UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Statistical Methods for Uncovering the Role of the Exposome in Complex Disease Risk

Permalink

https://escholarship.org/uc/item/8sk2c0wv

Author

Imani, Partow

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Statistical Methods for Uncovering the Role of the Exposome in Complex Disease Risk

by

Partow Imani

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sandrine Dudoit, Chair
Professor Haiyan Huang
Professor Stephen Rappaport

Spring 2024

Statistical Methods for Uncovering the Role of the Exposome
in Complex Disease Risk

Copyright 2024
by
Partow Imani

Abstract

Statistical Methods for Uncovering the Role of the Exposome
in Complex Disease Risk

by

Partow Imani

Doctor of Philosophy in Biostatistics

and the Designated Emphasis in

Computational and Genomic Biology

University of California, Berkeley

Professor Sandrine Dudoit, Chair

It has become increasingly apparent that the set of exposures a person experiences can have profound ramifications for increasing their risks for various diseases. While genetics explains some portion of this risk, more complex diseases require understanding the interaction of genetics with this set of dynamic exposures, also referred to as the exposome. To address this, researchers have turned to various technologies focused on detecting indicators for these exposures. Due to the nature of the technologies used as well as the dynamic and continuous nature of the features measured, these exposomic datasets often suffer from a large amount of technical noise which must first be removed to allow for downstream detection of the biological signal of interest. In addition, it is important to consider the characteristics of these datasets and how they were generated when performing the analysis to answer the biological, and resulting statistical, questions at hand. In this dissertation we present analyses of three different types of exposomic datasets and associated questions of interest, in each case first discussing the necessary pre-processing. The first chapter applies a prediction framework to assess whether folate pathway disregulation is involved in increasing risk for pediatric ALL using a targeted metabolomics dataset. The second includes both hypothesis generation and a limited hypothesis testing analysis for transformed untargeted metabolomic features to reactomic features for paired mother-child data in a pediatric ALL case-control study. For hypothesis generation, marginal and joint analyses of the two datasets are performed using an ensemble feature selection method. Chapter three focuses on the previously mentioned ensemble feature selection method, using an untargeted adductomics lung cancer case-control dataset. Options for automating the method are proposed and evaluated using a

simulation for bench-marking compared to standard hypothesis testing practices.

To Kian

Who kept me fed and relatively sane, and whose love and support made this possible.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **xi**

1 Introduction **1**

2 Targeted Metabolomics: Examining Folate Pathway Disregulation in Pediatric Acute Lymphoblastic Leukemia **4**
2.1 Introduction 4
2.2 Background 6
2.3 Materials and Methods 8
2.4 Results 17
2.5 Conclusions 20
2.6 Discussion 23
2.7 Notation 26
2.8 Abbreviations 27

3 Untargeted Reactomics: Feature Selection for Paired Mother-Child Data in Childhood Leukemia Case-Control Study **29**
3.1 Introduction 29
3.2 Materials and Methods 31
3.3 Results 42
3.4 Discussion 47

4 Untargeted Adductomics: Feature Selection in Lung Cancer Case-control Study and Simulation **53**
4.1 Introduction 53
4.2 Lung Cancer Analysis 54
4.3 Simulation and Automated Analysis 69
4.4 Discussion 87

5 Conclusion and Future Directions **90**

Bibliography **92**

A **107**
A.1 Supplementary Figures and Tables 107

B **116**
B.1 Supplementary Methods 116
B.2 Supplementary Figures and Tables 116

C **128**
C.1 Supplementary Methods 128
C.2 Supplementary Figures and Tables 130

List of Figures

2.1 One-carbon (folate) metabolism pathway. The data analyzed in our study correspond to a subset of 11 metabolites involved in this pathway [55]. 7
2.2 Data structure for targeted metabolomics dataset. 9
2.3 Flow chart showing the partitioning of the data into learning and test sets, followed by separate pre-processing, model training on the learning set using Monte Carlo cross-validation, and culminating in prediction evaluation on the test set. 10
2.4 a) Density plot of ICC values by batch. Three removed metabolites with ICC $\leq$ 0.2 shown. b) Percent missing values (y-axis) by metabolite (x-axis). Red line shows 30% cutoff. 12
2.5 Pairwise scatterplots showing the first four principal components of the abundance measures for the biological samples. The left column shows un-normalized data. The right column shows data post-normalization. For the top row, color corresponds to hemoglobin abundance. For the bottom row, color corresponds to run order. The subfigure captions indicate whether the data are pre- or post-normalization as well as what variable the points are colored by. 13
2.6 Boxplots of log abundances for each folate pathway metabolite, stratified by case-control status, following filtering and normalization. Blue = cases (n = 92) and red = controls (n = 92). Note the different y-axis scales. 17
2.7 a) Pairwise scatterplots of the first 5 components from PCA on only the 11 folate metabolites and all subjects, with plotting symbols for subjects colored by case-control status. b) Results of UMAP on only the 11 folate metabolites and all subjects plotted and colored by case-control status. Blue = cases (n = 92) and pink = controls (n = 92). 18
2.8 Pseudo-color image of Spearman correlation matrix for 11 folate pathway features, with rows and columns ordered by complete linkage hierarchical clustering. A darker red indicates a stronger positive correlation and a darker blue indicates a stronger negative correlation. As it is a symmetric matrix, the columns correspond to the same set and ordering of metabolites as the rows. . 18

2.9 Boxplots showing optimistic classification error rates for three classification methods. Results from using all the covariates with: all metabolites (pink), only the folate metabolites (green), and all metabolites except the folate metabolites (blue). The red horizontal line corresponds to a classification error rate of 0.5. . 19
2.10 Boxplots showing optimistic classification error rates for three classification methods on six subsets of the data based on sex, ethnicity, and age at diagnosis. Results from using all the covariates with: all metabolites (pink), only the folate metabolites (green), and all metabolites except the folate metabolites (blue). The red horizontal line corresponds to a classification error rate of 0.5. 20

3.1 Scatterplot of first two principal components of the serum samples (biological=green, pooled QC=orange, blank matrix=red). 34
3.2 Example MD-plots for one batch in DBS data. Panel a) shows metabolites found in all three blank samples. Quintile-specific cutoffs are shown with red line. Panel b) shows the MD-plot for features detected in fewer than three blank samples in the upper-left panel. The remaining three panels show close-ups of each of the three data clouds based on the number of blank samples the metabolite was detected. Green line indicates cutoff used in each of the three cases. 35
3.3 Example MD-plots for one batch in maternal serum data. Panel a) shows metabolites found in all three blank samples. Cutoffs are shown with green line. Panel b) shows the MD-plot for features detected in fewer than three blank samples in the upper-left panel. The remaining three panels show close-ups of each of the three data clouds based on the number of blank samples the metabolite was detected. Green line indicates cutoff used in each of the three cases. 36
3.4 Densities of ICC values estimated by batch from the random-effects model for (a) DBS and (b) maternal serum data. 37
3.5 Pairwise scatterplots of the first five components from PCA on static DBS data (both cases and controls) (a) without and with (b) pre-processing and normalization discussed in Section 3.2. Points colored according to spot age of DBS sample. 38
3.6 Boxplots of log abundances of qPMDs from DBS reactomics data (red) and maternal serum reactomics data (purple). X-axis shows PMD values ordered within each dataset. 39
3.7 Distributions of (a) means, (b) variances, and (d) CVs for DBS and maternal serum qPMDs. (c) scatterplot of means (x-axis) vs. variances (y-axis) for each qPMD. Ellipses are drawn using stat_ellipse in ggplot assuming multivariate t-distribution at level 0.95. DBS = green and maternal serum = purple). 40
3.8 Empirical CDF plot of the upper diagonal of the Spearman correlation matrices between qPMDs for neonatal qPMD (green) and maternal qPMDs (purple). Cumulative density at r=0.5 is indicated with a dashed vertical line. 44

3.9 Mean-difference plot of Spearman correlation values for qPMDs. a) Mean-difference plot of Spearman pairwise correlation coefficients for qPMDs in neonatal cases and controls, where each point corresponds to one of 3,003 pairs of qPMDs (for 78 individual qPMDs). b) Same as in a), for maternal qPMDs, where, to aid visualization, we only plot means and differences for a random subset of 10% (i.e., 3,605) of all possible pairwise correlation coefficients for 269 qPMDs. The x-axis shows the average correlation coefficient for cases and controls and the y-axis shows the difference (cases – controls). Blue line shows a difference of zero. 45
3.10 Volcano plot of (a) neonatal, (b) maternal, and (c) neonatal and maternal qPMDs associated with leukemia. Numbered qPMDs were selected by the ensemble feature selection method. Estimated case-control fold-change from the linear regression is plotted on the x-axis and the negative log of the associated nominal p-value of the t-test is plotted on the y-axis. Red points indicate the qPMD was selected through the concordance of the regression methods and blue points are qPMDs that were selected only by Random Forests. Orange points (in (c)) were also selected through the concordance of regression methods using a more lenient cutoff. The dashed horizontal line corresponds to a nominal p-value of 0.05 for reference. 51
3.11 a) Heatmap of hierarchical clustering results of Spearman correlation values of selected neonatal qPMDs with linolenic and linoleic acid. b) Heatmap of hierarchical clustering results of Spearman correlation values of selected neonatal qPMDs with linoleic and linolenic acid after adjustment. 52

4.1 Heatmap of pairwise Spearman correlation matrix for all features in HC cases and controls (left) with annotations and case-control fold-changes shown (right). Features selected for association with lung cancer are highlighted in red. 64
4.2 Heatmap of pairwise Spearman correlation matrix for all features in LUAD cases and controls (left) with annotations and case-control fold-changes shown (right). Features selected for association with lung cancer are highlighted in red. 65
4.3 Variable importance of covariates and selected features as ranked by Random Forests in HC cases and controls (left), and heatmap showing the corresponding pairwise Spearman correlations (right). 66
4.4 Variable importance of covariates and selected features as ranked by Random Forests in LUAD cases and controls (left), and heatmap showing the corresponding Spearman correlations (right). 66

4.5 a) Example variable importance plot from the bootstrapped LASSO model fits. The adduct features are sorted by the proportion of iterations that they were included in the model and shown on the y-axis. The x-axis shows the corresponding proportion of the 500 iterations that the feature was included in the LASSO model. Blue line shows the line segment used to determine the elbow of the plot (blue point). Red line indicates the cutoff for the top features selected according to the elbow method. b) Calculated adjacent differences from plot a). Red line shows lowest ranked adduct value where adjacent difference > 4× the median difference. c) Same variable importance plot from a) but this time red line indicates cutoff using difference method (from b). . . . 77

4.6 Simulation results with sample size $n = 200$, $\beta = 100\% \times \beta$, $k = 12$, and $k_{int} = 6$. Each of the fifteen panels corresponds to a given combination of the marginal distribution of X and distributions of X and $Y|X$, as indicated by the titles above the top row and along the right side (see Table 4.3). Mean TPR (y-axis) is plotted against mean TNR (x-axis) grouped by the method used for feature selection. Points are colored by method (BH adjusted logistic regression (light orange), unadjusted logistic regression (dark orange), ensemble method using λ_{1se} vs. λ_{min} (blue and green respectively) and mean RF vs. elbow RF (light and dark respectively) and the different cutoff point methods used for the regression features in the ensemble method is indicated by a circle for the difference method and a triangle for the elbow method. 82

4.7 Simulation results with sample size $n = 200$. In each figure, each of the fifteen panels corresponds to a given combination of the marginal distribution of X and distributions of X and $Y|X$, as indicated by the titles above the top row and along the right side (see Table 4.3). Across the columns, the number of features k and k_{int} associated with the outcome was varied (left: $k = 12, k_{int} = 6$, middle: $k = 8, k_{int} = 4$, right: $k = 6, k_{int} = 4$). Across the rows, the magnitude of the regression coefficients (β) was varied (top: 100% β, middle: 75% β, bottom: 50% β). So the panel with the most associated features with the strongest signal is in the top left (a) and the panel with the fewest associated features with the weakest signal is in the bottom right (i). Within each panel, mean TPR (y-axis) is plotted against mean TNR (x-axis) grouped by the method used for feature selection. Points are colored by method (BH adjusted logistic regression (light orange), unadjusted logistic regression (dark orange), ensemble method using λ_{1se} vs. λ_{min} (blue and green respectively) and mean RF vs. elbow RF (light and dark respectively) and the different cutoff point methods used for the regression features in the ensemble method is indicated by a circle for the difference method and a triangle for the elbow method. 83

4.8 Simulation results with sample size $n = 400$. In each figure, each of the fifteen panels corresponds to a given combination of the marginal distribution of X and distributions of X and $Y|X$, as indicated by the titles above the top row and along the right side (see Table 4.3). Across the columns, the number of features k and k_{int} associated with the outcome was varied (left: $k = 12, k_{int} = 6$, middle: $k = 8, k_{int} = 4$, right: $k = 6, k_{int} = 4$). Across the rows, the magnitude of the regression coefficients (β) was varied (top: 100% β, middle: 75% β, bottom: 50% β). So the panel with the most associated features with the strongest signal is in the top left (a) and the panel with the fewest associated features with the weakest signal is in the bottom right (i). Within each panel, mean TPR (y-axis) is plotted against mean TNR (x-axis) grouped by the method used for feature selection. Points are colored by method (BH adjusted logistic regression (light orange), unadjusted logistic regression (dark orange), ensemble method using λ_{1se} vs. λ_{min} (blue and green respectively) and mean RF vs. elbow RF (light and dark respectively) and the different cutoff point methods used for the regression features in the ensemble method is indicated by a circle for the difference method and a triangle for the elbow method. 89

A.1 Boxplots of classification error rates using k-NN on the Monte Carlo sample partitions of the full learning dataset for a range of k values (shown on x-axis). Boxplots are colored by the mean classification error rate by k. Darker red indicates a lower (better) mean classification error rate and dark purple indicates the highest. 107
A.2 Mosaic plot of 2×2 contingency table for joint distribution of case-control status and sex. 108
A.3 Barplot for distribution of age at diagnosis (in years) for cases. 109
A.4 Mosaic plot of 2×3 contingency table for joint distribution of case-control status and race. 110
A.5 Mosaic plot of 2×2 contingency table for joint distribution of case-control status and mode of delivery. 111
A.6 Boxplots of weight percentiles conditional on gestational age, stratified by case/-ontrol status. Calculated using the INTERGROWTH-21st standards. . . . 112
A.7 Boxplots of age at blood collection stratified by case-control status. 113
A.8 Boxplots of number of years from sample collection to sample processing, stratified by case-control status. 114
A.9 Pseudo-color image of Spearman correlation matrix for all 148 metabolites, with rows and columns ordered by complete linkage hierarchical clustering. Darker red indicates a stronger positive correlation and a darker blue indicates a stronger negative correlation. 115

B.1 Heatmaps of hierarchical clustering results using pairwise spearman correlation values between neonatal qPMDs for a) cases and b) controls. 117

B.2 Mean-difference plot of spearman correlation values for full set of maternal qPMDs with 36,046 correlation points (pairwise for 269 qPMDs). X-axis shows the mean and y-axis shows the difference (cases – controls). Blue line shows a difference of zero. 118
B.3 Heatmaps of hierarchical clustering results using pairwise spearman correlation values between maternal qPMDs for a) cases and b) controls. 119
B.4 Heatmap showing hierarchical clustering results for neonatal and maternal PMDs jointly based on pairwise spearman correlation values. The color bar along the left side indicates whether the qPMD is from the neonatal data (green) or from the maternal data (purple). 120
B.5 Clustering using Spearman correlation values of selected DBS qPMDs associated with childhood leukemia. a) Heatmap of hierarchical clustering results in three distinct correlation clusters. b) Silhouette widths for PAM clustering results on the same qPMDs. Red line indicates average silhouette width of 0.58. 121
B.6 Clustering using Spearman correlation values of selected maternal serum qPMDs associated with leukemia. a) Heatmap of hierarchical clustering results in three distinct correlation clusters. b) Silhouette widths for PAM clustering results on the same qPMDs. Red line indicates average silhouette width of 0.58. 121
B.7 Heatmaps of hierarchical clustering results using pairwise Spearman correlation values between selected neonatal and maternal PMDs. 122
B.8 Distribution of elemental carbon changes of qPMDs from the KEGG database compared to those selected qPMDs in DBS and maternal serum associated with child leukemia status. 123
B.9 Random Forests variable importance plot using the mean decrease in Gini index for prediction of case-control status using the selected neonatal qPMDs and covariates. 124
B.10 Clustering using Spearman correlation values of selected neonatal DBS qPMDs associated with childhood leukemia and continuous covariates. Left: Heatmap of hierarchical clustering results in four distinct correlation clusters. Right: Silhouette widths for PAM clustering clusters the same qPMDs. Red line indicates average silhouette width of 0.48. 125
B.11 Random Forests variable importance plot using the Mean decrease in Gini index for prediction of case-control status using the selected maternal serum qPMDs and covariates. 126
B.12 Clustering using Spearman correlation values of selected maternal serum qPMDs associated with childhood leukemia and continuous covariates. a) Heatmap of hierarchical clustering results in four distinct correlation clusters. b) Silhouette widths for PAM clustering clusters the same qPMDs. Red line indicates average silhouette width of 0.35. 127

