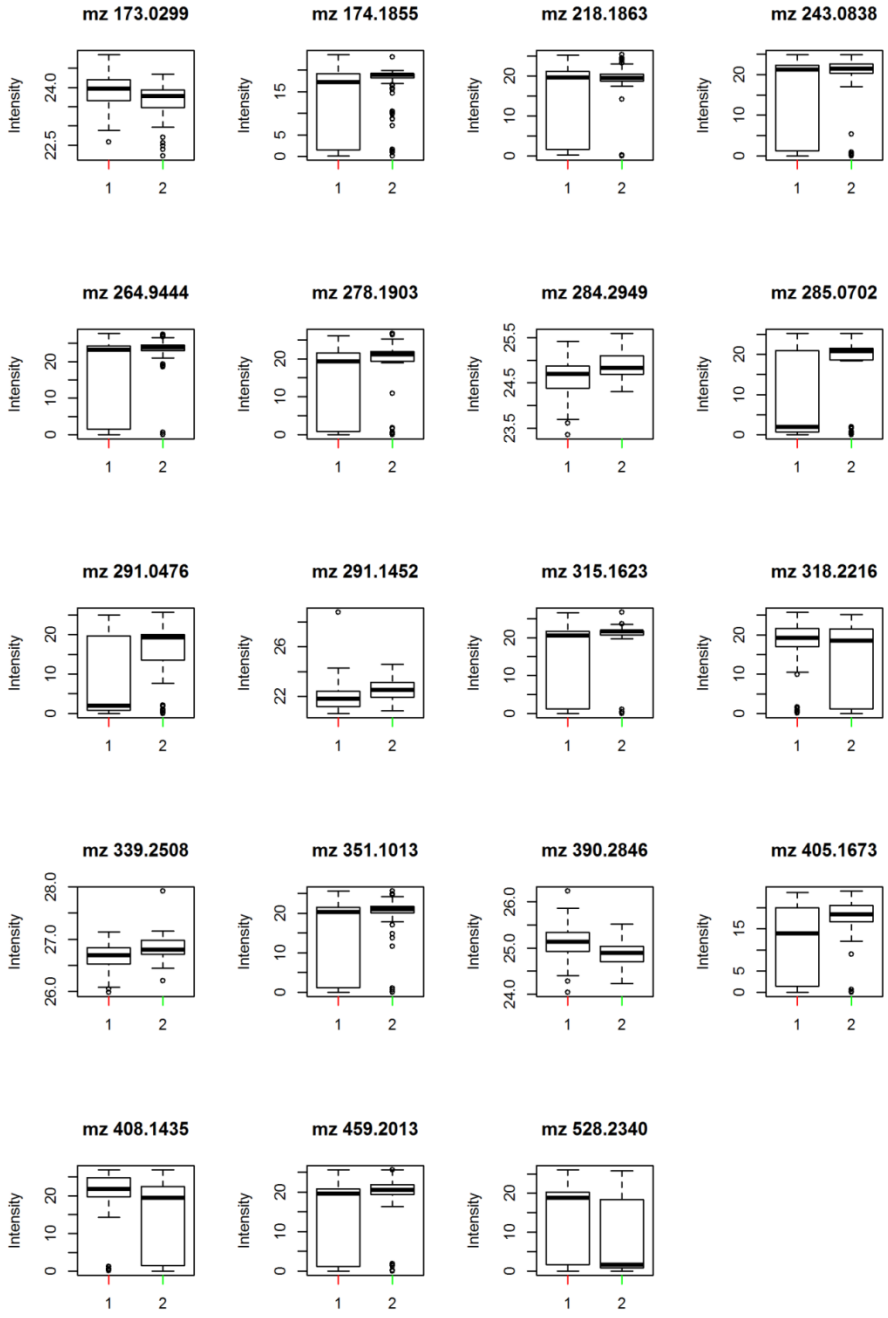


Figure S5. (Continued)

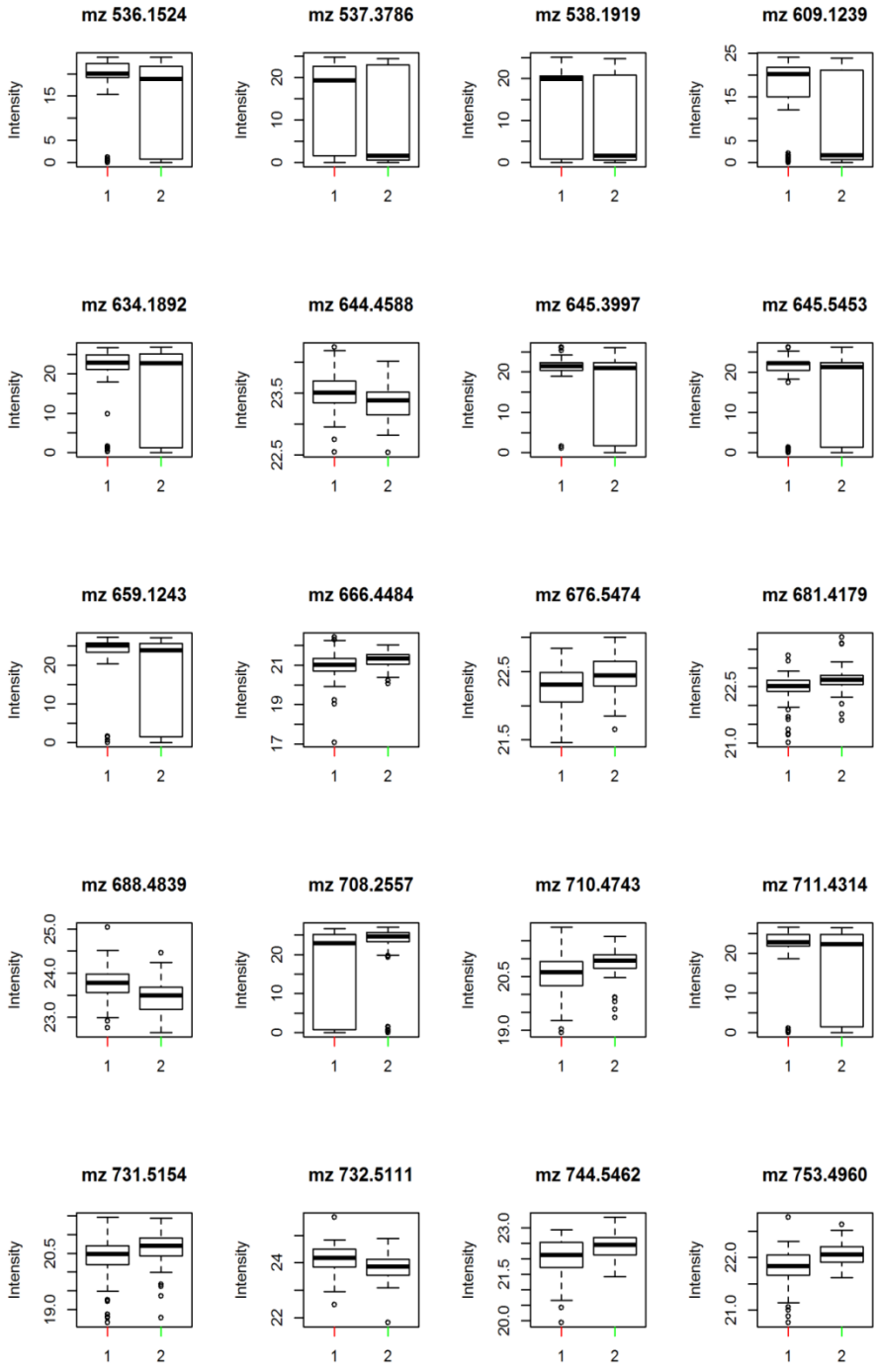


Figure S5. (Continued)

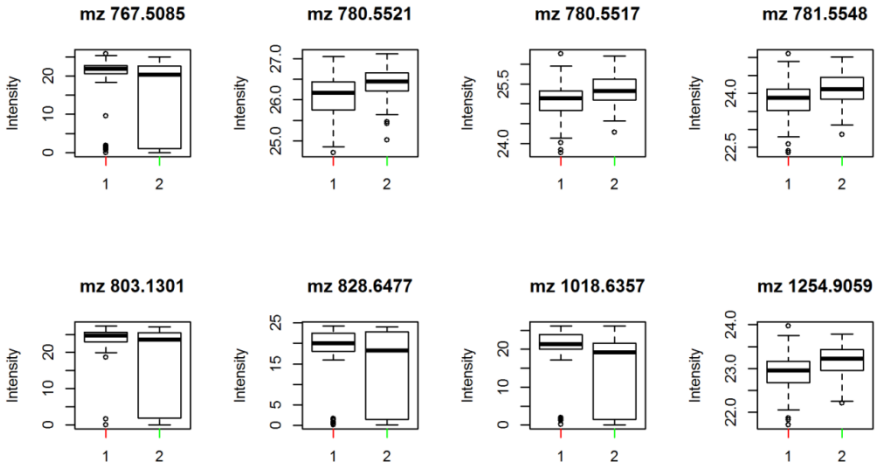
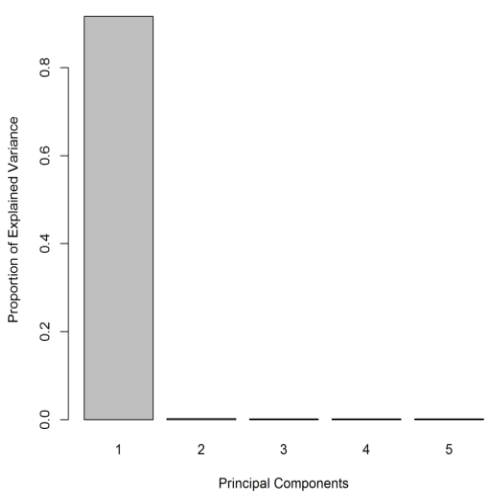
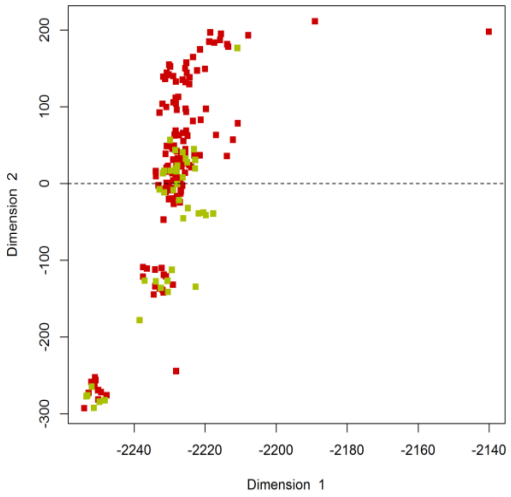


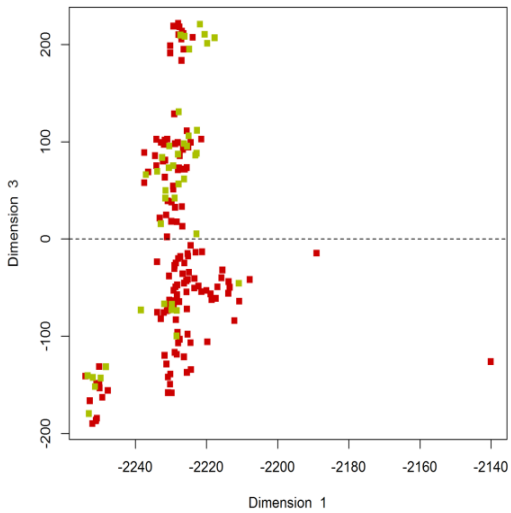
Figure S6. Part 2 Classification Method 2: Global PCA of the 11,202 features of the metabolome (red: low PBB, green: high PBB)



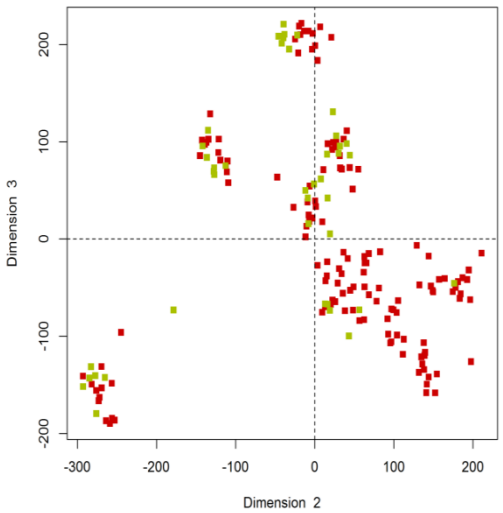
a) Scree plot



b) Score plot: PC1 vs. PC3

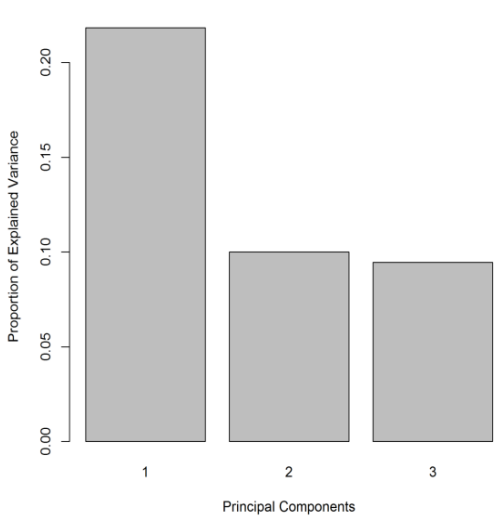


c) Score plot: PC1 vs. PC3

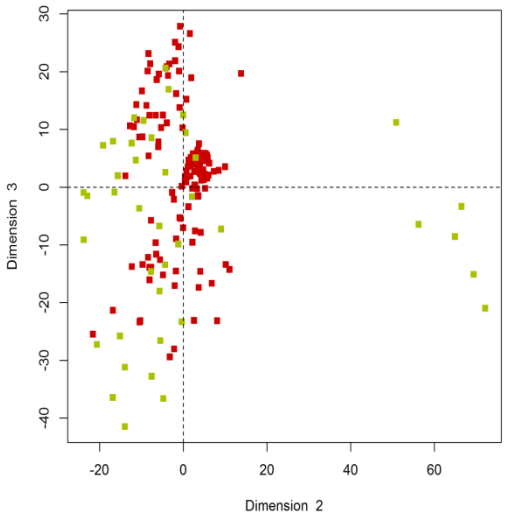


d) Score plot: PC2 vs. PC3

Figure S7. Part 2 Classification Method 2: Focused PCA of the 44 FDR significant features of the metabolome (red: low PBB, green: high PBB)



a) Scree plot



b) Score plot: PC2 vs. PC3

Figure S8. Part 2 Classification Method 2: Box-and-whisker plots of the 44 FDR significant features (1: low PBB, 2: high PBB)