C.1 T3 peptide (mostly Cys34) features with sufficient data for statistical analysis. (Mean peak areas are based on normalized data after imputation for missing values). 131
C.2 Lys525 peptide features with sufficient data for statistical analysis. (Mean peak areas are based on normalized data after imputation for missing values). 132
C.3 Ensemble of regression and classification methods for selection of adduct features associated with case-control status for HC cases. A) Volcano plot of nominal *p*-values for the case-controls status regression coefficient in a multivariate linear regression of each adduct's log abundance on case-control status and other covariates, with selected adducts shown in red (dashed line indicates nominal *p*-value=0.05); B) proportion of times that a given adduct was selected by a regularized logistic regression (LASSO) of HC case-control status; C) ranked variable importance measures from a Random Forests classification of case-control status. 136
C.4 Ensemble of regression and classification methods for selection of adduct features associated with case-control status for LUAD cases. a) Volcano plot of nominal *p*-values for the case-control status regression coefficient in a multivariate linear regression of each adduct's log abundance on case-control status and other covariates, with selected adducts shown in red (dashed line indicates p=0.05); b) proportion of times that a given adduct was selected by a regularized logistic regression (LASSO) of LUAD case-control status; c) ranked variable importance measures from a Random Forests classification of case-control status. 137
C.5 Plots of ln(case-matched-control fold change) versus time to diagnosis (ttd) in years from recruitment for adduct features selected for HC cases-controls. *P*-values are shown for slopes of the relationship. 138
C.6 Plots of ln(case-matched-control fold change) versus time to diagnosis (ttd) in years from recruitment for adduct features selected for LUAD cases-controls. *P*-values are shown for slopes of the relationship. 139

List of Tables

2.1 Classification error rates for each candidate class predictor after being trained on the learning set, and then evaluated on the hold-out test set. 22
2.2 Descriptive characteristics of the pregnancy and early-life demographics of the leukemia cases and controls (n = 244) in Chapter 2 and mother-child leukemia pairs used in Chapter 3 (n = 244 pairs). 25

3.1 Description of untargeted metabolomics and associated data for neonatal DBS and maternal serum samples. 33
3.2 Associations between log linolenic acid and linoleic acid metabolite abundances on pediatric ALL case-control status. Regression coefficients were estimated using univariate logistic regression models adjusted for batch, run order, spot age, hemoglobin, infant sex, and ethinicity as described in Petrick et al. [62] 43
3.3 Selected qPMDs, the pairs of *m/z* used to create them, element composition of reaction, potential enzyme involved, and KEGG reaction class ID. 48

4.1 Summary statistics for covariates measured on histologically-confirmed (HC) cases, LUAD (lung adenocarcinoma) cases, and controls. 62
4.2 Features selected by the ensemble pipeline for association with lung cancer. Selection criteria satisfied for each feature are shown in boldface. [Fold-change based on normalized and imputed data (FC); linear regression nominal p-value; bootstrap LASSO; Random Forests (RF)] 63
4.3 DGP Combinations of X and Y 74
4.4 Feature selection results on the HC subset of the lung cancer data using each of the eight previously described automated ensemble feature selection methods. A given list is read vertically, corresponding to the options listed above it. For example, the first column corresponds to using the λ_{1se} and the "difference" options for the LASSO and the "plots" option for RF. Features added by each automated approach compared to the manual implementation in Section 4.2 are listed below the double line. 85

4.5 Feature selection results on the LUAD subset of the lung cancer data using each of the eight previously described automated ensemble feature selection methods. A given list is read vertically, corresponding to the options listed above it. For example, the first column of features correspond to using the λ_{1se} and the "difference" options for the LASSO and the "plots" option for RF. Features added by each automated approach compared to the manual implementation in Section 4.2 are listed below the double line. 86

A.1 Spearman correlation coefficients between blood specimen age (defined as the numbers of years between the sample collection and processing at the laboratory) and measured metabolites in the folate pathway. 108

C.1 Intraclass correlation coefficients (ICC) and coefficients of variation (CV) from the random effects model for features abundances from duplicate nLC-HRMSMS injections. 133
C.2 Case-control fold changes (FC) and nominal *p*-values after normalization and imputation for HC lung cancer cases and controls. 134
C.3 Case-control fold changes (FC) and nominal *p*-values after normalization and imputation for LUAD cases and controls. 135

Acknowledgments

Thank you to Sandrine Dudoit, Stephen Rappaport, and Haiyan Huang for all their support and guidance throughout my time at Berkeley as well as their generous feedback in writing this dissertation.

Thank you to all my collaborators and coauthors for their help with the various analyses and manuscripts included here. In particular, thank you to Sandrine Dudoit, Catherine Metayer, Lauren Petrick, Libby Morimoto, Xiaomei Ma, and Joseph Wiemels for their help with the targeted metabolomics manuscript *One-Carbon (Folate) Metabolism Pathway at Birth and Risk of Childhood Acute Lymphoblastic Leukemia: A Biomarker Study in Newborns* which was published in *Cancers* in 2023 [71]. The second chapter of this dissertation is largely comprised of this work.

Thank you to Sandrine Dudoit, Catherine Metayer, Libby Morimoto, Lauren Petrick, and Miao Yu for their help with the reactomics analysis and draft, which largely comprises the third chapter of this dissertation.

Lastly, thank you to Sandrine Dudoit, Stephen Rappaport, Hasmik Grigoryan, Qing Lan, and Nathaniel Rothman for their help with the untargeted adductomics lung cancer manuscript *Albumin Adductomics in the Shanghai Women's Health Study Links Lung Cancer in Never-Smokers with Air Pollution, Redox Biology and One-Carbon Metabolism*. This work comprises part of the fourth chapter of this dissertation.

Chapter 1

Introduction

With the overall goal of better understanding disease etiology to aid in prevention or treatment, researchers have focused on various areas of study to try to tease out these relationships. A common area since the completion of the Human Genome Project, is the use of genetic markers, which allows researchers to examine relative gene expression levels between individuals with and without a disease [102]. Genomic technologies and analyses have seen rapid advancement in the subsequent decades and have provided invaluable insights into genetic causes for a range of diseases [31, 49]. However, it has become increasingly apparent that genetics is only a part of the equation, and to more fully understand the underlying risk factors will require looking at how an individual's exposures interact with their genome to lead to a disease phenotype [28].

The central dogma of biology states that genetic information is stored in DNA, which is transcribed to RNA, which is then translated to proteins. These proteins are then involved in and regulate metabolic processes throughout the body [90]. The way that genetic information gets passed on at each stage eventually reaching a phenotype can be influenced, not only by the actual genetic code itself, but also by exposures acting on and interacting with the process. The set of exposures an individual experiences, such as caffeine, sunlight, various pollutants, chemical agents, dietary factors, etc. comprise what is referred to as the exposome [67]. An individual's exposome dynamically interacts with their genome and causes various intermediate byproducts that can be associated with disease [15]. Exposomics, or the study of the exposome, is focused on better understanding this component of disease risk [28]. The past couple of decades have seen great advancements in various high-throughput technologies focused on different biological data sources that are designed to detect these intermediate measurements. Examples of areas of study that focus on this are proteomics, metabolomics, and epigenomics, to name a few [120]. In each case, studying exposomic data (whether it is the abundance of proteins, or DNA methylation levels, or small molecule levels between cases and controls) is meant to use information stored in these proxies for different exposures and shed light on what biological pathways seem relevant in understanding the complex disease of interest. Zhang et al. describe the motivations and range of applications for exposomic studies in

detail, which include better understanding causes of diseases and environmental effects, particularly relevant amidst climate change [137].

While exposomic data holds important information for uncovering the nuances of different disease etiologies, the technologies used to acquire the data can produce artifacts which threaten to mask the signal of interest. In large dimensional datasets such as untargeted metabolomics or DNA methylation data, there can be a large amount of uninformative features that should be removed before proceeding with analysis. In each of these cases applying a data-adaptive filtering process has been shown to improve the retention of useful features while removing uninformative ones [92]. There can also be the issue of data sparsity or missing values, whether at random, due to the limit of detection of the instrument, or otherwise. Imputation is often necessary for application of methods and a wide range of methods have been proposed (zero, mean or median imputation, imputing with a fraction of the limit of detection, *k*-NN, Random Forests, etc.) [114]. The researcher must carefully consider the properties of their data and consider the implications of applying their chosen method(s) as results can vary greatly depending on how data is imputed.

Technical artifacts can be created due to logistic necessities. For example, analyzing samples in separate batches and sequentially, can introduce artifacts based on when a sample happened to be analyzed [90]. Similarly, the abundance of metabolites can naturally depend on the volume of the sample that was analyzed. Hence, a normalizing variable, such as hemoglobin abundance for blood samples, can act as a proxy for blood volume to try to adjust for that effect [134]. There are also adjustments that need to be made to try to minimize the effects from differences in how the samples were taken or how long they were stored before analysis. In each of these cases, properly adjusting for these factors in the pre-processing stage is important to allow for the detection of the set of biological signals of potential interest in downstream analyses.

The issues regarding careful pre-processing and analysis of the data are intertwined with the notion of reproducibility. Historically, there have been challenges regarding reproducibility (or lack thereof) in high-throughput exposomic analyses. This concerns how data are created (samples prepared and analyzed), features are created (peak picking in untargeted metabolomics and ChIP-Seq, retention time aligment, sequencing aligment etc.) as well as the subsequent data analyses performed [78, 93, 92]. Employing data-adaptive and automated pre-processing steps and analyses help remove user bias and facilitate reproducibility. Careful documentation of methods and decisions made for each stage are necessary.

As the exposome is dynamic and variable across time and space, appropriate analyses that recognize this are necessary [137]. Thus, using traditional approaches applied for categorical and/or static data such as genome-wide association studies (GWAS), fail to recognize the vastly different characteristics of these two types of data and are, therefore, ill-suited for the exposomic analyses at hand. The measurement error inherent in the range of continuous exposomic features under study, as well as the potentially more complicated dose-response relationship, further complicate detection of the signal of interest.

In this dissertation we will discuss three such analyses using three different types of exposomic data. The first chapter discusses the application of a prediction framework using targeted metabolomics data to examine a potential relationship between an exposure of interest (folate pathway disregulation) and pediatric ALL. The second integrates reactomics (reaction-level) data transformed from untargeted metabolomics data of mother-child pairs to generate novel hypotheses for risk factors of pediatric ALL. The final chapter utilizes similar feature selection methodology for hypothesis generation of risk factors of lung cancer in never-smokers in an untargeted adductomics dataset as well as providing simulation results to benchmark methodology performance. In each case, we illustrate EDA, pre-processing, and analyses to uncover the biological signal of interest and discuss statistical considerations inherent to the problems at hand.

Chapter 2

Targeted Metabolomics: Examining Folate Pathway Disregulation in Pediatric Acute Lymphoblastic Leukemia

2.1 Introduction

Metabolomics, the exposomic area discussed in this chapter, is the study of low molecular weight molecules (including metabolites), found in a biological specimen, also referred to as the metabolome. These small molecules can be endogenously created as a byproduct of upstream metabolic processes or introduced exogenously from the environment [68]. As such, the metabolome holds information regarding an individual's exposures, and how these exposures may interact with ones genetics to lead to higher disease risk. Broadly, there are two types of metabolomics studies: targeted and untargeted. The goal of untargeted metabolomics studies is to detect whatever small molecules exist in a sample within a specified size range [90]. It provides an agnostic manner to detect any organic compounds present across a large range of biochemical properties and, through various analyses such as variable selection, gain insight into markers associated with the disease of interest [98, 133]. However, conclusions drawn are often hampered by the inability to confidently identify the selected biomarkers. On the other hand, targeted metabolomics uses an in house library of standards to specifically probe for a known set of targets. As such, while it no longer reflects measurements of the full range of molecules present, the identity of any detected metabolite is known, which increases ease of interpretation of findings. For the purpose of this chapter, we will discuss targeted metabolomics and will leave discussion of untargeted metabolomics largely for Chapter 3.

Metabolomics data are acquired through various technologies including nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry [90]. In the case of our data, liquid chromatography high-resolution (LC-HR) mass spectrometry was used with HILIC chromatography, which allowed for the detection of certain metabolites of interest. Be-

cause of the targeted nature, metabolites are annotated by matching them to in-house standards that have been previously benchmarked. Targeted metabolomics datasets, therefore, do not result in the large numbers of unknown features (a considerable portion of which are not reliable) that other datasets may suffer from. Therefore, the same level of heavy filtering necessary in an untargeted metabolomics dataset, for example, is not needed in this context. That being said, it is still important to examine the data and remove features that seem to show strong evidence of a high noise-to-signal ratio. Even after feature filtering, the data still suffers from a high amount of noise and other pre-processing measures previously described are also important to consider, as failure to do so will result in the masking of any biological signal with technical noise [92]. Blood volume, in particular for our samples, must be accounted for by using Hb abundance as a proxy [134].

As the researcher has confidence in the annotation of each feature, targeted metabolomics studies are often hypothesis driven and therefore lead to hypothesis testing. Beyond the often stringent distributional assumptions necessary depending on the test, using a hypothesis testing approach necessitates the clear identification of a statistical hypothesis of interest. This may be reasonable in a univariate setting where testing whether the abundance of a single metabolite differs between cases and controls. However, in the context of testing the hypothesis that a group of metabolites differs between cases and controls, this may be difficult to translate to a specific statistical test, given the complexity of the biochemical processes involved. This is especially the case where, as with the folate pathway discussed below, the posited directed graphical structure has cyclic components, shown in Figure 2.1. In addition, having only a subset of the relevant metabolites measured adds additional challenges. Although there are proposed methods to test hypotheses for multiple target molecules [99, 18, 89], they may not be interpretable in cases where relationships are complex and directionality differs between features.

Since the general biological question is whether a certain pathway (as evidenced by a subset of measured metabolites belonging to it) seems associated with disease status, we propose using a prediction framework to allow flexibility regarding relationships between metabolites and the outcome. Prediction has been used in many different contexts, including various -omics studies (though mainly to select features). These general principles can be applied to our question and data as well, with certain considerations [106, 32]. An essential component of building and evaluating predictive models is to have an independent dataset that can validate the models built using the learning data. This practice avoids overly optimistic assessments of predictive performance. Ideally, one would have data from a replicated experiment, but often in practice data is split from the onset into learning and test sets to maintain independence. An additional challenge in assessing prediction performance with metabolomic data, is the need for careful pre-processing steps to remove unwanted technical variation. Due to small sample sizes in most metabolomics studies, cross-validation is used after pre-processing and normalization (particularly with imbalanced classes) [32]. While this may be appealing because it allows all the data to

be used during the pre-processing, model building, and testing stages, the training and testing data are no longer independent. Thus, splitting the learning and testing dataset before pre-processing is necessary to obtain unbiased estimates of predictive performance.

Maintaining the independence of the training and testing set, however, poses some additional challenges. Because using a data-adaptive pre-processing pipeline is preferable, one needs to be aware of differences that may arise between the learning and testing data due to data specific decisions made during the process and consider potential downstream implications [93]. Alternatively, cross-validation can be used. However, it would require pre-processing each of the splits separately, a logistically lengthy process, which may also compound the issues raised with one splitting.

Assessing the impact of disregulation of a pathway using prediction also allows for flexibility in the models used and does not necessitate subscribing to either specific unrealistic underlying distributional assumptions or a hypothesis that is not reflective of the scientific question at hand. In addition, change in predictive performance can be examined, not just by how predictive a set of metabolites are, but also by how much predictive capability is lost when they are removed, allowing another perspective from which to view the question. In the event that the set of metabolites look promising using preliminary models, more complex models can be built to try to improve predictive performance, always using the independent test set for assessing classification performance. In this chapter, we discuss the analysis of one such targeted metabolomics dataset in a pediatric ALL case-control study and the prediction approach we took in determining whether disregulation of the folate pathway, reflected by a subset of the metabolites in the dataset, was involved in increasing disease risk.

2.2 Background

Leukemia is the most common cancer among children in industrialized countries, and its incidence has increased in the past 40 years [10, 60]. Acute lymphoblastic leukemia (ALL) accounts for 80-85% of all leukemia diagnosed in children (0-14 years), and despite improvements in treatment, not all children have benefited equally from a favorable prognosis and most survivors experience complications throughout their life [79]. Therefore, identification of modifiable risk factors of childhood leukemia is of high public heath relevance. Childhood ALL exhibits a peak in incidence at the age of 2-5 years, and several experiments have demonstrated that chromosomal abnormalities detected at diagnostic were also identified at birth (such as ETV6-RUNX1, RUNX1-RUNX1T1, and PML-RARA gene fusions, as well as high hyperdiploidy), providing strong support to the prenatal origin of many childhood leukemia subtypes [33, 118, 117, 119, 38]. In addition, a recent study on discordant twins for ALL supports a role for DNA methylation alterations in utero impacting leukemogenesis [76]; such DNA methylation patterns may be influenced by periconceptional nutrients including folate [77]. In epidemiologic studies using interview data, prenatal folic acid and vitamin intake from dietary sources

and supplementation [70, 96, 95, 2, 109] has consistently emerged as a protective factor for childhood leukemia. B-vitamins (e.g., folate) and amino-acids (e.g., methionine) are important micronutrients in the one-carbon (folate) metabolism pathway that supports the maintenance of DNA and provision of methyl groups for epigenetic control of DNA expression. In contrary, lack of these nutrients can lead to cell damages including hematotoxicity, as best illustrated by one of the many side effects of methotrexate, a folate antagonist drug that decreases nucleic acid synthesis, and thus compromise DNA synthesis, repair, and cellular replication [39]. However, despite a healthy folate intake during pregnancy overall, it is not clear whether inter-individual differences in the ability to metabolize folate may modulate risk. One-carbon metabolism nutrients support the embyro's DNA programming mostly during preconception and the first weeks after conception. Folate levels during the remainder of the first trimester (referred to as early pregnancy) are critical for prevention of birth defects (neural, heart, and lip), yet it is unknown at which point these nutrients are critical for reduction of childhood leukemia risk particularly towards the third trimester (referred to as late pregnancy). Direct measurement of downstream metabolites at birth can provide insight to this latter question. Our group has previously measured folate species in neonatal blood spots of 357 children with ALL and 405 controls and reported no association between neonatal folate concentrations at birth and risk of pediatric ALL [20]. Little is known, however, about other key metabolites in the folate metabolism pathway. For example, homocysteine concentration is a marker of folate status (with an inverse relationship) and a metabolite closer to DNA methylation in the one-carbon cycle, and increased levels have been associated with cancers in adults [73, 4]. Anecdotally, a child diagnosed with methylmalonic acidemia, a condition that leads to high levels of homocysteine, was reported to have multisystem disorders, including leukemia [142]. However, no population-based study has examined its potential role on development of childhood leukemia to date.

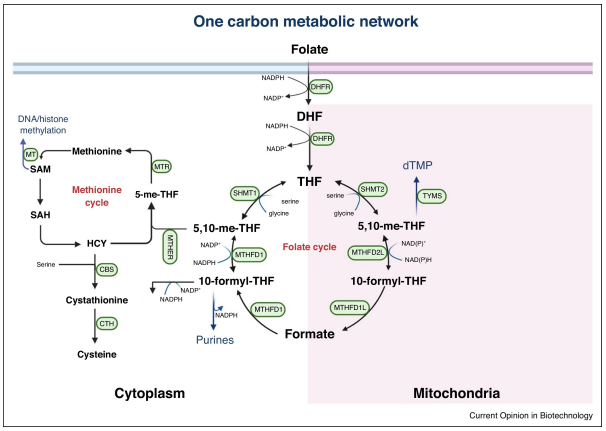


Figure 2.1: One-carbon (folate) metabolism pathway. The data analyzed in our study correspond to a subset of 11 metabolites involved in this pathway [55].

The goal of our study is to expand this characterization by using novel metabolomics techniques applied to neonatal blood specimens in a registry-based case-control study of childhood ALL conducted in California.

2.3 Materials and Methods

Study Population

Our study population of interest is children diagnosed with ALL before they are 15 years old. We utilized archived neonatal dried blood specimens (DBS) from a subset of children diagnosed with ALL at age 0-14 years (International Classification of Disease for Oncology (ICD-O) codes 9820, 9823, 9826, 9827, 9831-9837, 9940, 9948) and controls identified through a California registry-based mother-child study linking electronic records from the cancer registry (1988-2011) with birth data from vital statistics. Of the 137 ALL cases originally identified, we included 122 children diagnosed at age 1 to 14 years and with available specimens from the California Department of Public Health biobank program. Controls were 1-to-1 matched on sex, year and month of birth, and race/ethnicity.

Metabolite Extraction and Liquid Chromatography-High Resolution Mass Spectrometry (LC-HRMS) Analysis

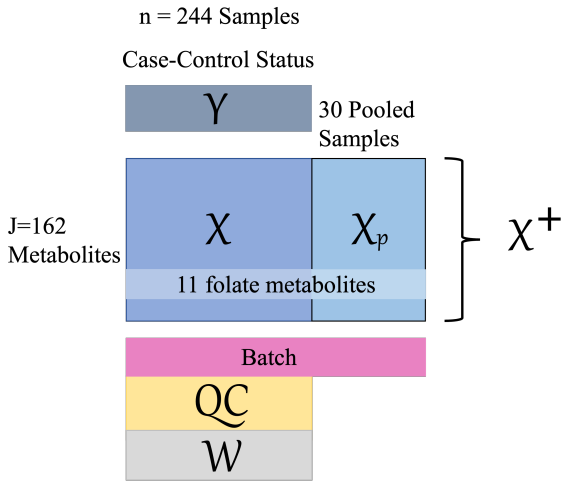


Figure 2.2: Data structure for targeted metabolomics dataset.

DBS were analyzed using a methodology developed by our team [87, 62]. In short, DBS punches were extracted with 100 μL of water at room temperature (15 min, 1400 rpm), and a 5 μL aliquot reserved for hemoglobin measurements to adjust for original blood volume [128], 400 μL of acetonitrile containing isotopically labeled internal standards (IS) was added to the remaining aqueous solution containing the DBS punch, samples agitated (1400 rpm, 37°C, 1 h), and protein precipitated at -20°C for 30 min. The supernatant was evaporated to dryness and stored at -80°C until analysis. Immediately prior to analysis, samples were reconstituted and analyzed LC-HRMS with HILIC chromatography to capture folate pathway metabolites [75]. A pooled quality control sample was prepared by combining aliquots of all of the biological samples of interest and was then injected routinely throughout the run to monitor instrument stability and facilitate batch and run order correction [112]. The biological and pooled samples were run in two sequentially-analyzed batches. The pooled sample was run 30 times (15 in each batch) and at least once every 10 biological samples. Eleven folate pathway metabolites as well as 151 additional metabolites across a broad range of biological processes were identified considering retention time, accurate mass and MS/MS matching, when available, with our in-house library of standards analyzed under the same LC-HRMS conditions using Profinder and Find by Formula software (Agilent Technologies). The folate pathway metabolites measured in this study included methionine, betaine/valine, homocysteine, choline, taurine, dimethylglycine/2-aminoisobutyrate, cysteine, glycine, serine, adenosylhomocysteine, and cystathionine. Metabolites with multiple annotations are those that co-eluted and could not be distinguished.

Data

The structure of the data is shown in Figure 2.2. The data consist of a $J \times n$ matrix X of metabolite abundances for J metabolites and n samples, where the folate pathway related metabolites are a subset of the J features (i.e., $X_{folate} \in X$). The data also include abundance

measures for each of the J metabolites in each of the 30 runs of the pooled sample, the X_p matrix. The joint X and X_p dataset is referred to as X^+. For each sample in X^+, biological or pooled technical, there is a binary batch variable b. In addition, for each of the biological samples in X, there are four measured continuous quality control variables in QC: *DBS spot age*, *age at collection*, *runorder*, and *hemoglobin*, as well as six baseline covariates in the W matrix: *sex* (binary), *race* (categorical), *education* (ordinal categorical), *delivery* (binary), and *birth weight conditioned on gestational age* $\geq 90^{th}$ *percentile based on INTERGROWTH21 standards* [82] (binary). Lastly, there is the binary outcome vector Y, with entries in $\{0, 1\}$, which indicates case-control status. Our data then consists of n identically-distributed random draws of $(X, W, Y) \sim P_0 \in \mathcal{M}$.

Data Pre-processing

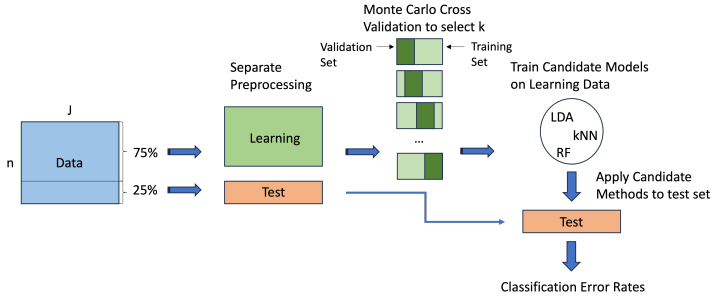


Figure 2.3: Flow chart showing the partitioning of the data into learning and test sets, followed by separate pre-processing, model training on the learning set using Monte Carlo cross-validation, and culminating in prediction evaluation on the test set.

All data visualization and analyses were performed in the R environment [1]. For the purpose of prediction performance assessment, the 244 samples (122 cases and 122 controls) were randomly separated into learning (75%; $n_L = 186$) and test (25%; $n_T = 60$) sets, maintaining the matched pairing. An overview of the partitioning and modeling process is illustrated in Figure 2.3. The pooled samples were also randomly partitioned by batch (for each batch, 10 for learning set and 5 for test set). Random partitioning of the data in this manner maintains the study design and is standard when attempting to estimate predictive ability. It allows for training of classifiers on a learning dataset and then evaluation of the candidate classifiers on an independent test dataset. This extra precaution is to ensure that prediction estimates are not overly optimistic due to overfitting that can occur when evaluating classifiers on the same data that they were trained on. Unless otherwise mentioned, this next portion describes analyses performed on the learning set.

First, the 162 metabolites were filtered based on Intraclass Correlation Coefficient (ICC) and percent missing. To perform the ICC filtering, the following random effects model

was fit separately for each metabolite using the pooled samples as technical replicates in order to estimate the proportion of the variance that was due to technical effects as opposed to biological effects [93].

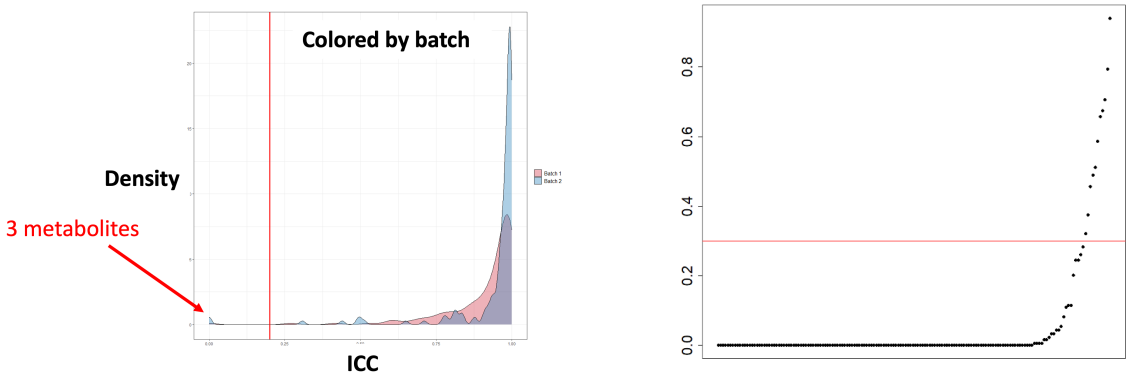
$$X_{i,j,k} = \mu_j + \alpha_{i,j} + \epsilon_{i,j,k}, \tag{2.1}$$

where $X_{i,j,k}$ is the log abundance of metabolite j in subject i's kth replicate (including the pooled samples which are treated here as "subjects"), μ_j is the overall mean log abundance of metabolite j, $\alpha_{i,j}$ is the subject-specific random effects term, and $\epsilon_{i,j,k}$ is the random error for replicate measurement k for metabolite j in subject i. Note that the pooled samples are the only ones for which we have technical replicates. The model was fit separately for each of the two batches in b and ICC values were computed for each metabolite J as the estimated fraction of the variance that is attributed to the between-subject variance, i.e.,

$$\hat{ICC}_{batch} = \frac{\hat{\sigma}^2_{between-subject}}{\hat{\sigma}^2_{between-subject} + \hat{\sigma}^2_{technical-replicates}}.$$

Upon visually inspecting the distribution of ICC values, per the recommendation of Schiffman et al. [93], a conservative empirical ICC cutoff of 0.2 is used for each batch and only metabolites that passed for both batches were retained, removing three metabolites for a remainder of 159 (Figure 2.4a). The more conservative cutoff seemed appropriate given that, as the data was generated from a targeted (rather than untargeted) metabolomics analysis, there are not that many features to begin with that need be filtered and their peak quality can be manually verified.

Following this, a two-step process was used to remove metabolites with too many missing values for reasonable imputation. The percentage of missing values for each metabolite across all study samples was calculated and any metabolite with more than 60% missing values was removed. For any metabolite with 30-60% missing values, a Fisher's exact test was used to check for differential missingness by case-control status [93]. This two-step process allows for retaining metabolites with lower percentages of missing values, where missing values appear to be related to the biological condition of interest. As there did not appear to be evidence for differential missingness, all metabolites with more than 30% missing values were removed (shown in Figure 2.4b), leaving all 11 folate pathway metabolites and an additional 137 metabolites in the dataset for analysis. Imputation was then performed using the k-nearest neighbor algorithm with the number of neighbors set to 5 [105]. Finally, the scone package was used to evaluate the relative ability of various normalization schemes to remove unwanted variation while maintaining the biological signal of interest [21]. In addition to the imputed data, scone was provided with batch, case-control status, and a QC matrix, which included variables that were thought to help control for various sources of unwanted variation, including blood spot age, run order, hemoglobin, and the newborn's age (in hours) when the sample was taken. Scone was also provided with five scaling options [identity (no scaling), DESeq, Upper



(a) (b)

Figure 2.4: a) Density plot of ICC values by batch. Three removed metabolites with ICC $\leq$ 0.2 shown. b) Percent missing values (y-axis) by metabolite (x-axis). Red line shows 30% cutoff.

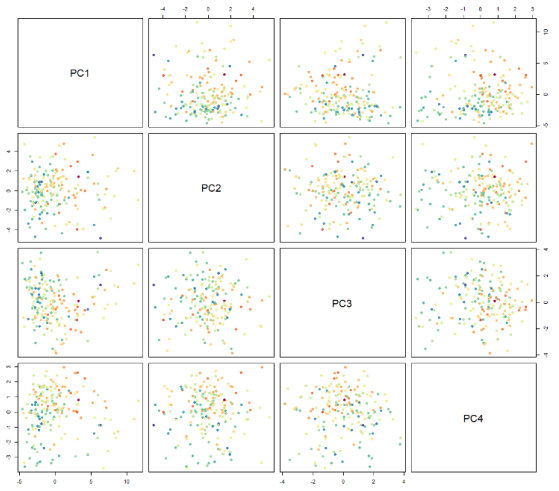
Quartile (UQ), Variance Stabilization (VSN), and Trimmed Mean of M-values (TMM)] [7, 108, 46, 91]. Based on the scone performance measures, UQ scaling while adjusting for case-control status, batch, and the full QC matrix was selected and implemented to acquire the normalized data.

Figure 2.5 shows the removal of technical effects for two of the quality control variables: hemoglobin abundance (top) and run order (bottom). In each of the four panels (a-d), pairwise scatterplots are shown for the first four principal components using PCA. We can therefore see the removal of the respective technical effects from the data post-normalization by looking left to right. For hemoglobin, we can see that it is contributing to the first and fourth principal component pre-normalization, while that effect is removed post-normalization. Similarly, run order seems to be predominantly contributing to the first principal component (based on the plot of pre-normalized data), while that effect is no longer visible in the post-normalized data. This further illustrates the necessity to carefully remove technical artifacts whenever possible, to enable more successful downstream analyses.

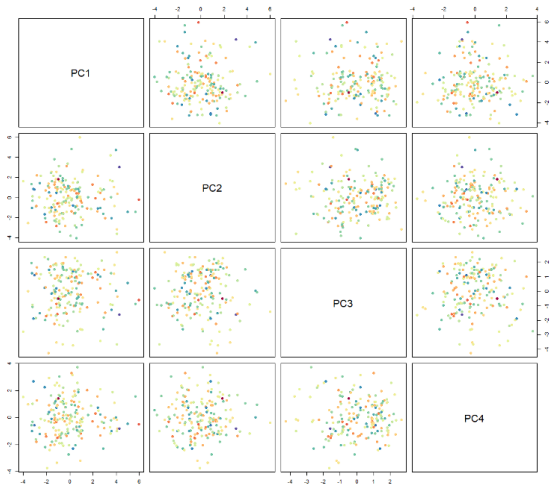
Statistical Analysis

Learning Data

Using the normalized data, the abundances of the 11 folate pathway features were visualized univariately as well as by using dimensionality reduction techniques such as Principal Components Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). After visualization of the data, we consider the question of whether and



(a) Pre-normalization: Hemoglobin



(b) Post-normalization: Hemoglobin



(c) Pre-normalization: Run order



(d) Post-normalization: Run order

Figure 2.5: Pairwise scatterplots showing the first four principal components of the abundance measures for the biological samples. The left column shows un-normalized data. The right column shows data post-normalization. For the top row, color corresponds to hemoglobin abundance. For the bottom row, color corresponds to run order. The subfigure captions indicate whether the data are pre- or post-normalization as well as what variable the points are colored by.

to what extent the folate pathway metabolites are predictive of childhood ALL. In our prediction framework, we would therefore like to find a mapping $\psi : X, W \rightarrow Y$ which minimizes the true risk $\Theta_L(\psi, P_0) \equiv \int L((y|x,w),\psi)dP_0(x,w,y)$, where here we focus on the 0-1 loss function. As the true underlying data distribution P_0 is unknown, we will try to minimize the empirical risk with a limited number of candidate models $\hat{\psi}_n$ and use cross-validation when necessary to avoid overfitting. Namely, observing P_n we want to select $\hat{\psi}_n = \arg\min_{\psi' \in \Psi} \Theta_L(\psi', P_n(y|X,W))$. Our goal is to select our candidate class predictors from a diverse set of predictors (both linear and non-linear) to allow for the detection of various types of relationships within the data. As such, the candidate class prediction methods used were Linear Discriminant Analysis (LDA), Random Forests (RF), and k-nearest neighbors (k-NN) using a range of number of neighbors. In this case, k-NN refers to the classification algorithm rather than the imputation algorithm. Note that, if the number of features was large, a variable selection step may be beneficial to help some of the methods avoid overfitting. As we only have 11 folate related features, this is not necessary.