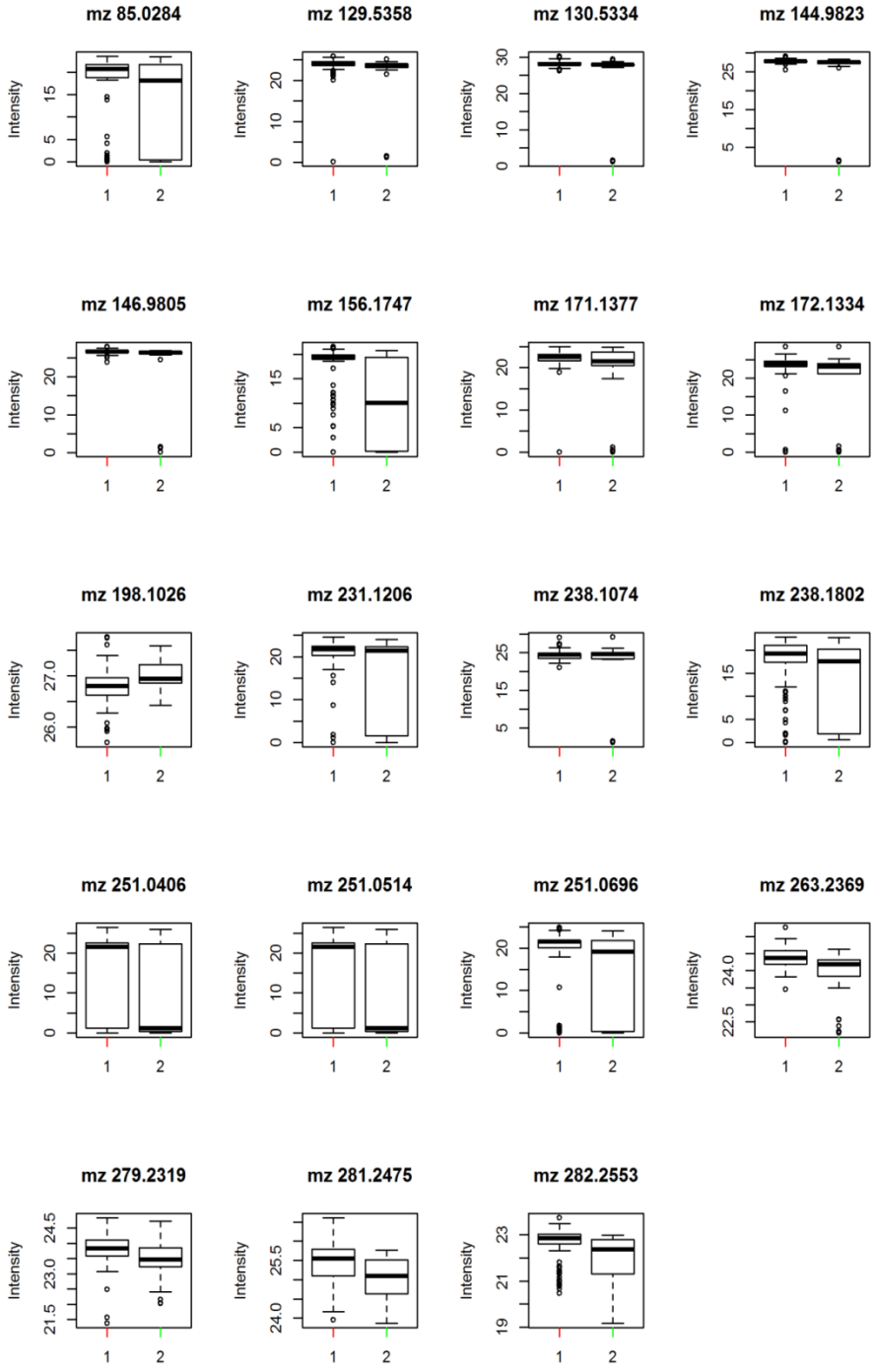


Figure S8. (Continued)

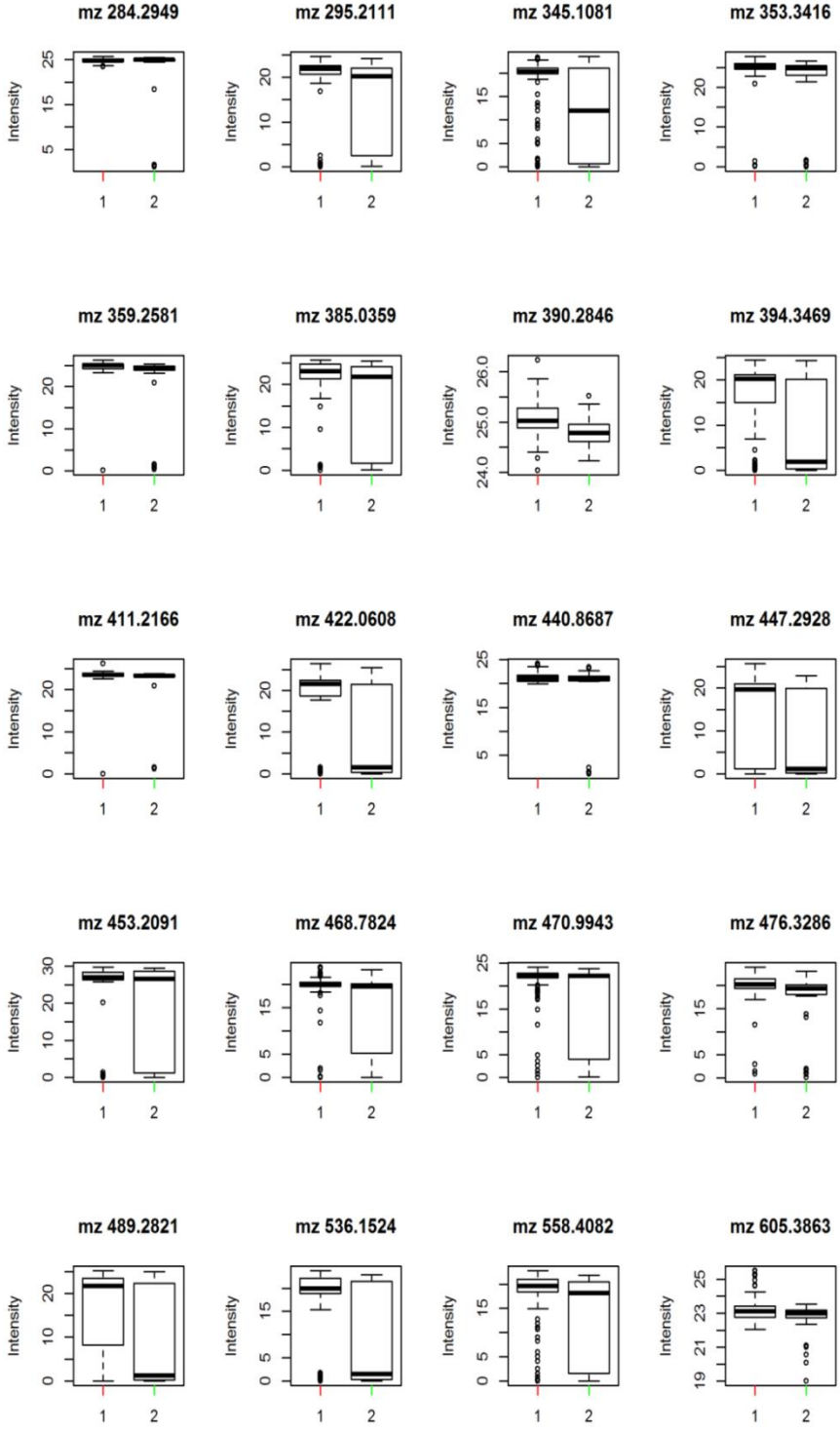


Figure S8. (Continued)