Linear Discriminant Analysis. LDA classification models each class conditional density as a multivariate Gaussian, assuming a common covariance matrix for all classes ($\Sigma_k = \Sigma$ $\forall$ class k), estimates necessary parameters from the learning data, and creates a decision boundary based on the log-ratio of the class posterior probabilities for the two-class scenario which takes the form of a hyperplane in the covariate space. Specifically, LDA models each class k's density as a multivariate Gaussian:

$$f_k(x) = \frac{1}{(2\pi)^{J/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

and computes the posterior class probabilities $P(Y = k|X = x, W = w)$ using Bayes Rule and prior probabilities π_k of belonging to class k, where $\sum_k \pi_k = 1$. The log-ratio of the posterior class probabilities for two different classes k_1 and k_2 is then given by:

$$log\frac{P(Y = k_1|X = x, W = w)}{P(Y = k_2|X = x, W = w)} = log\frac{\pi_{k_1}}{\pi_{k_2}} - \frac{1}{2}(\mu_{k_1} + \mu_{k_2})^T\Sigma^{-1}(\mu_{k_1} - \mu_{k_2}) + x^T\Sigma^{-1}(\mu_{k_1} - \mu_{k_2}),$$

where the covariance matrix for each group is assumed to be equal ($\Sigma_{k_1} = \Sigma_{k_2} = \Sigma$). This defines a decision rule that is linear in X. Necessary parameters π_k, μ_k, and Σ are estimated from the learning data $\forall$ k and in the two-class classification setting, the LDA rule classifies to case 2 when

$$x^T\Sigma^{-1}(\hat{\mu_2} - \hat{\mu_1}) > \frac{1}{2}(\hat{\mu_2} + \hat{\mu_1})^T\hat{\Sigma}^{-1}(\hat{\mu_2} - \hat{\mu_1}) - log\frac{N_2}{N_1},$$

where N_1 and N_2 are the number of observations in each group respectively. Though LDA assumes approximately Gaussian data and equal covariance structures between the classes, it has been shown to work perhaps surprisingly well in practice, even compared to more complex models, when these assumptions are not met. This may stem from the

fact that, though it has relatively high bias due to the assumption of a linear decision boundary, it benefits from the lower variance in estimating this boundary. In addition, simpler decision boundaries may be all that the data can support, and therefore make the estimates more stable [40]. As LDA requires the notion of relative distances in the projected space, centering and scaling predictors is often common practice.

Random Forests. Random Forests, on the other hand, does not specify a particular parametric model, and rather builds a large number of decision trees through resampling both the samples and features. It first bootstrap samples from the n observations and then, for each bootstrapped sample, randomly samples a subset of the features J for each decision node from which it greedily makes the best split. This is repeated over many iterations and as such, grows a forest of decision trees. The Random Forests classifier Ψ_{RF} then classifies new data points based on a majority rules voting scheme where each decision tree casts a vote. It makes no distributional assumptions on the data and can handle a mix of variable types on different scales [14]. If allowed to grow to a sufficient depth, trees have low bias, though are noisy, which is why they benefit from aggregating (hence a forest). As an individual tree has low bias, so does an aggregation of a set of independently grown trees, with the benefit that the aggregation serves to average over some of the noise and lower the variance. Each of the trees then votes on what the outcome for a given sample should be and majority rules [40]. Random Forests classifiers are appealing for a range of reasons. As previously mentioned, in addition to not needing any distributional assumptions, they are flexible in the types of outcomes and predictors they allow, tend to avoid overfitting due to the resampling process, and are robust to outliers.

Nearest Neighbor Classification. The final method we consider is k-nearest neighbor (k-NN) classification. Using k-NN, a test sample is labeled by a vote of their k nearest neighbors according to some distance metric, such as the Euclidean distance used here. The tuning parameter k determines how adaptive the decision boundary is. The smaller the number of neighbors k, the noisier the decision boundary and the larger the risk of overfitting. On the flip side, larger k results in more stable results though at the risk of using neighbors that are no longer similar enough to the sample whose label is being predicted. It is typical to use cross-validation to identify an appropriate value of k, which we describe below [40]. Similar to LDA, and especially in the case of predictors on different scales, it is often preferable to center and scale predictors to create more meaningful distances in the feature space.

By using three very different prediction methods that have traditionally performed well under a variety of data distributions, we can cast a wide net in trying to identity a predictive relationship between the folate pathway features and our outcome. For each method, the normalized abundances of the 11 folate pathway metabolites as well as sociodemographic and reproductive factors that have been associated with ALL risk [116]

were provided as input. These covariates are sex (binary), race/ethnicity (categorical), mode of delivery (binary: vaginal vs. cesarean), mom's education level (categorical), and birth weight conditional on gestational age (binary: above or below the 90th percentile using INTERGROWTH 21 standards) [82]. In the case of LDA and *k*-NN, because they rely on distances between observations to decide similarity, each predictor was first mean centered and standardized by its standard deviation. In the case of non-ordinal categorical variables, dummy variables were first created. In order to select the optimal number of neighbors *k* in the *k*-NN algorithm, Monte Carlo cross-validation was used for a range of *k* values from 3 to 25. For each iteration, the data were split into 60% training and 40% validation sets, each classifier was trained on the training set and classification errors were calculated on the validation set. The *k*-NN classifier with the number of neighbors *k* which resulted in the lowest classification error rate was selected for future evaluation in the testing set along with the LDA and RF predictors. The three classifiers were also trained and evaluated using this Monte Carlo cross-validation using (1) only the folate features, (2) all features, and (3) all features except the folate features, to examine if there were any noticeable differences in predictive ability, though classification error rates in this context are optimistic. The "optimistic" qualifier is used because these partitions all used the learning data (which were pre-processed together, therefore not independent from one another). Distributions of these optimistic cross-validated classification error rates were also visualized for all three classifiers and predictor sets using boxplots, to get a sense of how well they were performing in the easier learning set scenario. These three classifiers were then trained on the full learning set. These analyses were repeated on stratified datasets by sex, race/ethnicity, and age at diagnosis (<=2 years vs. >2 years), in case heterogeneities between the groups were masking the signal. The sex and race/ethnicity variables were dropped in their respective stratified analyses.

Testing Data

Once the candidate classifiers were selected using the learning set, the independent test set was used to acquire an unbiased estimate of the various predictors' classification error rates. The pre-processing pipeline of the testing data was identical to that of the learning data mentioned previously with the exception that any metabolites of interest were retained even if they failed one of the filtering steps. This was the case for 2 metabolites, one of which was a folate pathway metabolite, leaving 151 metabolites in the testing dataset for analysis. Following *k*-NN imputation with the number of neighbors set to 5, scone evaluation selected DESeq scaling adjusting for case-control status, batch, and the full QC matrix. This differed from the scheme selected for the learning data only in the scaling method used (DESeq as opposed to UQ). Using the normalized dataset, each of the classifiers trained on and selected from the learning set was applied to predict case-control status on the testing set and classification error rates were reported.

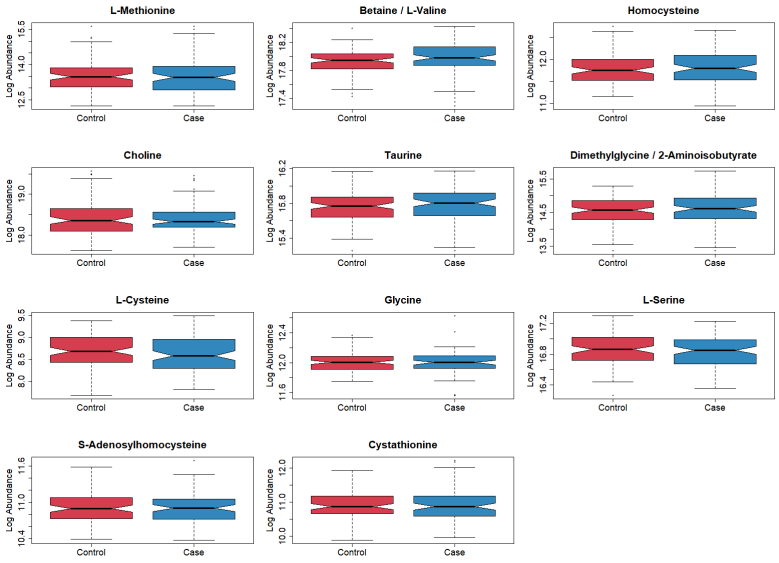
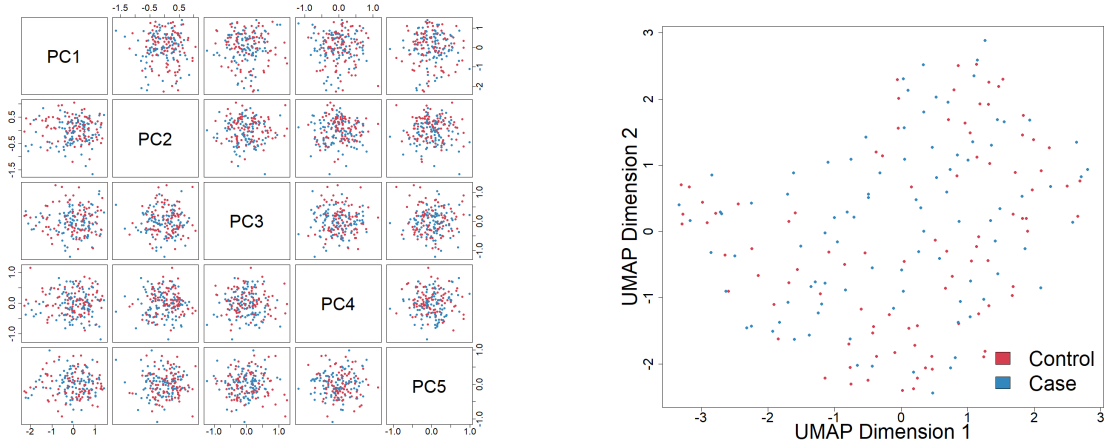


Figure 2.6: Boxplots of log abundances for each folate pathway metabolite, stratified by case-control status, following filtering and normalization. Blue = cases (n = 92) and red = controls (n = 92). Note the different y-axis scales.

2.4 Results

Table 2.2 shows demographic summary statistics for cases and controls separately. Visualizations of these variables stratified by case and control are also shown in Figures A.2 to A.8. There was evidence of a difference in the distributions of the timing of the newborn's blood spot collection after their birth (in hours) between cases and controls, which was a variable adjusted for during the normalization process. There was no correlation between blood spot age (defined as the number of years between the sample collection and processing at the laboratory) and folate pathway metabolite levels (Spearman correlation coefficients ranged from -0.04 to 0.06) (Table A.1). Boxplots of the log abundances of each folate pathway metabolite showed no noticeable differences in median values between cases and controls for any of the folate pathway metabolites (Figure 2.6).

Visualizations of the folate metabolite data using dimensionality reduction on all subjects are presented in Figure 2.7. Specifically, Figure 2.7a shows pairwise plots of the first five components from PCA and Figure 2.7b the two components from UMAP, with each point colored by case-control status. Figure 2.7 did not reveal any clustering by case-control status. In addition, the folate pathway metabolites themselves did not appear to be correlated, as indicated by the pseudo-color image of the Spearman correlation matrix, with rows/columns ordered by hierarchical clustering (Figure 2.8). When comparing the folate pathway features to the other metabolites present in the dataset, the folate pathway metabolites did not appear to cluster together (Figure A.9).



(a) (b)

Figure 2.7: a) Pairwise scatterplots of the first 5 components from PCA on only the 11 folate metabolites and all subjects, with plotting symbols for subjects colored by case-control status. b) Results of UMAP on only the 11 folate metabolites and all subjects plotted and colored by case-control status. Blue = cases (n = 92) and pink = controls (n = 92).

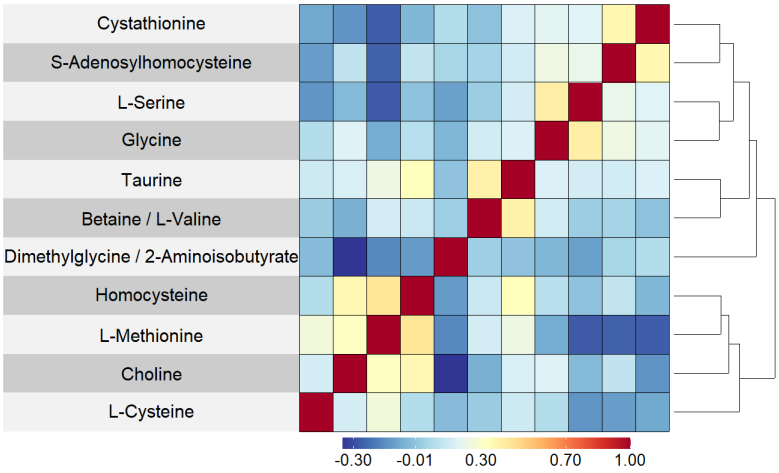


Figure 2.8: Pseudo-color image of Spearman correlation matrix for 11 folate pathway features, with rows and columns ordered by complete linkage hierarchical clustering. A darker red indicates a stronger positive correlation and a darker blue indicates a stronger negative correlation. As it is a symmetric matrix, the columns correspond to the same set and ordering of metabolites as the rows.

An example of results that were used to select the optimal k for the k-nearest neighbor classifier are shown in Figure A.1 for the unstratified analysis. Each boxplot shows the distribution of the Monte Carlo cross-validated error for a given number of neighbors k. In each instance, the k value that returned the lowest average cross-validated classification error (darkest red boxplot) was used (in this case $k = 10$). In general, there did not seem to be a large difference in the error distributions across k values.

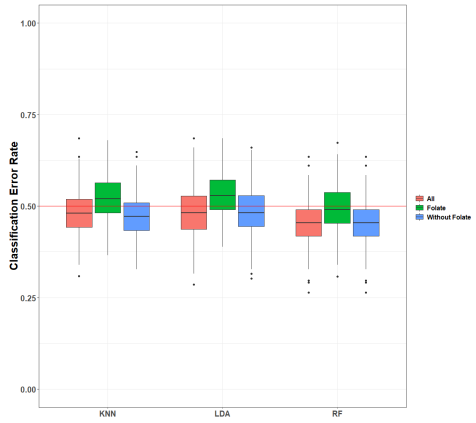
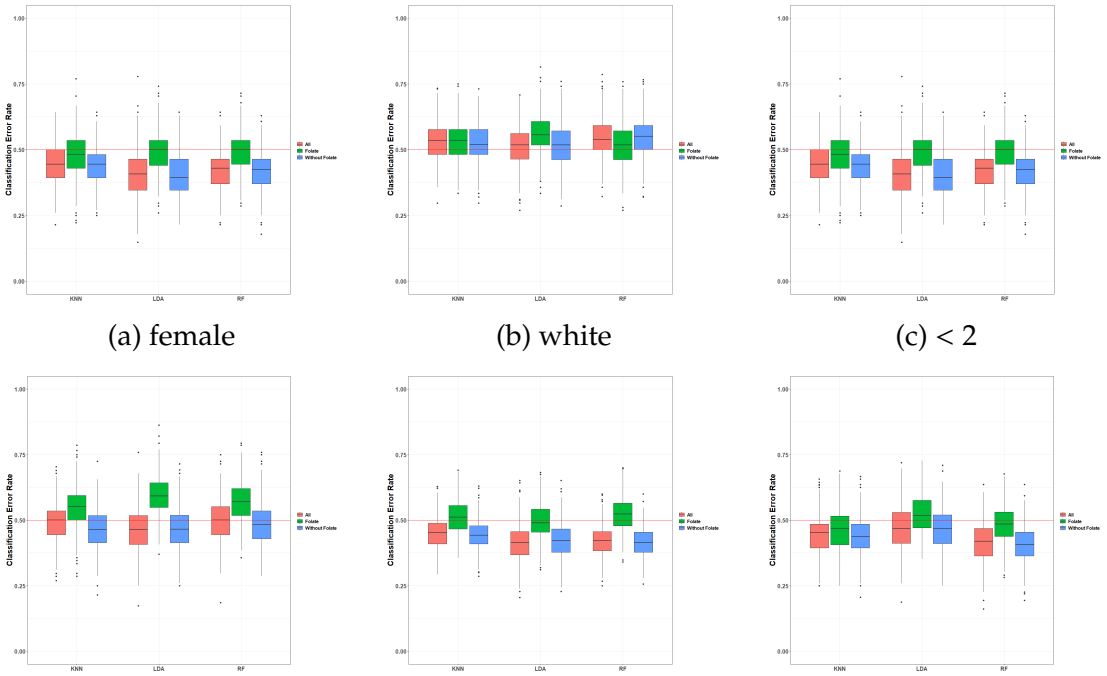


Figure 2.9: Boxplots showing optimistic classification error rates for three classification methods. Results from using all the covariates with: all metabolites (pink), only the folate metabolites (green), and all metabolites except the folate metabolites (blue). The red horizontal line corresponds to a classification error rate of 0.5.

Boxplots of the optimistic classification error rates on the Monte Carlo 60/40 split into training and validation sets are shown in Figure 2.9 for the full dataset and in Figure 2.10 for each of the six stratified analyses. Note that, since the partitions of the learning set (training and validation sets) were pre-processed and normalized together, and therefore are not independent datasets, these are optimistic classification error rates. In each boxplot, the classification error rate is shown on the y-axis with the three methods on the x-axis. Higher classification error rates indicate worse performance by the method, so ideally the boxplots would be close to zero. The horizontal red line indicates a classification error rate of 0.5, or the expected classification error rate if the outcome was predicted using a coin flip (based on chance alone). Within each panel, there are optimistic classification error rates shown for models fit using three subsets of the data: all of the features (pink), only the folate features (green), and all of the features except the folate ones (blue). In this way, we can not only see the performance of models using the folate features, but also if there seems to be a loss of predictive ability in the absence of the folate metabolites. In all cases, models fit using the folate features did not seem to be predictive of the outcome (green boxplots). In addition, the absence of the folate features did not seem to worsen prediction compared to including them (blue vs. pink boxplots). Lastly, Table 2.1 lists the classification error rates of the prediction algorithms on the hold-out testing set, which shows little to no evidence of predictive ability of the folate pathway metabolites for childhood ALL (a classification error rate of 0.5 corresponding to chance alone). Note that training and evaluating Random Forests on the mean-centered and standard deviation-scaled training and test data (the same input for LDA and k-NN) did not improve performance (data not shown).



(a) female (b) white (c) < 2

(d) male (e) latino (f) ≥ 2

Figure 2.10: Boxplots showing optimistic classification error rates for three classification methods on six subsets of the data based on sex, ethnicity, and age at diagnosis. Results from using all the covariates with: all metabolites (pink), only the folate metabolites (green), and all metabolites except the folate metabolites (blue). The red horizontal line corresponds to a classification error rate of 0.5.

2.5 Conclusions

Epidemiologic studies have reported a reduced risk of childhood leukemia following self-reported prenatal folate and vitamin intake from supplementation and diet [70, 96, 95, 2, 109], yet the mechanistic pathways and the windows of sensitivity during the prenatal period are not well understood. Our study characterized downstream one-carbon(folate)-related metabolites at birth and found no predictive value of 11 nutrients in the development of childhood ALL. Null findings were observed overall, as well as in stratified analyses by sex, race/ethnicity, and age at diagnosis. Our interest in examining these groups stemmed from the known higher incidence of childhood leukemia in boys and LatinX children [10, 60], and the difference in molecular characteristics by age at onset [48, 121]. Our observations are consistent with a previous report from our group using an independent study population of childhood ALL cases and controls showing no association between neonatal hemoglobin-normalized folate concentrations measured by the Lactobacillus casei microbiologic growth assay and risk of childhood ALL (n = 313)

[20]. To our knowledge, no other childhood leukemia studies have directly measured downstream folate pathway metabolites, neither at birth nor in cord blood or pregnancy blood samples. In a cohort study conducted in England [103], DNA methylation levels measured in cord blood samples were associated with self-reported folic acid supplementation during pregnancy. However, these folate-associated CpG sites did not overlap with known ALL-associated CpG sites, providing no support that folic acid acts via DNA methylation in preventing childhood leukemia [103]. Most other biomarker studies to date have examined the role of genetic variants in the folate pathway. Meta-analyses have consistently reported that the MTHFR C677T polymorphism, which regulates folate and homocysteine metabolisms, is associated with childhood ALL risk, however homocysteine as measured in our study was not predictive of childhood ALL; the evidence for other folate-related genes is less strong mostly due to lack of studies with sufficient sample sizes and validation in independent test sets [124, 54, 69]. Measuring the downstream metabolites in the folate pathway should account for inter-individual variations that may exist in terms of genetic susceptibility, methylation processes, and other lifestyle factors.

Our study focused on the neonatal period, as a surrogate of last weeks of the pregnancy, and negative findings may be explained by the fact that nearly all women in the US take prenatal vitamins by the end of their pregnancy [56], limiting our ability to find case-control differences in newborn levels of one-carbon nutrients; studies implemented in other countries where prenatal folate/vitamin supplementation is less complete than in the US could be informative. Also, there is a chance that the null findings could be attributed to our relatively small sample size. Alternatively, the timing of a potential impact of folate-related nutrients on blood cell development and epigenome at large in the fetus may occur in early pregnancy as shown in studies of nutritionally-deprived women during early pregnancy [12, 83], although no studies to date has directly tested this yet for childhood leukemia. Our group has recently published results on the relationship of DNA methylation patterns associated with peri-conception folate intake in children with ALL vs. healthy controls [77]. This epigenome-wide association analysis revealed a differentially methylated region (DMR) in the promoter region of DUSP22, a protein phosphatase and known tumor suppressor, in response to total and food preconceptional folate intake. While the direction of effect was the same between cases and controls, the strength of association was greater for cases. Additional relationships between dietary periconceptional folate and DNA methylation by case status were identified via differentially methylated probes in genes CERK, CUTA, WDFY4, SART1, TCF20, and ERGIC1. These results lend support for a role of folate as a drive of early pregnancy changes in DNA methylation that affect future ALL development.

Our case-control study has several strengths. Childhood leukemia is a rare disease, and investigating its etiology is hampered with several methodological challenges. Here, we assembled 122 cases and 122 controls with unique access to blood specimen collected at birth to better characterize the perinatal window of susceptibility. The use of these archived neonatal blood specimens allowed the examination of events before the leukemia occurred, thus preserving temporality in investigating causation. The cases and controls

	Total	Female	Male	≤ 2 years old at ddx	> 2 years old at ddx	White	Latino
	$n = 54$	$n = 17$	$n = 37$	$n = 25$	$n = 29$	$n = 17$	$n = 28$
k-NN	0.5185	0.5882	0.4595	0.56	0.4482	0.4706	0.3929
LDA	0.5556	0.5882	0.4595	0.4	0.5172	0.5882	0.3571
RF	0.537	0.5882	0.4595	0.56	0.5517	0.5294	0.3929

Table 2.1: Classification error rates for each candidate class predictor after being trained on the learning set, and then evaluated on the hold-out test set.

Abbreviations: k-NN = k-nearest neighbors; LDA = Linear discriminate analysis; RF= Random Forests, ddx = diagnosis

in this study were age-matched by year and month of birth to reduce any potential bias in the results due to storage conditions. Nevertheless, all metabolites were adjusted for 'blood spot age' as well as 'hemoglobin' which also accounts for storage aging factors [133]. Lastly, we confirmed that storage conditions likely did not bias our results since blood spot age was not correlated with the levels of metabolites in cases and controls. We developed and validated laboratory methods to measure exposomic features with extremely small amounts of blood from the neonatal spots. Our statistical approach focused on predictive ability of the set of folate-related metabolites rather than hypothesis testing for individual metabolites and assessed the performance of three distinct classes of predictors using sample-splitting. Our data are derived from a registry-based study with no contact with participants, therefore reducing the potential for differential participation between cases and controls. The subset included in this study represented socio-demographic and reproductive characteristics similar to the larger source California registry-based study [110]. Folate deficiency can cause chromosome breakage [13], and DNA-damaging agents have been associated with childhood leukemia harboring chromosome structural abnormalities (such as KMT2A gene fusion in infant leukemia, and ETV6-RUNX1 translocation) [121, 5, 72]. However, information on molecular subtypes of leukemia was not available in this series derived from a cancer registry database. Additionally, the type of specimen available at birth (i.e., archived dried blood spot) does not allow for measuring metabolites separately in different cellular blood components (like lymphocytes, monocytes, etc.). The statistical methods used did not utilize any information about the cyclic structure of the folate pathway; future development of statistical methods to address this issue could improve predictive performance. The analyses were limited to only 11 folate pathway metabolites with semi-quantitative measures.

In our study, the 11 analytes involved in the one-carbon (folate) metabolism pathway that were measured in newborn blood specimens did not appear to predict subsequent ALL in children. These negative findings are reflective of the last weeks of pregnancy and alteration of the DNA methylation under the influence of various nutrients, including folic

acid and their metabolites during early gestation could be a more important factor. They do not rule out the possible impact of these nutrients at the time of conception or during the first trimester of pregnancy that are critical times for the embryo's DNA methylation programming.

2.6 Discussion

While our study findings did not support the hypothesis that folate pathway disregulation was associated with increased risk of pediatric ALL, at least at the timepoint measured, it does offer an example of applying a prediction framework to a targeted metabolomics dataset, rather than a hypothesis testing one. In this way, there is flexibility in what models can be used to detect various complex relationships, and certainly more advanced modeling and machine learning algorithms could be employed in a scenario where there seems to be any signal in the data. Importantly, we also reinforce and give an example of how to appropriately maintain an independent test set to acquire unbiased estimates of prediction performance.

Additionally, we adapted the filtering, imputation, and normalization pipeline to a targeted metabolomics dataset and showed the importance of appropriate normalization to remove unwanted technical variation. In our case, the data were limited by only having a subset of the relevant metabolites involved in the pathway and perhaps having a more complete picture would provide additional insight. Biochemical pathways such as these, however, are quite complex and, as previously mentioned, the cyclic graphical structure made detection of an association more difficult. Development of methods that can incorporate this structure may also aid in predictive ability.

That being said, preliminary exploratory data analysis already suggested the potential lack of signal. This was reinforced during model building, as even the optimistic error rates across cross-validated splits were not promising. This may also explain the apparent robustness in error rates across various values of k using the kNN model, where one would normally expect to see a variation in performance based on whether the defined neighborhood was too focused. If there is no signal, then trying to optimize the scope of the neighborhood provides little benefit.

Better understanding the causes of pediatric ALL and other complex diseases requires examining the full set of potential risk factors, as genetics can only explain so much. Using targeted metabolomics data allows for examining a hypothesis related to a particular pathway, and can provide a focused look at a potential risk factor. To generate a novel set of hypotheses, one can similarly use metabolomics or other exposomic analyses in an untargeted capacity. The following chapter will examine one such analysis and the necessary considerations along with insights it provides.

Covariate		Combined	Cases	Controls	*p*-values*
	n	244	122	122	
Infant ethnicity	Hispanic	138	67	71	0.422
	Non-Hispanic White	73	35	38	0.422
	Other	33	20	13	
Sex of child	Male	138	67	71	1
	Female	106	53	53	
Child's age at leukemia diagnosis (count)	1 yr		17		N/A
	2 yr		34		
	3 yr		32		
	4 yr	N/A	17	N/A	
	5 yr		9		
	6 yr		10		
	7 yr		3		
Child birth year (year)	Min	2000	2000	2000	1
	Mean	2004	2004	2004	
	Median	2004	2004	2004	
	Max	2008	2008	2008	
Time of heal prick after child birth (hours)	Min	0	0	12	0.049
	Mean	33	35	31	
	Median	28	29	27	
	Max	131	131	120	
Child birth weight (grams)	Min	1899	2165	1899	0.571
	Mean	3376	3369	3384	
	Median	3380	3345	3407	
	Max	131	131	120	
Gestational age at birth (weeks)	Min	32	32	34	0.248
	Mean	39	39	39	
	Median	39	39	39	
	Max	44	43	44	
	N/A	8	3	5	
Large for gestational age (LGA)	$< 90^{th}$ percentile	188	93	95	0.675
	$\geq 90^{th}$ percentile	48	26	22	
	N/A	8	3	5	
Gestational age at maternal collection (weeks)$^{+}$	Min	12	14	12	0.897
	Mean	17	17	17	
	Median	17	17	17	
	Max	39	39	39	
	N/A	9	5	4	

continued . . .

. . . continued

Covariate		Combined	Cases	Controls	*p*-values*
	n	244	122	122	
Maternal age at collection (years)$^{+}$	Min	15	17	15	0.167
	Mean	28	28	28	
	Median	29	29	28	
	Max	39	39	39	
Mode of delivery	Vaginal	163	75	88	0.103
	Caesarean	81	47	34	
Blood spot age at time of laboratory analysis (years)	Min	11	11	11	1
	Mean	15.33	15.33	15.33	
	Median	15	15	15	
	Max	19	19	19	
Maternal highest level of education	8^{th} grade or less	24	9	15	0.465
	9^{th}-12^{th} grade	111	55	56	
	Higher than 12^{th} grade	103	53	50	

N/A, not available

*The *p*-value corresponds to a test for differences in covariate distributions between case and control groups. For binary variables, Fisher's exact test was used to test for differences between groups; for categorical variables, a Chi-squared test was used; and for continuous variables, a Wilcoxon rank-sum test was used.

$^{+}$ Covariates only relevant to mother-child paired analyses in Chapter 3.

Table 2.2: Descriptive characteristics of the pregnancy and early-life demographics of the leukemia cases and controls (n = 244) in Chapter 2 and mother-child leukemia pairs used in Chapter 3 (n = 244 pairs).

2.7 Notation

We present notation consistent with that presented in the previous chapter that will be used throughout the rest of this dissertation and can be used as a reference.

We use n to denote the number of observations (i.e., subjects) in our data and J to denote the total number of observed exposomic features. We then denote an $n \times J$ matrix X for the observed J feature abundances in our n samples. X_j refers to the n-vector of measured abundances for the j^{th} feature, while $X_{i,j}$ refers to the observation of the j^{th} feature for the i^{th} sample. In our applications, features are continuous. W refers to the $n \times w$ matrix of covariates which can be indexed by W_{w^*} for a given covariate w^*. These covariates can be continuous, categorical (ordinal or nominal), as well as binary. The outcome vector Y is of length n and in our application is binary $\in \{0, 1\}$, where 1 refers to a case and 0 to a control (as our applications are in the context of case-control studies).

2.8 Abbreviations

List of abbreviations utilized throughout the dissertation:

- ALL: Acute lymphoblastic leukemia
- BMI: Body mass index
- DBS: Dried blood spot
- ddx: Diagnosis
- DEE: Diesel engine exhaust
- DGP: Data generating process
- EDA: Exploratory data analysis
- ETS: Environmental tobacco smoke
- HC: Histologically confirmed
- HILIC: Hydrophilic interaction liquid chromatography
- HSA: Human serum albumin
- ICC: Intraclass correlation coefficient
- ICD-O: International classification of disease for oncology
- *k*-NN: *k*-nearest neighbor
- LASSO: Least Absolute Shrinkage Operator
- LC-HRMS: Liquid chromatography high resolution mass spectrometer
- LC-MS: Liquid chromatography mass spectrometry
- LDA: Linear Discriminant Analysis
- LGA: Large for gestational age
- LUAD: Lung adenocarcenoma
- MS: Mass spectrometry
- *m/z*: Mass to charge ratio
- OCM: One-carbon metabolism

- PMD: Paired mass difference
- PAH: Polycyclic aromatic hydrocarbons
- PCA: Principle components analysis
- QC: Quality control
- qPMD: Quantitative paired mass difference
- RF: Random Forests
- ROS: Reactive oxygen species
- RSD: Relative Standard Deviation
- sn: Skew-normal
- SWHS: Shanghai Women's Health Study
- TMM: Trimmed mean of M values
- ttd: Time to diagnosis
- UMAP: Uniform manifold projections
- UQ: Upper quartile

Chapter 3

Untargeted Reactomics: Feature Selection for Paired Mother-Child Data in Childhood Leukemia Case-Control Study

3.1 Introduction

In this chapter, we continue our discussion of the metabolome in the context of pediatric acute lymphoblastic leukemia (ALL), however now mainly in the untargeted metabolomics setting. Our analysis in Chapter 2 found little evidence supporting the hypothesis that disregulation of the folate pathway is a risk factor for pediatric ALL. It would therefore seem beneficial to broaden the set of candidate exposures/risk factors under scrutiny. Untargeted metabolomic studies are able to detect a large range and number of metabolites agnostically, resulting in a dataset that reflects the breadth of molecules present. Metabolites can then be selected by strength of association with the disease outcome to discover novel biomarkers and potential exposures that may shed light on risk factors for the disease of interest. This is important in informing more targeted future directions of study, and therefore, primarily serves in hypothesis generation. Note that this breadth comes at the cost of no longer having benchmarked standards making any annotations putative.