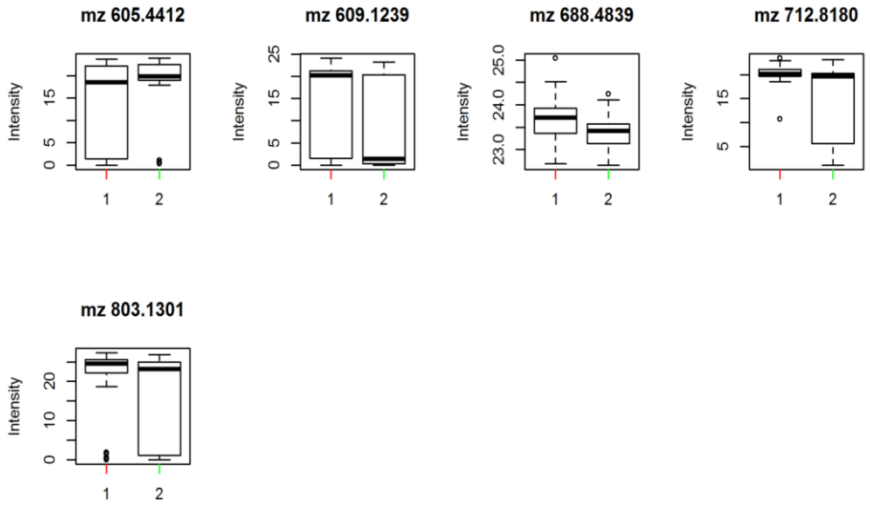
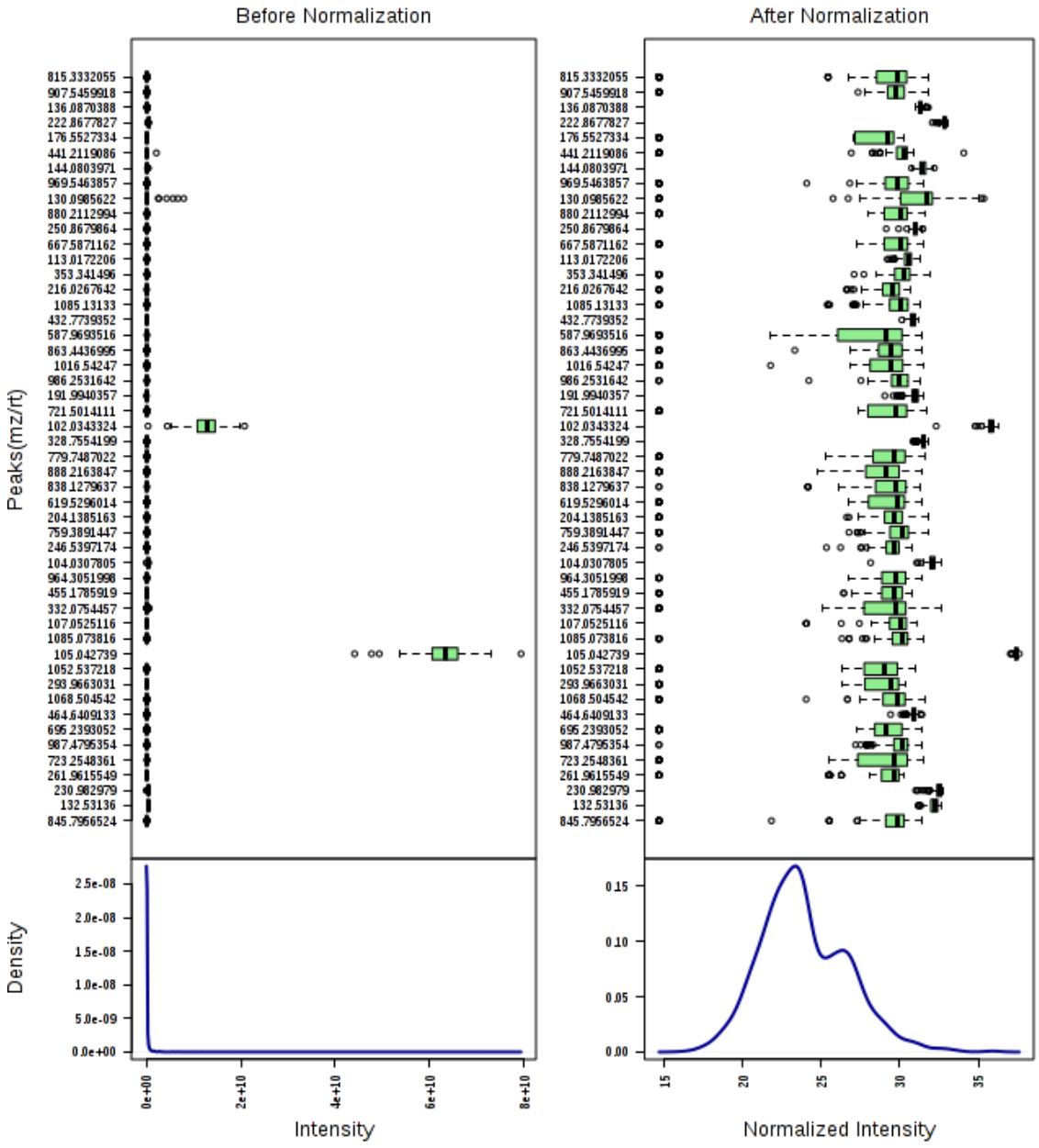
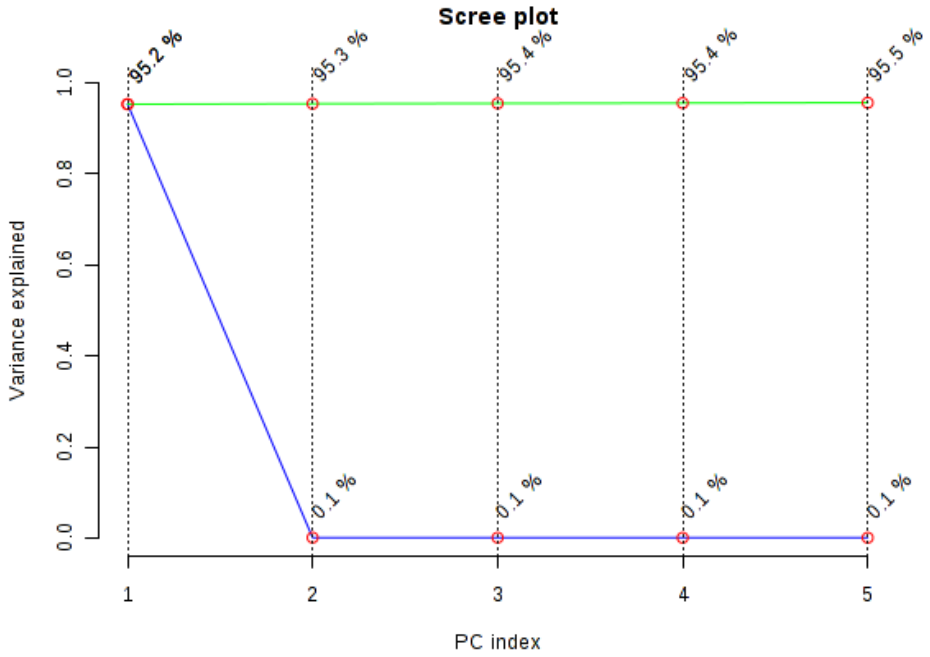


Figure S9. Part 2 Classification Method 3: Data normalization [before (left) and after (right) normalization]*



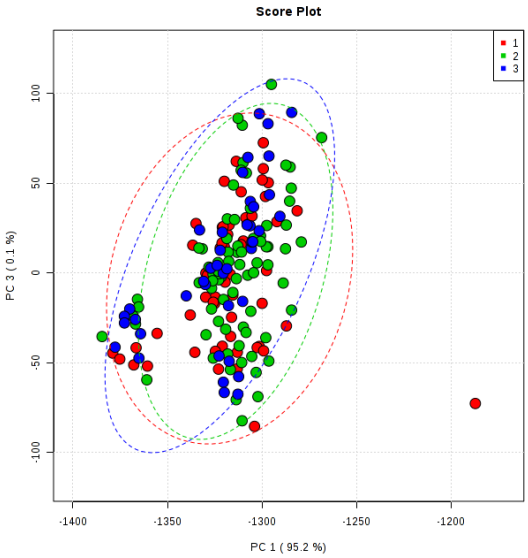
* The density plots are based on all 2,999 features, but the box plots are only represented for 50 features due to the space limit. It can be seen that features follow a more bell-shaped curve after normalization.

Figure S10. Part 2 Classification Method 3: PCA of the 2,999 features included in the analysis (red: low PBB, green: medium PBB, blue: high PBB)

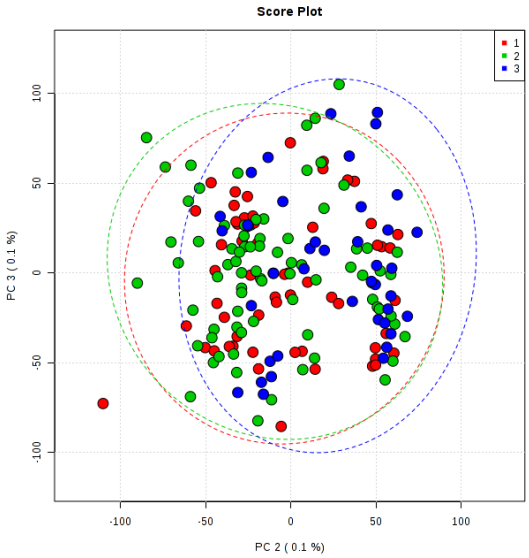


The green line on the top shows the accumulated percentage of variance explained with the addition of each PC, and the blue line beneath shows the variance explained by each individual PC.

a) Scree plot



b) Score plot: PC1 vs. PC3



c) Score plot: PC2 vs. PC3

Figure S11. Part 2 Classification Method 3: PLS-DA score plot (PC2 vs. PC3) of the 2,999 features included in the analysis (red: low PBB, green: medium PBB, blue: high PBB)

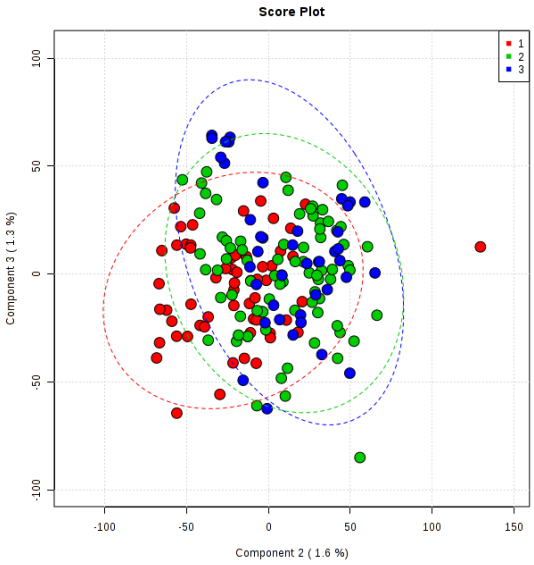


Figure S12. Classification Method 3: PLS-DA model validation by permutation tests based on separation distance (model based on 2,999 features initially included in the analysis)