In addition to difficulties in downstream interpretation arising from uncertainty in metabolite identification, untargeted metabolomics data can suffer from a large amount of technical noise which will mask the signal of interest, even more so than targeted metabolomics data. As the net cast is much wider, resulting datasets have on the order of thousands of metabolites, a large number of which are uninformative due to the nature of peak-picking in the data acquisition process. Removing these "bad" peaks while maintaining high quality ones is an essential step in data cleaning [92]. The data are also

subject to the same technical effects discussed in the context of targeted metabolomics in Chapter 2, related to sample collection, storage, or processing, and measures to remove their effects are paramount to the resulting quality of signal in the data [93].

Previous analyses often employed traditional multiple testing approaches to select the candidate set of associated features. This was inherited largely from genomic analyses. A key difference between data acquired from the two technologies, however, is that genomic allele data are discrete while metabolomic peak abundances are continuous, and have a larger amount of inherent error from various sources [92]. Additionally, with the large number of metabolite features detected in untargeted metabolomic datasets, multiple testing approaches will impose harsh p-value cutoffs which make selection of "significant" features unnecessarily difficult, since the goal is hypothesis generation rather than testing to begin with [86]. As such, we will utilize an ensemble feature selection approach previously described [62, 86, 128]. This allows for the selection of the top features associated with the outcome of interest according to two parametric regression methods (linear and penalized logistic regression) and aggregated decision trees (Random Forests). In this way, there is more flexibility in the types of associations detected and no need to select arbitrary p-value cutoffs.

This, however, still does not speak to the limitation of solely having putative annotations. To address this, we instead transform the metabolite features into reactomic features on a higher enzymatic level, as discussed by Yu et al., and perform feature selection on this new reactomics dataset instead [133]. This grants us the ability to look at the data on a reaction level without needing to annotate the intermediate metabolite features. This transformation still relies on peak abundances from the original metabolomics data, however. So all pre-processing and normalization considerations discussed previously are still relevant.

For a complex disease such as pediatric ALL, analyzing both newborn and maternal data from pregnancy may be paramount to detecting relevant early-stage risk factors. Previous work has shown that neonatal metabolomic profiles mirror those of the mother during late pregnancy [84] and there is also evidence for vertical transmission of the metabolome [80, 65]. Recognizing the potential importance of the information contained within the maternal metabolome, increasingly researchers are using maternal samples from during pregnancy to try to better understand early exposomic risk factors for various diseases and conditions in their offspring [45, 129]. In our case, the child data is from DBS samples mentioned previously (Chapter 2), while the maternal data is from serum samples collected during the first trimester of pregnancy. Integrating these two datasets together requires data-adaptive decisions during the pre-processing stage as well as careful consideration of their differences upon downstream analysis. By doing this, we can apply feature selection methodology to select features across both sample types and time points that seem to show the strongest evidence of an association with pediatric ALL, getting a more comprehensive view than analyzing either dataset separately. To our knowledge this is the first mother-child paired dataset for which feature selection has been done for reactomics transformed untargeted metabolomics data.

In addition to the hypothesis generation framework for our untargeted reactomics transformed datasets, we also have a small targeted analysis following up on results from a previous study. Through feature selection of untargeted metabolomics data for pediatric ALL cases and controls, our team previously found two metabolites, linoleic and linolenic acid, which were highly ranked in their association with disease status for late diagnosis children [62]. As such, a second aim of our study is to use a hypothesis-driven approach to determine whether the previously reported trend is reproducible.

3.2 Materials and Methods

Study Population

As in Chapter 2, our population of interest is children diagnosed with pediatric ALL before turning 15 years old and our sample of children is the same as described in Chapter 2 as well. As a reminder, children diagnosed with ALL at age 0–14 years (International Classification of Disease for Oncology (ICD-O) codes 9820, 9823, 9826, 9827, 9831–9837, 9940, 9948) and 1-to-1 controls matched on sex, year and month age of birth, and race/ethnicity were identified through a California registry-based study linking electronic records from the cancer registry (1988–2011) with birth data from vital statistics and archived biospecimens from the Biobank Program at the California Department of Public Health. We included 122 cases and 122 controls with available neonatal dried blood specimens (DBS) and maternal pregnancy serum collected as part of a newborn and prenatal screening programs in California [52]. Five 14-mm diameter blood spot specimens were collected from newborns on S&S filter paper (also known as Guthrie cards) by heel-stick after 12 hours and usually no later than 6 days of age (median age at collection: 29 hours). Venous blood was collected between 15 and 20 weeks of gestation, and leftover maternal specimens (1–2 mL) were refrigerated for 1–2 days and then processed for long-term storage. The current analysis includes a total of 488 individual samples for 244 mother-child pairs (122 case pairs and 122 control pairs), including leukemia status and potential confounding variables (Table 2.2).

Metabolite Extraction and Liquid Chromatography-High Resolution Mass Spectrometry (LC-HRMS) Analysis

Neonatal DBS were included for the present analysis using a methodology developed by our team [87, 62]. In short, 4.7-mm diameter DBS punches were extracted with 100 μL of water at room temperature (15 min, 1400 rpm), and a 5 μL aliquot was reserved for hemoglobin measurements to adjust for original blood volume [134]. 400 μL of acetonitrile containing isotopically labeled internal standards (IS) was added to the remaining aqueous solution containing the DBS punch, samples were agitated (1400 rpm, 37°C, 1 h), and protein was precipitated at −20°C for 30 min. The supernatant was evaporated to dryness

and stored at −80°C until analysis. Stored maternal serum samples were thawed on ice, and 50 μL aliquots were combined with 150 μL of methanol containing IS. The sample was vortexed, then incubated at −20°C for 30 min., and evaporated to dryness and stored at −80°C until analysis. Immediately prior to analysis, samples were reconstituted and analyzed with liquid chromatography-high resolution mass spectrometry (LC-HRMS) with reverse-phase chromatography in negative ionization mode. Additional instrument parameters are provided in the Supplementary Materials (Section B.1). DBS and serum samples were analyzed separately. Pooled quality control (QC) samples were prepared by combining aliquots of the sample extracts (i.e., a QC pool of DBS extracts and a separate QC pool of serum extracts) that were injected routinely throughout the run. These were used to monitor instrument stability and facilitate batch and run order correction.

Data Processing for Linolenic Acid and Linoleic Acid

Semi-quantitative measures for alpha-linolenic acid/gamma-linolenic acid (co-eluting) and linoleic acid were extracted from the DBS using Profinder software (Agilent Technologies, Santa Clara, USA) considering retention time, accurate mass, isotope distribution, and MS/MS fragmentation pattern matching with reference analyzed under the same conditions (matching retention time < 5 sec and ppm < 5).

Processing for Untargeted Data

Data extraction and pre-processing of the two datasets for neonatal and maternal specimens follows methods previously described. Raw data were converted into mzxml format and analyzed using the R programming platform (version 4.0.4) [1]. The Xcms package was used to extract peaks, align the peaks, group peaks into features and fill the baseline for grouped peaks as a final peak table using optimized parameters by IPO package [97, 59].

Data

The data structure generally resembles that from Chapter 2 (with two sets of metabolites measured for each subject, one from the DBS samples and one from the maternal serum) and is summarized in Table 3.1. For each of our n child cases or controls, we have an $n \times J_b$ matrix X_b which refers to metabolite abundances from the DBS samples for the child subjects as well as an $n \times J_m$ matrix X_m which refers to the metabolite abundances from the maternal serum samples. The DBS samples were analyzed in two sequential run batches while the maternal serum samples were analyzed in four. For each of these metabolite abundance matrices, we then have corresponding pooled QC samples that we name $X_{b,p}$ and $X_{m,p}$, with dimensions $30 \times J_b$ (15 pooled samples per run batch) and $34 \times J_m$ (nine pooled samples for run batches 1–3 and seven for run batch four), respectively. For each run batch, we have three blank samples which result from processing samples without

Neonatal DBS		Maternal Serum	
Data Object	Description	Data Object	Description
X_b	$n \times J_b$ matrix of DBS metabolite abundances	X_m	$n \times J_m$ matrix of serum metabolite abundances
$X_{b,p}$	$30 \times J_b$ matrix of pooled QC DBS metabolite abundances	$X_{m,p}$	$34 \times J_m$ matrix of pooled QC serum metabolite abundances
batch	binary run batch	$batch_r$ $batch_s$ *batch*	run batch with four levels binary storage location batch 4 $batch_r$× 2 $batch_s$ = 8 batches
QC_b	$n \times 4$ matrix of of quality control metrics	QC_m	$n \times 4$ matrix of quality control metrics
W	$n \times 9$ matrix of baseline covariates	W	$n \times 9$ matrix of baseline covariates
Y	n length binary vector indicating case-control status	Y	n length binary vector indicating case-control status

Table 3.1: Description of untargeted metabolomics and associated data for neonatal DBS and maternal serum samples.

any biological specimen added (to pick up background). In addition, for the biological serum samples, we have a binary storage location batch variable that indicates the storage site for the sample. As such, we effectively have eight batches for the serum data (4 run × 2 storage), which we refer to as the batches moving forward.

Figure 3.1 shows the first two components of PCA on the maternal serum data including all samples (biological, pooled QC, and blank). We can see that the blank samples cluster tightly away from the rest of the samples and the pooled QC samples cluster tightly between the two groups of biological samples. The biological samples, however, show a large split into two groups based on the storage batch, showing the importance of including the site information for the serum samples as it clearly creates the largest artifact in the biological samples (more so than the run batches).

For the DBS samples, we have the same QC_b matrix as in Chapter 2 with measurements for *run order*, *DBS spot age*, *age at collection*, and *hemoglobin*. For the serum samples, we similarly have a QC_m matrix that contains the continuous variables *run order*, *serum sample age*, and *gestational age at collection*. We have our matrix of baseline covariates for mother-child pairs W which includes *sex of the child* (binary), *infant ethnicity* (categorical), *child birth year* (continuous), *child birth weight* (continuous), *gestational age at birth* (continuous), *Large for gestational age (LGA)* (binary), *maternal age at collection* (continuous), *mode of delivery* (binary), and *maternal highest level of education* (categorical ordinal).

Lastly, we have our binary outcome vector Y of length n which indicates the case-

control status of the child. We assume that we have n identically distributed random draws of $(X_{b,i}, X_{m,i}, W_i, Y_i) \sim P_0 \in \mathcal{M}$.

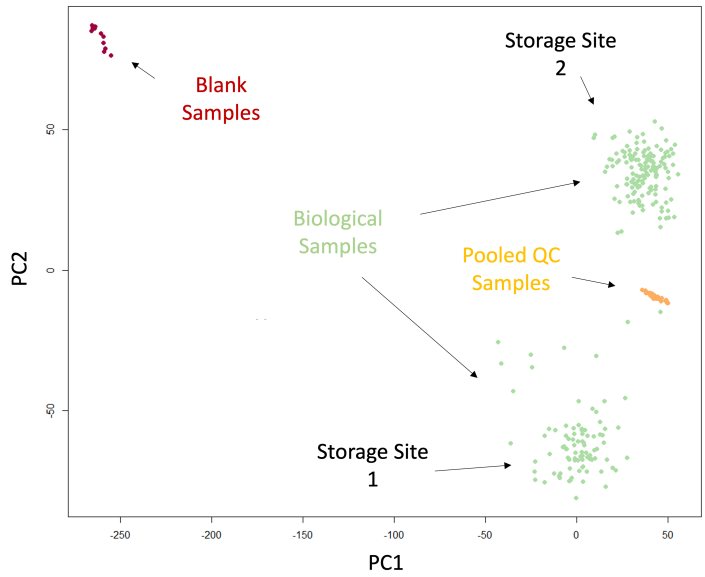
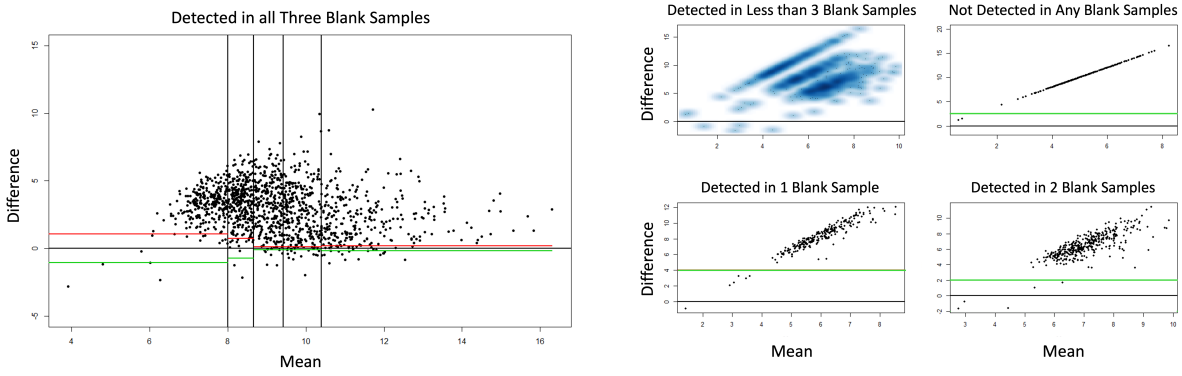


Figure 3.1: Scatterplot of first two principal components of the serum samples (biological=green, pooled QC=orange, blank matrix=red).

Data Pre-processing

As mentioned previously, careful consideration in removing unwanted technical effects is necessary to allow for downstream detection of the biological signal of interest. Pre-processing was performed separately for the DBS data and the serum data. For the 244 neonatal DBS data, 2142 metabolite features were measured (J_b = 2142), while 3802 were measured in the maternal serum samples (J_m = 3802). Metabolite features detected above the noise (determined by mean-difference (MD) plots by batch) were retained adapting the filtering scheme from Schiffman et al. with some modifications [93]. Examples for one batch are shown for the DBS and serum data in Figures 3.2 and 3.3, respectively. In each case, panel a) shows the MD-plot for metabolites detected in all three blank samples. Panel b) shows the MD-plot for the rest of the metabolites in the upper-left panel, and then zoomed in views of the three data clouds (based on how many blank samples a metabolite was detected in) in the remaining three panels. In general, the data clouds tended to be much higher (higher difference values) than those used in the examples shown by Schiffman et al. with much fewer points below the zero-difference line [93]. This was particularly true for the serum samples. As we did not have any benchmarking

regarding the true labeling of "good" or "bad" peaks, we opted for more conservative cutoffs (to not unintentionally remove too many "good" features). In the case of the DBS data (Figure 3.2), the absolute value of the quintile-specific 25 percentile below the zero-difference line was used as a cutoff (shown by the red line in Figure 3.2a). The other cutoffs for features detected in fewer than all three blank samples (shown in green in Figure 3.2b) were drawn at reasonable separation points between the lower data points and the majority of the data cloud. Similar logic was applied to the serum samples (Figure 3.3). As the differences tended to be higher, however, a global cutoff of the absolute value of the 25^{th} percentile below the zero-difference line was used for features detected in all three blank samples (Figure 3.3a, green line). This left 1930 and 3576 metabolites in the DBS and serum datasets, respectively.

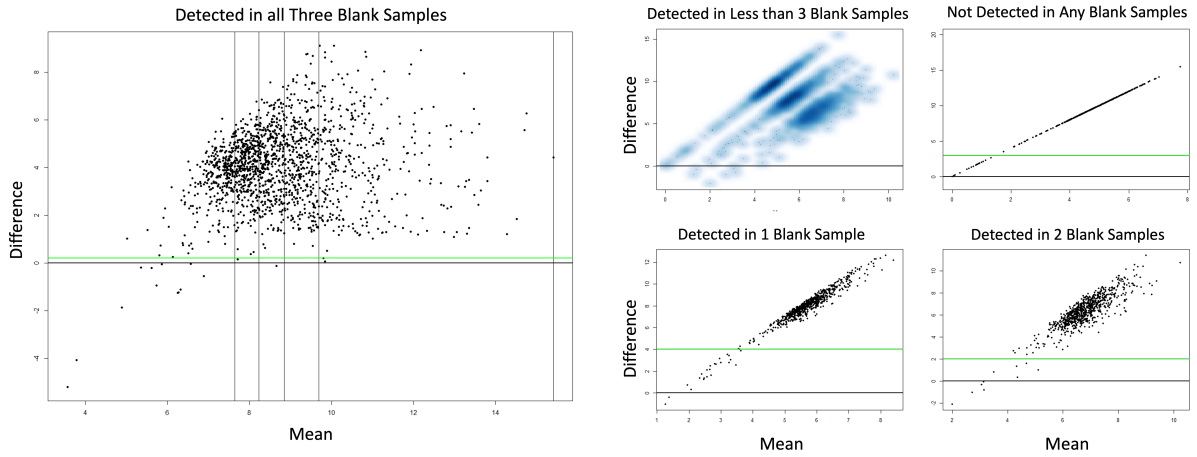


(a) (b)

Figure 3.2: Example MD-plots for one batch in DBS data. Panel a) shows metabolites found in all three blank samples. Quintile-specific cutoffs are shown with red line. Panel b) shows the MD-plot for features detected in fewer than three blank samples in the upper-left panel. The remaining three panels show close-ups of each of the three data clouds based on the number of blank samples the metabolite was detected. Green line indicates cutoff used in each of the three cases.

Next, a random effects model (Equation 2.1) was fit to estimate ICC values by batch, similar to Chapter 2, for the DBS and serum features separately. As a reminder, only the pooled QC samples had replicate injections. Figures 3.4a and 3.4b show density plots of the estimated ICC values for DBS and serum metabolites, respectively. The red line indicates the cutoff used to remove features with a low ICC value (0.75 and 0.7 for DBS and serum, respectively) [93]. After ICC filtering, there were 1481 metabolites left in the DBS dataset and 2460 left in the serum data.

The final filtering step was by percentage of missing values by feature. Both datasets had relatively low levels of missingness and using a $< 30\%$ missing values threshold only removed two features from the DBS data and one from the serum data (none of



(a) (b)

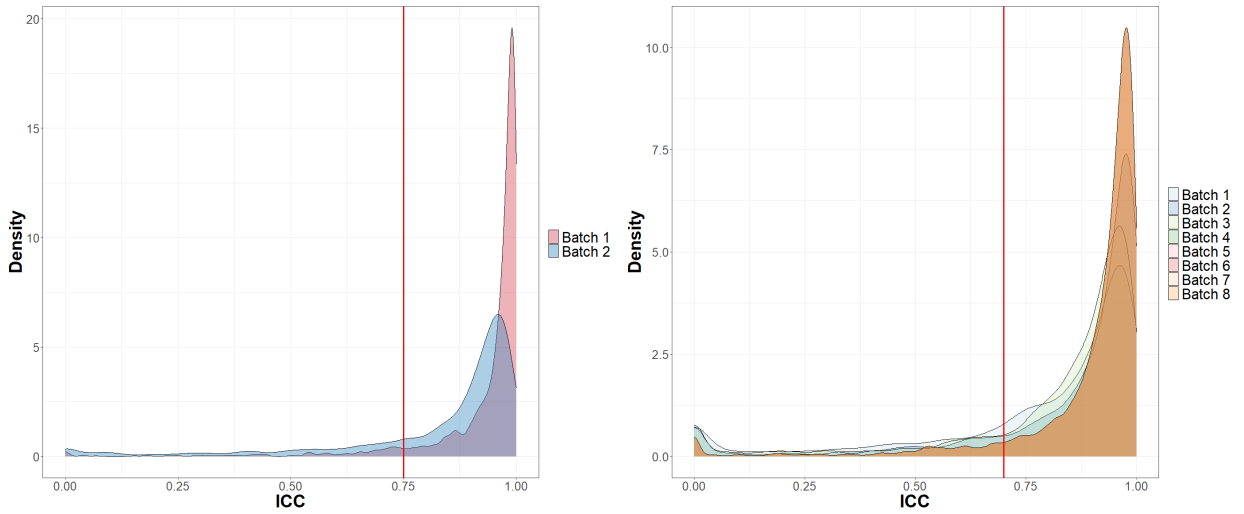
Figure 3.3: Example MD-plots for one batch in maternal serum data. Panel a) shows metabolites found in all three blank samples. Cutoffs are shown with green line. Panel b) shows the MD-plot for features detected in fewer than three blank samples in the upper-left panel. The remaining three panels show close-ups of each of the three data clouds based on the number of blank samples the metabolite was detected. Green line indicates cutoff used in each of the three cases.

which showed evidence of being differentially missing by case-control status according to Fisher's exact test). The final datasets used in downstream analyses consisted of 1479 metabolites for the DBS data and 2459 for the maternal serum data (300 of which appeared in both datasets).

k-nearest-neighbor was used for imputation (number of neighbors set to four) [105]. Normalization was selected by scone for both datasets [21]. For each dataset, scone was provided the matrix of imputed feature abundances, the corresponding *QC* matrix, batch information, case-control status, and a set of empirically derived negative controls that were consistent across all batches for *RUV* analysis [92]. For the DBS data, scone included variables for case-control, batch, spot age, age at collection, run order, hemoglobin, and TMM scaling in the normalization scheme. For the maternal serum data, it was normalized using identity scaling (no scaling) and the variables: case-control, batch, and the first three factors of unwanted variation using *RUV* analysis.

Reactomics

As mentioned, one goal with this analysis was to transform the metabolomic data to a new set of features that would speak to higher-order enzyme activity, thereby hopefully creating a new set of more robust features for hypothesis generation that do not rely on



(a) DBS (b) Serum

Figure 3.4: Densities of ICC values estimated by batch from the random-effects model for (a) DBS and (b) maternal serum data.

imprecise putative annotations. The reactomics algorithm is described by Yu et al. and is summarized briefly below [133]. Reactomics data analysis was performed separately on the normalized untargeted metabolomics datasets for neonates and mothers. The GlobalStd algorithm was used to remove redundant peaks such as isotopologues, adducts, as well as multi-chargers [132]. Paired mass distances (PMDs) of the remaining peaks were computed (by taking the paired differences of the *m/z* values) and only the PMDs in the KEGG reaction database were retained, creating the new list of candidate features [133]. Quantitative PMDs (qPMDs), which represent reaction-level changes across samples instead of individual metabolite-level changes, were calculated as follows (i.e., static PMDs) [133]. Specifically, for each PMD the relative standard deviation (RSD) of the ratio of abundances for the pair of metabolites across samples was calculated and any with RSD $\geq$ 30% were removed. For the PMDs that remained, the abundances of the pair of metabolites whose *m/z* was used to create the PMD were summed to create the observed measurement for the qPMD feature. The motivation of the consistent proportionality requirement being that it shows relative stability in the ratio of substrate and outcome, hence indicating the pair are regulated by enzymatic activity. We will use X_b^R and X_m^R to refer to the matrix of qPMD abundances for the DBS and serum data, respectively, where X_b^R has dimension $n \times J_b^R$ and X_m^R has dimension $n \times J_m^R$, where 56 features appear in both datasets.

Figure 3.5 illustrates the importance of appropriate upstream pre-processing and normalization in the resulting quality of the qPMD datasets. Figure 3.5a shows pairwise scatterplots of the first five principal components of the DBS reactomics data colored by



(a) Without pre-processing and normalization (b) With pre-processing and normalization

Figure 3.5: Pairwise scatterplots of the first five components from PCA on static DBS data (both cases and controls) (a) without and with (b) pre-processing and normalization discussed in Section 3.2. Points colored according to spot age of DBS sample.

.

spot age without any of the pre-processing and normalization described in Section 3.2. We can see there is a clear trend captured in the first two principal components (looks like a rainbow). Upon pre-processing and normalization, that effect is no longer visible (Figure 3.5b). This highlights the importance of taking appropriate measures in the removal of unwanted technical variation in the upstream metabolomic data, as the reactomics algorithm transforms metabolite peak abundance information, and will therefore carry forward any technical effects that plague it.

Figure 3.6 shows boxplots of log abundances of qPMDs for the DBS and maternal serum reactomics data side by side (DBS in red and serum in purple). The qPMDs within each set are ordered by PMD (shown on the x-axis). The most noticeable difference is in the range of PMD values. The observed DBS PMDs range until about 200 while the maternal serum PMDs extend all the way to around 470. This may reflect a true difference in the range of reactions present in maternal serum versus neonatal DBS, or is the result of the larger set of candidate metabolites used in the maternal data to generate the corresponding qPMD matrices.

In addition, when examining the two static qPMD datasets (X_b^R and X_m^R), we notice that the means and variances for the J_m^R qPMDs in X_m^R tend to be smaller than the J_b^R qPMDs in X_b^R (Figures 3.7a and 3.7b). The relationship between means and variances for the qPMDs in either dataset is also visualized in Figure 3.7c, where ellipses were drawn for each group (DBS = purple, maternal serum = green) assuming a multivariate t-distribution. This

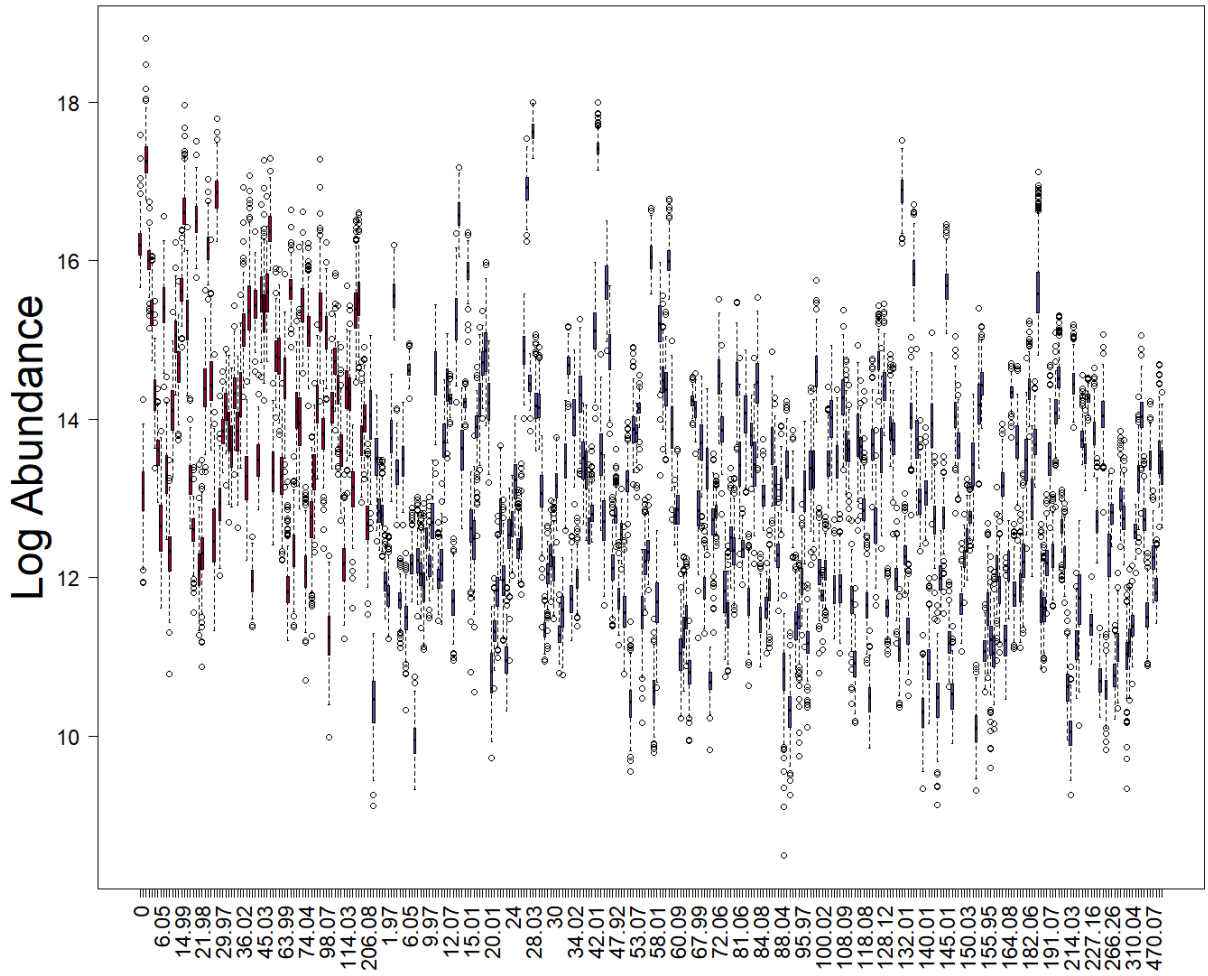
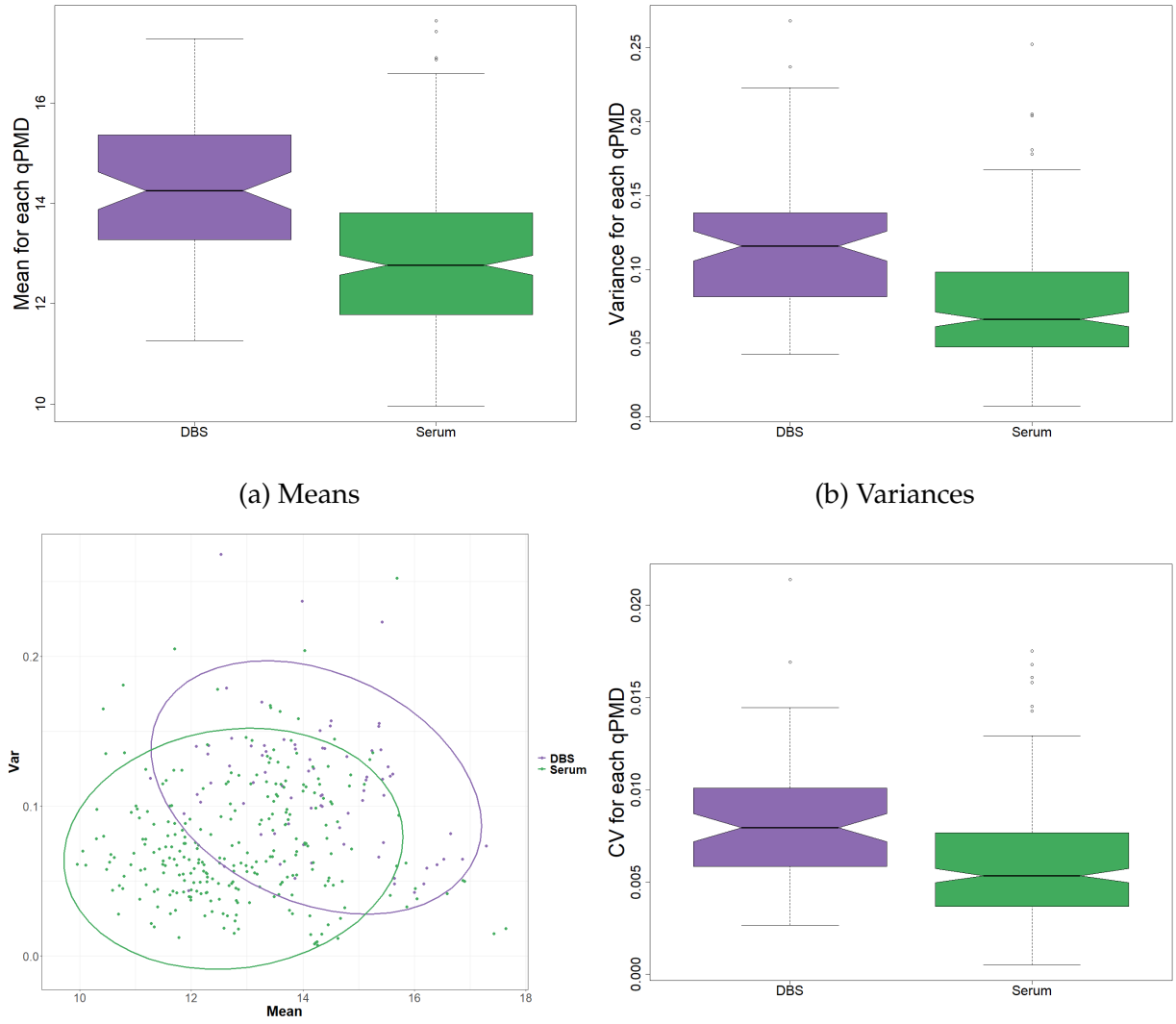


Figure 3.6: Boxplots of log abundances of qPMDs from DBS reactomics data (red) and maternal serum reactomics data (purple). X-axis shows PMD values ordered within each dataset.

trend still persists albeit slightly dampened when incorporating mean abundance values by looking at distributions of coefficients of variation for the qPMD sets in Figure 3.7d, where the variance is scaled by the mean value. This can arise both due to the different sample sources (DBS versus serum) as well as differences in upstream pre-processing, especially the differences in scalings, and will be kept under consideration as we proceed with analysis.



(a) Means

(b) Variances

(c) Mean vs. Variance

(d) Coefficients of Variation

Figure 3.7: Distributions of (a) means, (b) variances, and (d) CVs for DBS and maternal serum qPMDs. (c) scatterplot of means (x-axis) vs. variances (y-axis) for each qPMD. Ellipses are drawn using stat_ellipse in ggplot assuming multivariate *t*-distribution at level 0.95. DBS = green and maternal serum = purple).

Statistical Analysis

Similar to Chapter 2, differences between distributions of baseline covariates between cases and controls were examined using a Wilcoxon rank-sum test for continuous variables, Fisher's exact test for binary variables, and a Chi-squared test for categorical vari-

ables (Table 2.2).

The relationship of linolenic and linoleic acid with case-control status was investigated using linear regression where, for each metabolite, its log abundance was regressed on case-control status adjusting for batch, run order, spot age, hemoglobin, infant sex, and ethnicity, as described in Petrick et al. [62]. For each, the regression coefficient associated with the case-control variable ($\hat{\beta}$) was exponentiated to acquire an estimate of the fold-change and an approximate 95% confidence interval was constructed as $[\exp(\hat{\beta} - 1.96 \times \hat{se}_{\hat{\beta}}), \exp(\hat{\beta} + 1.96 \times \hat{se}_{\hat{\beta}})]$, where $\hat{se}_{\hat{\beta}}$ is the estimated standard error of $\hat{\beta}$.