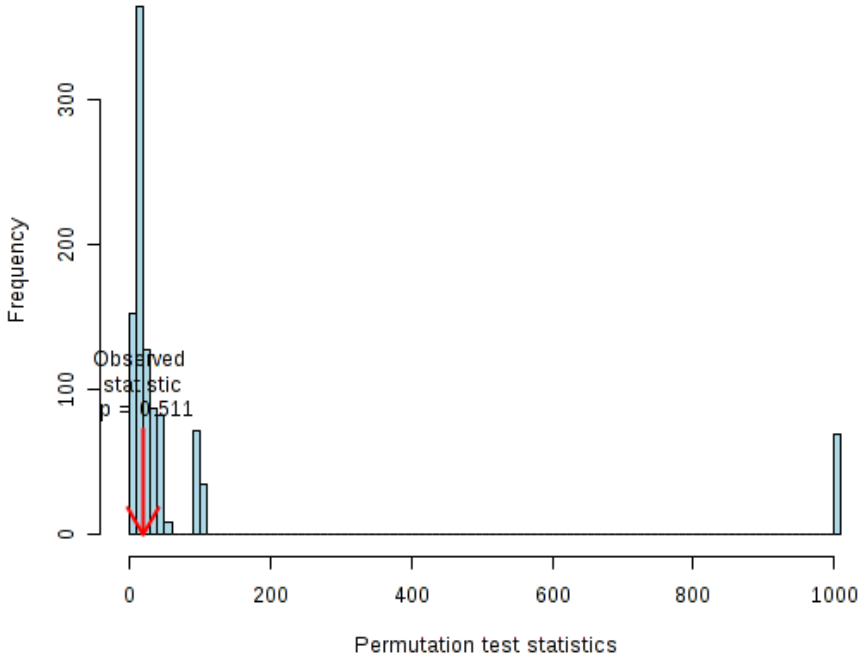
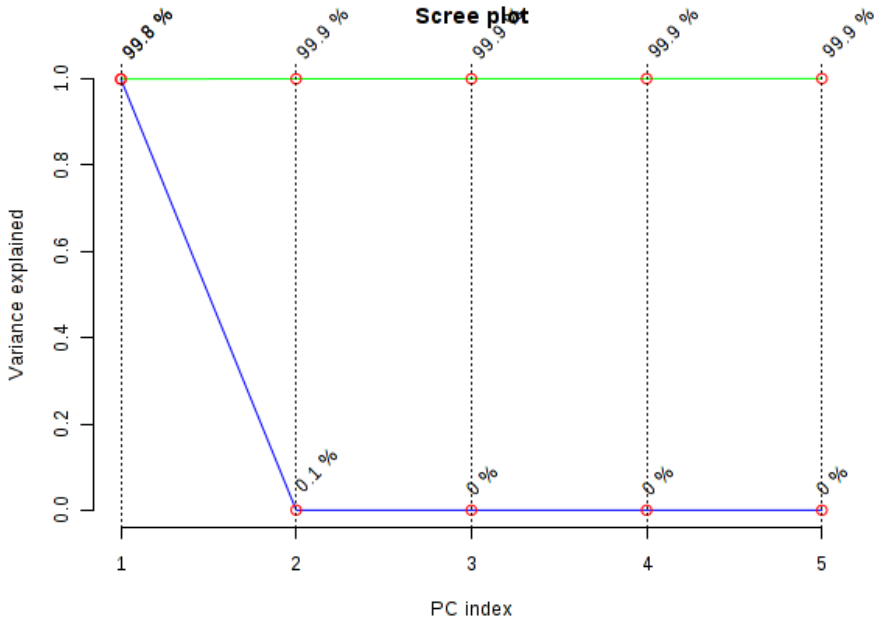
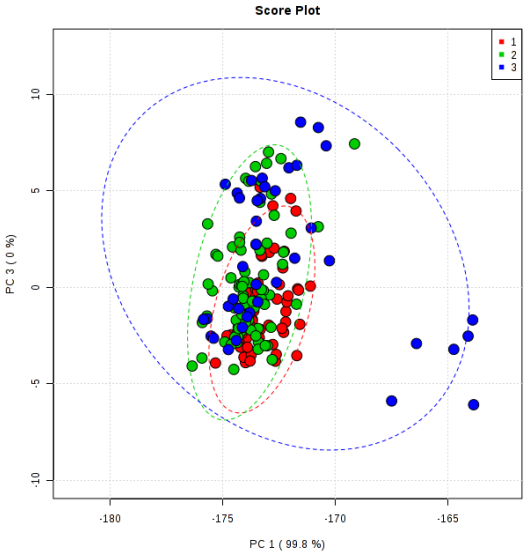


Figure S13. Part 2 Classification Method 3: PCA of the 43 selected features (red: low PBB, green: medium PBB, blue: high PBB)

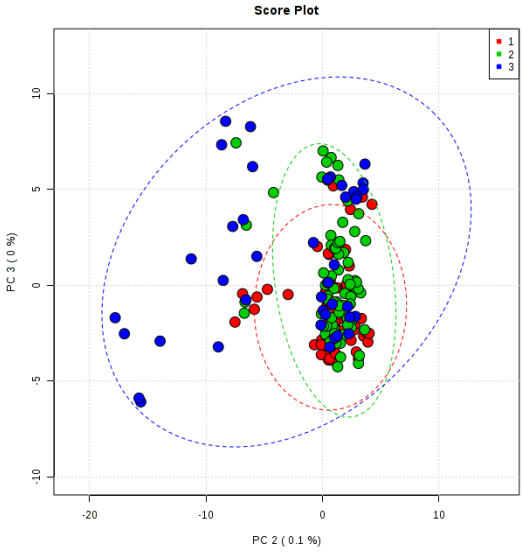


The green line on the top shows the accumulated percentage of variance explained with the addition of each PC, and the blue line beneath shows the variance explained by each individual PC.

a) Scree plot



b) Score plot: PC1 vs. PC3



c) Score plot: PC2 vs. PC3

Figure S14. Part 2 Classification Method 3: PLS-DA score plot (PC2 vs. PC3) of the 43 selected features (red: low PBB, green: medium PBB, blue: high PBB)

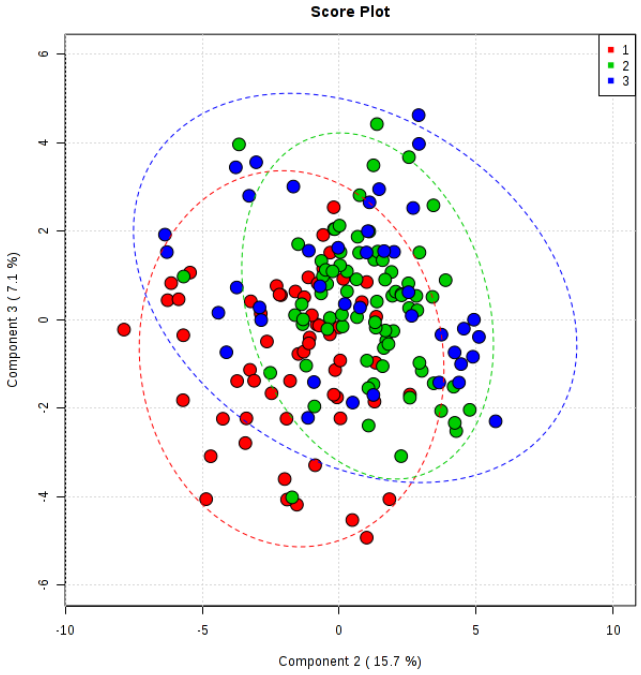
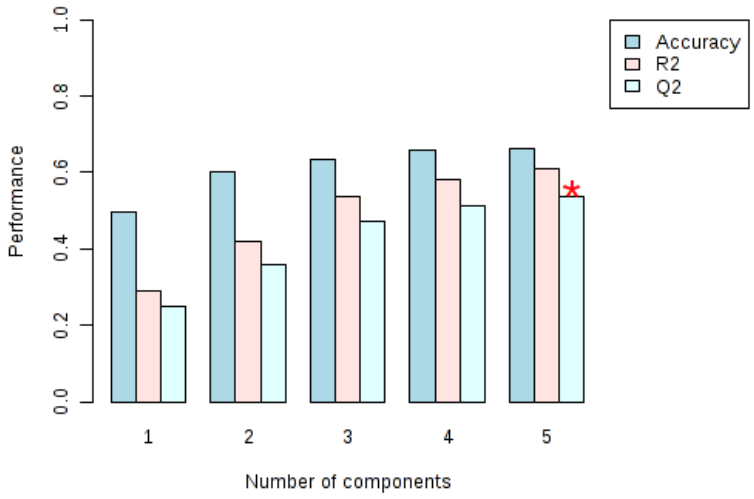
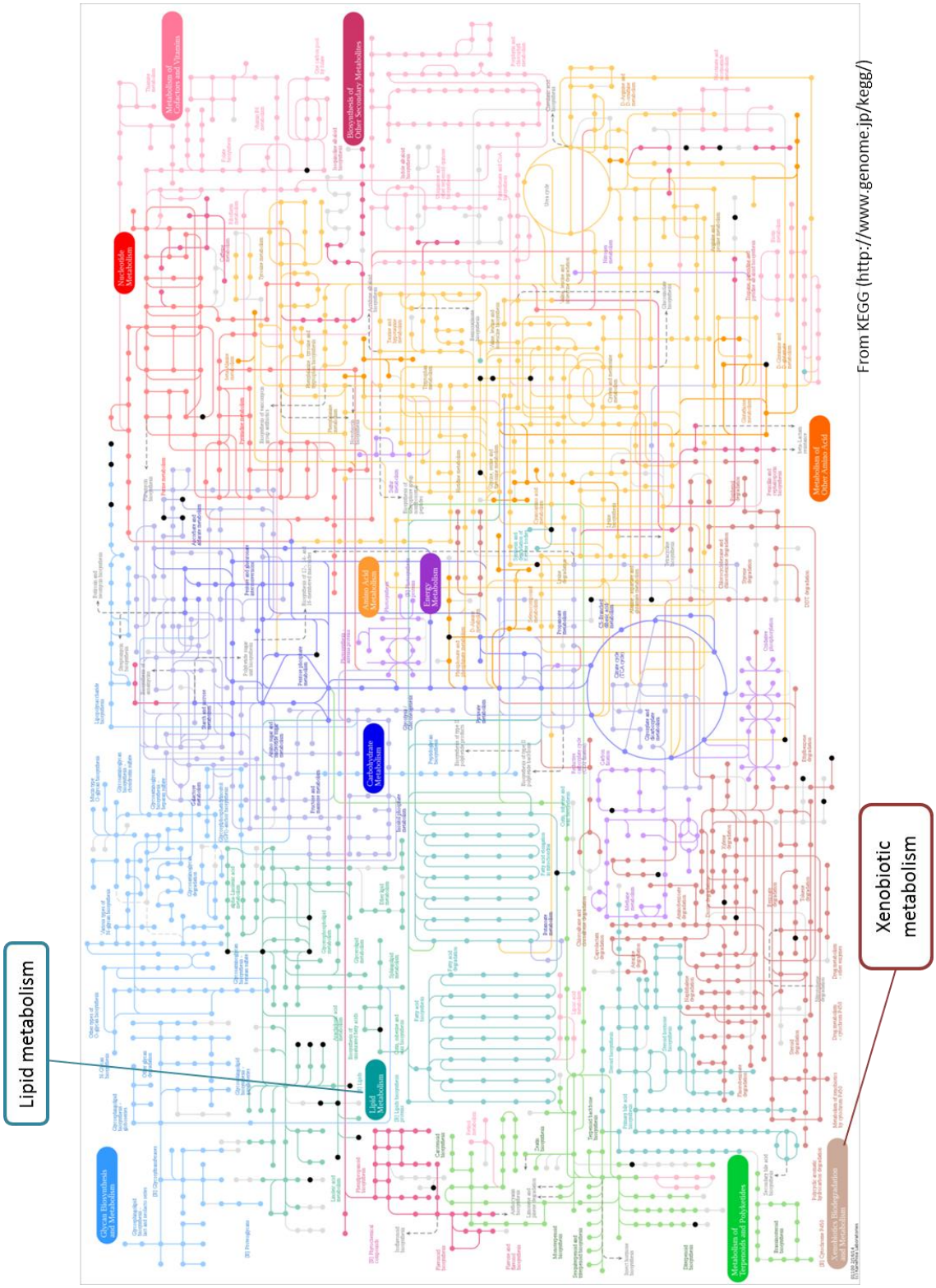


Figure S15. Part 2 Classification Method 3: Determination of the optimal number of components used in the PLS-DA model*



* The optimal number of components is assessed based on three performance measures: the sum of squares captured by the model (R^2), the cross-validated R^2 (i.e., Q^2), and the prediction accuracy. The default is Q^2. The red star indicates the optimal number of components.

Figure S16. Overview of the 162 discriminatory metabolites selected from Part 1 and 2 collectively (matched metabolites are indicated as black circles)



Appendix C. Supplementary Tables

Table S1. Part1. Discriminatory features mutually selected by Classification Method 1 and 2 with matched compound names in MSEA

Input mz*	dppm*	KEGG ID	MSEA match*
113.024	4	C01546	2-Furoic acid
	5	C00815	Citramalic acid
	5	C02614	(S)-2-Methylmalate
	5	C01087	D-2-Hydroxyglutaric acid
	5	C02266	D-Xylono-1,5-lactone
	5	C03826	2-Dehydro-3-deoxy-D-xylonate
	5	C00684	2-Dehydro-3-deoxy-L-arabinonate
	5	C01114	L-Arabinono-1,4-lactone
	5	C02630	2-Hydroxyglutarate
	5	C06032	D-erythro-3-Methylmalate
	5	C02612	(R)-2-Methylmalate
	5	C03196	L-2-Hydroxyglutaric acid
	0	C00490	Itaconic acid
	0	C00433	2,5-Dioxopentanoate
	0	C02800	Gamma-delta-Dioxovaleric acid
	0	C01732	Mesaconic acid
	0	C02214	Glutaconic acid
	0	C02226	Citraconic acid
146.980	8	C03167	Phosphonoacetaldehyde
149.095	7	C06577	Cuminaldehyde
	7	C10452	1-Methoxy-4-(2-propenyl)benzene
	8	C07287	Sulcatone
	7	C14147	Cyromazine
177.164	2	C03527	4-(2,6,6-Trimethyl-1-cyclohexen-1-yl)-2-butanone
198.091	4	C17249	7-Methylthioheptanaldoxime
198.091	1	C06551	N-Ethyl trans-2-cis-6-nonadienamide
269.248	0	C16536	(Z)-9-Heptadecenoic acid
	0	C12100	Cyclohexaneundecanoic acid
284.295	0	C13846	Octadecanamide
	1	C00836	Sphinganine

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S1. (Continued)

Input mz*	dppm*	KEGG ID	MSEA match*
339.251	6	C14772	5,6-DHET
	6	C14773	8,9-DiHETrE
	6	C14774	11,12-DiHETrE
	6	C14775	14,15-DiHETrE
	0	C15988	(9S,10S)-9,10-dihydroxyoctadecanoate
	8	C06475	Prostaglandin F1a
710.474	1	C00350	PE(O-18:1(1Z)/20:4(5Z,8Z,11Z,14Z))
666.448	2		
758.570	0	C00157	PC(16:0/16:0)
780.552	2		

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S2. Part2 Classification Method 3: ANOVA post hoc test (Tukey's Honestly Significant Difference)

mz	FDR	Tukey's HSD
784.584	4.61E-06	2-1; 3-1
780.552	3.25E-05	2-1; 3-1
339.251	3.25E-05	2-1; 3-1
228.232	1.75E-04	2-1; 3-1
806.567	1.75E-04	2-1; 3-1
758.570	1.99E-04	2-1; 3-1
759.574	1.99E-04	2-1; 3-1
146.980	3.04E-04	3-1; 3-2
782.570	3.17E-04	2-1; 3-1
783.574	3.17E-04	2-1; 3-1
198.103	3.17E-04	3-1; 3-2
759.573	5.73E-04	2-1; 3-1
200.201	6.75E-04	2-1; 3-2
144.982	1.15E-03	3-1; 3-2
784.584	1.47E-03	2-1; 3-1
760.583	1.47E-03	2-1; 3-1
158.154	2.44E-03	2-1; 3-1
806.569	2.49E-03	2-1; 3-1

* input detected *m/z* value (mz), FDR adjusted p-value (FDR), significantly different pair-wise comparisons (1: low PBB, 2: medium PBB, 3: high PBB) (Tukey's HSD)

Table S2. (Continued)

mz	FDR	Tukey's HSD
782.570	2.56E-03	2-1; 3-1
198.091	3.75E-03	3-1; 3-2
783.573	3.86E-03	2-1; 3-1
113.024	4.15E-03	2-1; 3-1
758.569	4.48E-03	2-1; 3-1
453.209	8.21E-03	3-1; 3-2
130.533	9.81E-03	3-1; 3-2
281.247	9.88E-03	3-1; 3-2
803.130	1.14E-02	3-1; 3-2
1189.064	1.14E-02	2-1; 3-1
411.770	1.24E-02	2-1; 3-1
808.583	1.43E-02	2-1; 3-1
359.258	1.49E-02	3-1; 3-2
190.982	1.91E-02	2-1; 3-1
353.341	1.91E-02	3-1; 3-2
123.079	2.24E-02	2-1; 3-2
169.134	2.47E-02	2-1; 3-2
352.898	2.67E-02	3-1; 3-2
142.123	2.72E-02	2-1; 3-1
786.601	2.75E-02	2-1; 3-1
238.107	2.94E-02	3-1; 3-2
154.123	3.22E-02	2-1; 3-1
106.035	3.71E-02	3-1; 3-2
780.590	3.96E-02	2-1; 3-1
679.639	4.90E-02	2-1

* input detected *m/z* value (mz), FDR adjusted p-value (FDR), significantly different pair-wise comparisons (1: low PBB, 2: medium PBB, 3: high PBB) (Tukey's HSD)

Table S3. Part2. Discriminatory features mutually selected by at least two of Classification Method 1, 2, and 3 with matched compound names in MSEA

Input mz*	dppm*	KEGG ID	MSEA match*
146.980	8	C03167	Phosphonoacetaldehyde
198.103	0	C14416	3-Amino-1-methyl-5H-pyrido[4,3-b]indole
339.251	6	C14772	5,6-DHET
	6	C14773	8,9-DiHETrE
	6	C14774	11,12-DiHETrE
	6	C14775	14,15-DiHETrE
	0	C15988	(9S,10S)-9,10-dihydroxyoctadecanoate
	8	C06475	Prostaglandin F1a
780.552	2	C00157	PC(16:0/16:0)
	0		
	0	C00350	PE(O-18:1(1Z)/20:4(5Z,8Z,11Z,14Z))

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S4. Part2. All discriminatory features selected by Classification Method 1, 2, and 3 with matched compound names in MSEA

Input mz*	dppm*	KEGG ID	MSEA match*
405.167	5	C07567	Sertindole
	0	C05950	20-Carboxy-leukotriene B4
	0	C05961	6-Keto-prostaglandin F1a
284.295	0	C13846	Octadecanamide
	1	C00836	Sphinganine
291.048	0	C07836	D-glycero-D-manno-Heptose 7-phosphate
	0	C07838	D-glycero-D-manno-Heptose 1-phosphate
	0	C06222	Sedoheptulose 1-phosphate
	0	C05382	Sedoheptulose 7-phosphate
	4	C01736	Nebularine
	4	C05512	Deoxyinosine

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S4. (Continued)

Input mz*	dppm*	KEGG ID	MSEA match*
666.448	2	C00350	PE(O-18:1(1Z)/20:4(5Z,8Z,11Z,14Z))
710.474	1		
758.569	0		
758.570	0		
760.583	2		
780.552	0		
780.590	0, 3		
782.570	3		
782.570	4		
784.584	1		
784.584	1		
828.648	0		
339.251	6	C14772	5,6-DHET
	6	C14773	8,9-DiHETrE
	6	C14774	11,12-DiHETrE
	6	C14775	14,15-DiHETrE
339.251	0	C15988	(9S,10S)-9,10-dihydroxyoctadecanoate
281.248	3		
281.247	4		
339.251	8	C06475	Prostaglandin F1a
758.569	0	C00157	PC(16:0/16:0)
758.570	0		
760.583	2		
780.552	0, 2		
782.570	0, 1, 3, 4		
784.584	1		
786.601	0		
806.567	0, 2		
806.569	0, 2		
808.583	0, 2		
315.162	0	C06918	Clomipramine
	8	C02035	Gibberellin A20
	8	C11864	Gibberellin A4
	8	C11865	Gibberellin A51
318.222	0	C08067	Butenafine
	3	C08012	Levomethadyl Acetate

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S4. (Continued)

Input mz*	dppm*	KEGG ID	MSEA match*
351.101	0	C08126	Penicillin V
	0	C16517	Indolylmethyl-desulfoglucosinolate
	5	C10283	2',3,4',5-Tetrahydroxy-4-prenylstilbene
	7	C07322	Olanzapine
408.144	2	C00415	Dihydrofolic acid
	0	C11154	Trimetrexate
459.201	5	C07007	Fluocinonide
	1	C11132	2-Methoxyestrone 3-glucuronide
681.418	7	C08596	Fucoxanthin
291.145	0	C10896	Benomyl
	0	C01965	Trimethoprim
278.190	0	C06824	Amitriptyline
	0	C07107	Maprotiline
	0	C16659	2-Ethylidene-1,5-dimethyl-3,3-diphenylpyrrolidine
708.256	0	C06371	Lacto-N-tetraose
263.237	9	C00517	Palmitaldehyde
	9	C14499	Hexadecane
	4	C08365	Ricinoleic acid
281.248	1	C08365	Ricinoleic acid
281.247	2		
263.237	2	C04056	Bovinic acid
281.248	0		
281.247	0		
263.237	2	C01595	Linoleic acid
281.248	0		
281.247	0		
146.980	8	C03167	Phosphonoacetaldehyde
345.108	2	C04079	D-Pantothenoyl-L-cysteine
	4	C10514	Phaseollin
	5	C09499	Matricin
422.061	0	C08409	Glucoerucin
489.282	9	C10009	Mucronine D

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S4. (Continued)

Input mz*	dppm*	KEGG ID	MSEA match*
279.232	0	C08364	Punicic acid
	0	C06427	Alpha-Linolenic acid
	0	C07289	Crepenynate
	0	C06426	Gamma-Linolenic acid
	8	C00249	Palmitic acid
	3	C14828	9,10-DHOME
	3	C14829	12,13-DHOME
	1	C14825	9,10-Epoxyoctadecenoic acid
	1	C14767	Alpha-dimorphecolic acid
	1	C14762	13S-hydroxyoctadecadienoic acid
	1	C14826	12,13-EpOME
171.138	1	C07276	Limonene-1,2-diol
	1	C18066	Menthone lactone
	1	C18202	8-Methylnonenoate
	1	C17621	10-Hydroxygeraniol
	1	C16462	Citronellic acid
198.103	0	C14416	3-Amino-1-methyl-5H-pyrido[4,3-b]indole
251.041	1	C03872	L-2-Aminoethyl seryl phosphate
251.051	0	C01691	Diflunisal
	3	C01817	DL-Homocystine
251.070	0	C00669	Gamma-Glutamylcysteine
	7	C14285	Oxybenzone
	7	C09314	Trioxsalen
	6	C09312	Seselin
	6	C09925	5,6-Dihydro-5-hydroxy-6-methyl-2H-pyran-2-one
	6	C03582	Resveratrol
	6	C09833	Sakuranetin
	6	C09282	(R)-Oxypeucedanin
	6	C10274	3,4-Dihydro-8-hydroxy-3-(3-hydroxy-4-methoxyphenyl)-1H-2-benzopyran-1-one
	4	C00858	Formononetin
	4	C12125	Isoformononetin
	1	C19254	2-Amino-3,4-dimethylimidazo[4,5-f]quinoline

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S4. (Continued)

Input mz*	dppm*	KEGG ID	MSEA match*
85.028	0	C06144	3-Butynoate
	0	C17601	4-Hydroxy-2-butenoic acid gamma-lactone
	6	C06002	(S)-Methylmalonic acid semialdehyde
	6	C00232	Succinic acid semialdehyde
	6	C00164	Acetoacetic acid
	6	C00109	2-Ketobutyric acid
	6	C00349	2-Methyl-3-oxopropanoic acid
784.584	2, 3	C02737	PS(16:0/16:0)
198.091	4	C17249	7-Methylthioheptanaldoxime
	1	C06551	N-Ethyl trans-2-cis-6-nonadienamide
113.024	4	C01546	2-Furoic acid
	5	C00815	Citramalic acid
	5	C02614	(S)-2-Methylmalate
	5	C01087	D-2-Hydroxyglutaric acid
	5	C02266	D-Xylono-1,5-lactone
	5	C03826	2-Dehydro-3-deoxy-D-xylonate
	5	C00684	2-Dehydro-3-deoxy-L-arabinonate
	5	C01114	L-Arabinono-1,4-lactone
	5	C02630	2-Hydroxyglutarate
	5	C06032	D-erythro-3-Methylmalate
	5	C02612	(R)-2-Methylmalate
	5	C03196	L-2-Hydroxyglutaric acid
	0	C00490	Itaconic acid
	0	C00433	2,5-Dioxopentanoate
	0	C02800	Gamma-delta-Dioxovaleric acid
	0	C01732	Mesaconic acid
	0	C02214	Glutaconic acid
	0	C02226	Citraconic acid
123.079	8	C05853	2-Phenylethanol
	8	C06757	4-Methylbenzyl alcohol
	8	C07112	1-Phenylethanol
	8	C07213	2-Methylbenzyl alcohol acetate
	8	C07216	(3-Methylphenyl)methyl acetate
	8	C11348	(S)-1-Phenylethanol
	8	C13637	4-Ethylphenol
	8	C14386	3-Ethylphenol

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

Table S4. (Continued)

Input mz*	dppm*	KEGG ID	MSEA match*
169.134	5	C01259	3-Hydroxy-N6,N6,N6-trimethyl-L-lysine
142.123	1	C00729	Tropine
	1	C06179	Hygrine
	1	C06182	(+/-)-Pelletierine
	1	C10864	Physoperuvine
154.123	0	C10865	psi-Pelletierine
679.639	0	C02530	CE(16:1(9Z))

* input detected *m/z* value (Input mz), delta = theoretical *m/z* – detected *m/z* (dppm), compound name matched in MSEA (MSEA match)

References

1. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomarkers Prev.* Aug 2005;14(8):1847-1850.

2. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet.* Dec 15 2012;380(9859):2224-2260.

3. Rappaport SM. Discovering environmental causes of disease. *J. Epidemiol. Community Health.* Feb 2012;66(2):99-102.

4. Pleil JD, Stiegel MA. Evolution of environmental exposure science: using breath-borne biomarkers for "discovery" of the human exposome. *Anal. Chem.* Nov 5 2013;85(21):9984-9990.

5. Rappaport SM. Implications of the exposome for exposure science. *Journal of exposure science & environmental epidemiology.* Jan-Feb 2011;21(1):5-9.

6. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. *Toxicol. Sci.* Jan 2014;137(1):1-2.

7. Athersuch TJ. The role of metabolomics in characterizing the human exposome. *Bioanalysis.* Sep 2012;4(18):2207-2212.

8. Vineis P, van Veldhoven K, Chadeau-Hyam M, Athersuch TJ. Advancing the application of omics-based biomarkers in environmental epidemiology. *Environ. Mol. Mutagen.* Aug 2013;54(7):461-467.

9. Wild CP, Scalbert A, Herceg Z. Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. *Environ. Mol. Mutagen.* Aug 2013;54(7):480-499.

10. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am. J. Public Health.* 2005;95 Suppl 1:S144-150.

11. Vineis P, Khan AE, Vlaanderen J, Vermeulen R. The impact of new research technologies on our understanding of environmental causes of disease: the concept of clinical vulnerability. *Environ. Health.* 2009;8:54.

12. Rappaport SM. Biomarkers intersect with the exposome. *Biomarkers.* Sep 2012;17(6):483-489.

13. Chadeau-Hyam M, Athersuch TJ, Keun HC, et al. Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers.* Feb 2011;16(1):83-88.

14. U.S. Environmental Protection Agency. Toxic Substance Control Act Chemical Substance Inventory. http://www.epa.gov/oppt/existingchemicals/pubs/tscainventory/basic.html. Accessed Jan 10, 2014.

15. *Fourth National Report on Human Exposure to Environmental Chemicals.*: Centers for Disease Control and Prevention;2013.

16. Birnbaum LS, Staskal DF. Brominated flame retardants: cause for concern? *Environ. Health Perspect.* Jan 2004;112(1):9-17.

17. Terrell ML, Berzen AK, Small CM, Cameron LL, Wirth JJ, Marcus M. A cohort study of the association between secondary sex ratio and parental exposure to polybrominated biphenyl (PBB) and polychlorinated biphenyl (PCB). *Environ. Health.* 2009;8:35.

18. *Toxicological Profile for Polybrominated Biphenyls and Polybrominated Diphenyl Ethers.* U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substance and Disease Registry;2004.

19. Landrigan PJ, Wilcox KR, Jr., Silva J, Jr., Humphrey HE, Kauffman C, Heath CW, Jr. Cohort study of Michigan residents exposed to polybrominated biphenyls: epidemiologic and immunologic findings. *Ann. N. Y. Acad. Sci.* May 31 1979;320:284-294.

20. Terrell ML, Manatunga AK, Small CM, et al. A decay model for assessing polybrominated biphenyl exposure among women in the Michigan Long-Term PBB Study. *Journal of exposure science & environmental epidemiology.* Jul 2008;18(4):410-420.

21. Small CM, DeCaro JJ, Terrell ML, et al. Maternal exposure to a brominated flame retardant and genitourinary conditions in male offspring. *Environ. Health Perspect.* Jul 2009;117(7):1175-1179.

22. Blanck HM, Marcus M, Tolbert PE, et al. Age at menarche and tanner stage in girls exposed in utero and postnatally to polybrominated biphenyl. *Epidemiology.* Nov 2000;11(6):641-647.

23. Rosen DH, Flanders WD, Friede A, Humphrey HE, Sinks TH. Half-life of polybrominated biphenyl in human sera. *Environ. Health Perspect.* Mar 1995;103(3):272-274.

24. Miceli JN, Nolan DC, Marks B, Hariharan M. Persistence of polybrominated biphenyls (PBB) in human post-mortem tissue. *Environ. Health Perspect.* May 1985;60:399-403.

25. Yu T, Park Y, Johnson JM, Jones DP. apLCMS--adaptive processing of high-resolution LC/MS data. *Bioinformatics.* Aug 1 2009;25(15):1930-1936.

26. Park YH, Lee K, Soltow QA, et al. High-performance metabolic profiling of plasma from seven mammalian species for simultaneous environmental chemical surveillance and bioeffect monitoring. *Toxicology.* May 16 2012;295(1-3):47-55.

27. Uppal K. User's Manual: Package 'MetBN'. 2014.

28. Smyth GK. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* 2005:397-420.

29. Xia J, Wishart DS. Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*. Jun 2011;Chapter 14:Unit 14 10.

30. Xia J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.* Jun 2011;6(6):743-760.

31. Shinde DD, Kim KB, Oh KS, et al. LC-MS/MS for the simultaneous analysis of arachidonic acid and 32 related metabolites in human plasma: Basal plasma concentrations and aspirin-induced changes of eicosanoids. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* Dec 12 2012;911:113-121.

32. Human Metabolome Database: Arachidonic acid (HMDB01043). http://www.hmdb.ca/metabolites/HMDB01043. Accessed Apr 15, 2014.

33. Sacerdoti D, Gatta A, McGiff JC. Role of cytochrome P450-dependent arachidonic acid metabolites in liver physiology and pathophysiology. *Prostaglandins Other Lipid Mediat.* Oct 2003;72(1-2):51-71.

34. Hakk H, Letcher RJ. Metabolism in the toxicokinetics and fate of brominated flame retardants--a review. *Environ. Int.* Sep 2003;29(6):801-828.

35. Berg JM, Tymoczko JL, Stryer L. *Biochemistry*. New York: W. H. Freeman and Company; 2007.

36. Osborn MP, Park Y, Parks MB, et al. Metabolome-wide association study of neovascular age-related macular degeneration. *PLoS One.* 2013;8(8):e72737.

37. Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: progress in addressing complexity in diet and health. *Annu. Rev. Nutr.* Aug 21 2012;32:183-202.