For data-driven discovery, we used a feature selection ensemble method described previously to select qPMDs associated with case-control status [62]. Briefly, a linear regression was fit for each j^{th} qPMD with the following model:

$$E[log(X_j)|W, Y] = \beta_{case-control} Y + W\beta,$$

where for the DBS data, $X = X_b^R$ and W contains the variables *sex*, *race*, *mom's education*, *delivery mode*, and *gestational birth weight category* (*LGA* [82]). For maternal serum, $X = X_m^R$ and W contains *sex*, *race*, and *mom's education*. X and W will serve as shorthand for the aforementioned appropriate counterparts here on out. For both datasets, qPMD features were ranked by the nominal p-value associated with the t-test for $\beta_{case-control}$.

Next, a bootstrapped regularized regression (LASSO) was fit where case-control status was regressed on all the qPMDs along with the previously mentioned covariates ie.

$$E[Y|X, W] = logit^{-1}(X\beta_\alpha + W\beta_\gamma),$$

subject to the constraint $||\beta_\alpha||_1 + ||\beta_\gamma||_1 < \lambda$. This was done separately for the DBS static qPMD data and the maternal serum static qPMD data. In either case, the features were ranked based on the proportion of the 500 bootstrapped iterations that their associated β coefficient in the model was non-zero.

The same input was given to Random Forests and variables were ranked by the mean decrease in Gini index and selected if there was a separation between them and other features further down the ranking. The concordance of the rankings from the parametric regression methods was used to select a set of features which was then supplemented by the features selected by Random Forests to allow for non-linear relationships and interactions.

For the neonatal data, hierarchical clustering was performed on Spearman correlation values with the log abundance values of linoleic and linolenic acid both with and without adjusting for *run order*, *spot age*, *batch*, and *hemoglobin*. For both neonatal and maternal data, descriptive analysis was performed using hierarchical clustering of pairwise Spearman correlation values using the R package superheat as well as PAM clustering using one-minus-correlation as the distance measure and selecting the number of clusters (between 2 and 15) maximizing the silhouette width [11]. Empirical cumulative density functions (CDF) of correlation values were visualized for each dataset. Relative variable importance

measures using the mean decrease in the Gini index from Random Forests models predicting outcome status Y using the respective set of selected features and covariates W were examined. Note that Random Forests was used during the feature selection process so high predictive performance of at least a subset of the selected features is to be expected.

In addition to the individual analyses detailed above, we also repeated the descriptive and statistical analysis on the dataset generated by combining neonatal and maternal features together, concatenating them for the mother-child pair. To examine whether the previously mentioned differences in feature variances between the neonatal and maternal data was influencing the results, feature selection of the joint data was repeated after mean-centering and scaling each qPMD by its standard deviation to create a scaled normalized dataset.

3.3 Results

There were 122 pediatric leukemia cases and 122 matched healthy controls and their mothers recruited to this study. Early life and pregnancy demographics of the child–mother pairs are described in Table 2.2. Cases were diagnosed with ALL between one to seven years old and born between 2000–2008. There were more males than females (138 and 106, respectively) and more Hispanic children than non-Hispanic White children (138 and 73, respectively) in the study. Leukemia cases and healthy controls at birth were similar in birth weight, gestational age at birth, delivery mode, and numbers of infants born large for gestational age (LGA). The neonatal dried blood spot was collected four hours earlier in cases compared to controls (mean of 31 h versus 35 h after birth, respectively, p-value = 0.049). Mothers of the child cases and controls were similar in age, week of gestation, and highest level of education at blood sample collection.

Associations Between Linolenic and Linoleic Acid and Pediatric ALL

We previously reported a positive association between alpha/gamma-linolenic acid and linoleic acid measured in newborn DBS with the development of pediatric ALL in children diagnosed $\geq$ 6 years old, but not children diagnosed ages 1-5. Using a similar analytical method, we tested this hypothesis in the current study population. When all cases diagnosed at 0–7 years old were included in the analysis, we found slightly higher levels of both alpha/gamma-linolenic acid and linoleic acid in DBS of children that went on to develop leukemia than DBS of children that did not develop ALL (Table 3.2). Interestingly, there was no difference between levels of linolenic and linoleic acid in the subset of cases that developed leukemia between the ages of 2 to 5 years, but imprecise positive associations were found between linolenic acid and linoleic acid and ALL in the subset of children diagnosed ages 6–7, consistent with previous findings by our team.

Population	Alpha/gamma-linolenic acid	linoleic acid
Full population of cases (n = 122) and controls (n = 122)	1.010 [0.906, 1.127]	1.031 [0.950, 1.119]
Cases diagnosed ≥ 6 years old (n = 13) and controls (n = 13)	1.071 [0.768, 1.49]	1.108 [0.867, 1.417]
Cases diagnosed between 2-5 years old (n = 92) and controls (n = 92)	0.974 [0.856, 1.107]	1.003 [0.914, 1.101]

Table 3.2: Associations between log linolenic acid and linoleic acid metabolite abundances on pediatric ALL case-control status. Regression coefficients were estimated using univariate logistic regression models adjusted for batch, run order, spot age, hemoglobin, infant sex, and ethinicity as described in Petrick et al. [62]

Spearman Correlation Between qPMDs for Leukemia Cases and Controls in DBS and Maternal Serum

We then used the untargeted reatomics data to identify possible reactive molecular biomarkers in the neonatal DBS and maternal serum. After filtering, 78 and 269 qPMDs for neonatal DBS and maternal serum, respectively, were used for analysis. Pairwise Spearman correlation coefficients of qPMDs were used to compare correlation distributions in DBS and maternal serum datasets. Compared to maternal serum, DBS had a larger proportion of highly correlated qPMDs (mean $r \geq 0.5$), Figure 3.8.

Spearman correlation coefficients between pairs of qPMDs were compared between cases and controls in the DBS and maternal serum using mean-difference plots (Figure 3.9). For highly correlated qPMDs in the DBS (mean r from 0.5–1.0), there are a higher number of qPMDs below the zero-difference line indicating overall higher correlation between qPMDs in the controls compared to cases. For moderately correlated qPMDs in the DBS (mean r from 0.25–0.5 on the x-axis) there are a higher number of qPMDs above the zero-difference line indicating lower correlation in the controls compared to cases. Individual heatmaps of DBS qPMDs in the cases and controls show similar findings, with fewer qPMDs in the upper-right correlation cluster of the cases compared to controls (Figure B.1). In the maternal serum mean-difference plot between cases and controls, highly correlated qPMDs (mean r from 0.5–1) were above the zero-difference line indicating more correlation in the cases than the controls (Figure 3.9b, Figure B.2), opposite to what is observed in the newborns (Figure 3.9a). In the maternal serum, there were no observable differences between cases and controls for moderately correlated qPMDs (mean r from 0.25-0.5), with a similar distribution above and below the zero-difference line. However, in the maternal serum, slightly negatively correlated qPMDs (mean r from –0.25–0) were below the zero-difference line. Similar results can be seen in individual heat maps of case and control maternal serum, where there are dark bands of slightly negatively correlated

qPMDs in controls but not in cases (Figure B.3). The results from the DBS MD-plots may reflect a relationship between disregulation of highly regulated processes in children who would develop ALL compared to the healthy controls. This pattern does not seem to yet be present when examining the maternal serum correlation patterns.

Hierarchical clustering results for the joint analysis of neonatal and maternal qPMDs together, indicate little overlap between the qPMDs in DBS and serum (Figure B.4), suggesting unique reaction-level data from the DBS and maternal serum.

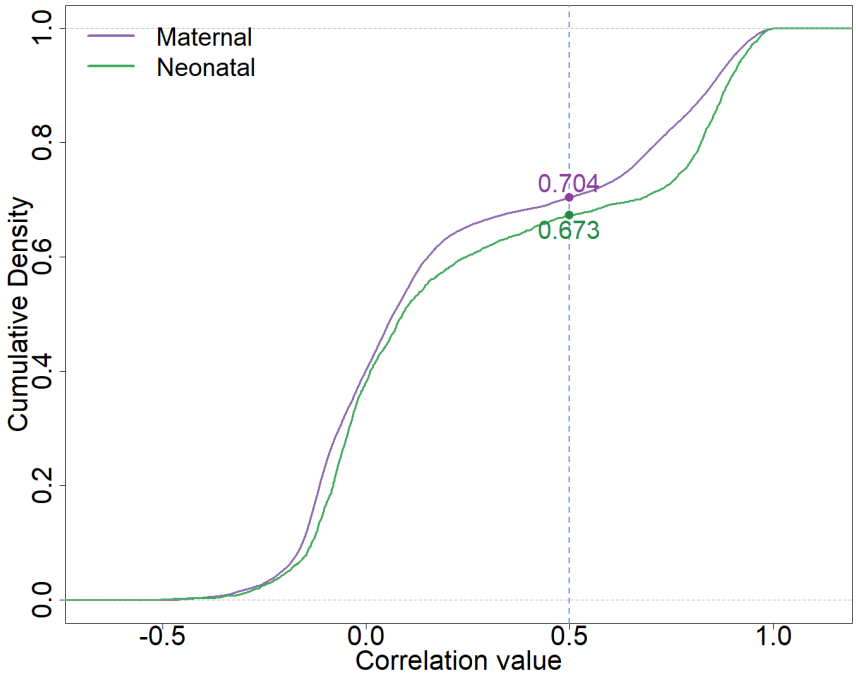
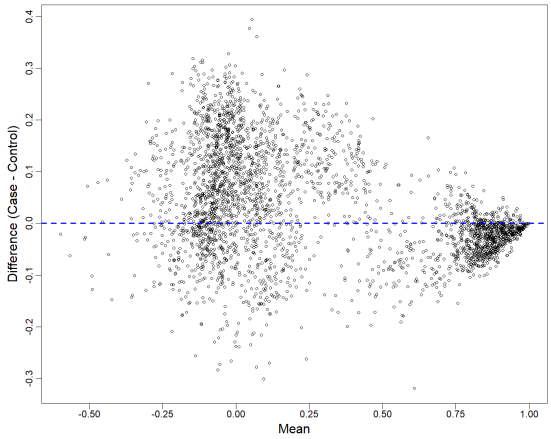


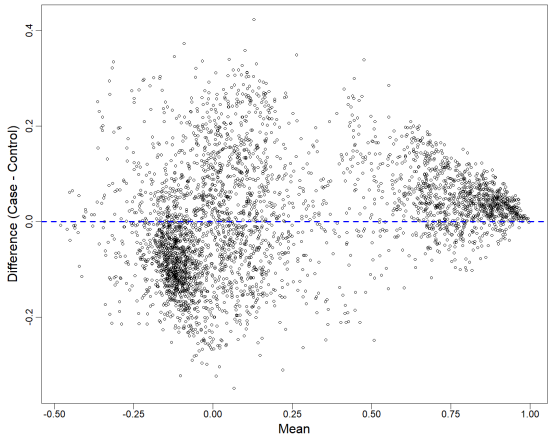
Figure 3.8: Empirical CDF plot of the upper diagonal of the Spearman correlation matrices between qPMDs for neonatal qPMD (green) and maternal qPMDs (purple). Cumulative density at r=0.5 is indicated with a dashed vertical line.

Reaction-level Differences (qPMDs) Associated with Development of Leukemia

To identify qPMDs in neonates associated with later leukemia diagnosis, we performed an ensemble feature selection on the DBS data. We identified nine qPMDs associated with case-control status (Figure 3.10a, Table 3.3). Six qPMDs, 149.08, 206.08, 5.99, 72.02, 64.04, and 41.06, were in higher abundance in DBS for the cases than the controls. Three



(a) Neonatal



(b) Maternal

Figure 3.9: Mean-difference plot of Spearman correlation values for qPMDs. a) Mean-difference plot of Spearman pairwise correlation coefficients for qPMDs in neonatal cases and controls, where each point corresponds to one of 3,003 pairs of qPMDs (for 78 individual qPMDs). b) Same as in a), for maternal qPMDs, where, to aid visualization, we only plot means and differences for a random subset of 10% (i.e., 3,605) of all possible pairwise correlation coefficients for 269 qPMDs. The x-axis shows the average correlation coefficient for cases and controls and the y-axis shows the difference (cases – controls). Blue line shows a difference of zero.

qPMDs, 136.04, 98.00, and 30.01, were in lower abundance in the DBS for the cases than the controls. These nine selected qPMDs formed three correlation clusters (Figure B.5). qPMDs 64.01 and 72.02 were moderately correlated (r = 0.46), while 30.1, 126.04, and 98.00

(r = 0.56–0.91) were strongly correlated, and 149.08, 206.08, 5.99, and 41.06 (r = 0.15–0.87) were weak to strongly correlated.

Feature selection in the maternal serum of offspring with leukemia versus healthy controls resulted in 16 qPMDs associated with case-control status (Figure 3.10b, Table 3.3), one of which, qPMD 149.08 was also selected in the DBS analysis. qPMDs 21.00, 147.07, 127.1, 0.08, 149.08, 236.05, 41.91, 129.06, 84.12, 52.07, 64.03, 99.05, 89.0, and 118.08 were positively associated with leukemia status, while qPMD 266.26 and 90.05 were negatively associated with leukemia status. These 16 qPMDs formed 4 correlation clusters (Figure B.6). qPMDs 236.05, 0.08, and 47.07 were highly correlated (r>0.72) while qPMDs 52.07, 129.0, 84.12, 149.08, 21.00, 89.0, 64.03 and 127.1 formed a cluster with low to high correlation (r = 0.25–0.95). qPMD 266.26 did not cluster with any of the other qPMDs (Figure B.6).

Feature selection in the joint datasets resulted in a total of 11 qPMDs, with seven associated with leukemia risk and four associated with leukemia resilience (Figure 3.10c, Table 3.3). Selected qPMDs in the DBS had larger fold-changes and smaller nominal *p*-values than the selected qPMDs from the maternal serum. In the joint analysis, only one new feature from the neonatal data (29.97) was selected in the joint analysis while it had not been in the DBS analysis. In fact, only two qPMDs selected from the individual DBS analysis were not selected (41.06 and 64.02) in the joint analysis, while 13 qPMDS from the individual serum analysis were not selected (90.05, 127.1, 0.08, 41.91, 236.05, 129.06, 84.12, 52.07, 64.03, 89, 118.08, 149.08, and 99.05) in the joint analysis. Indeed, qPMD 149.08 in the maternal serum was dropped from the joint feature selection while qPMD 149.08 from the neonatal blood was retained. Further, selected qPMDs associated with leukemia from maternal serum have moderate to low correlation with qPMDs associated with leukemia from DBS (Figure B.7, r <= 0.21). This suggests that qPMDs from the maternal serum provide less information related to leukemia, albeit new information is provided by both matrices. Repeated analysis on the joint dataset that was centered and scaled showed equivalent results (data not shown).

qPMDs associated with leukemia in DBS and maternal serum are summarized in Table 3.3. In DBS, *m/z* metabolite feature 383.1888 was measured in both qPMD 98.00 and 136.04, which may be a product involved in both reactions. In serum, *m/z* feature 131.0347 was measured as a potential product of qPMDs 52.07, 129.06, and 127.10, while *m/z* feature 336.1164 was measured as a potential reactant of both qPMDs 89.00 and 21.00. Since the levels of these *m/z* metabolite features change with multiple qPMDs, they may be particularly relevant for leukemia and are potential targets for future investigations.

The element composition changes of DBS qPMDs involved addition of 1, 3, 4 and 9 carbon atoms, while the serum qPMDs involved addition of carbon chains of 3, 4, 5, 7, 8, 9 or 18. In particular, compositional changes involving 8 and 9 carbon chains were enriched in the selected qPMDs compared to the KEGG database (Figure B.8), indicating that these may be particularly important in the development of pediatric leukemia. Indeed, the qPMD of 149.08 linked with the elemental +9C11HON product was selected in both maternal serum and DBS. While the *m/z* features of this qPMD were different for the two sample types, this further points to the potential importance of nine carbon reactions in

early leukemia development which may transfer from mothers to fetus.

Correlations Between Selected qPMDs and Covariates

To investigate factors that could potentially confound relationships between selected qPMDs and ALL status, the continuous covariates of maternal age, birth weight, and gestational age at birth, as well as categorical covariates of infant sex, race, LGA, and maternal education were evaluated because they have been suggested as risk factors for ALL. The relative predictive ability of these covariates to the selected qPMDs was assessed with mean decrease in Gini Index using Random Forests. All selected qPMDs in the neonatal samples were better predictors of ALL than the available covariates, albeit birth weight, gestational age, and maternal age showed much higher mean decrease in the Gini Index than the other covariates (Figure B.9). Low correlation between maternal age, birthweight, and gestational age with selected neonatal qPMDs (Figure B.10) suggests that the neonatal qPMDs have independent relationships with ALL which were not confounded by the covariates.

For the maternal serum, 14 out of the 16 selected qPMDs were better predictors of ALL than covariates, but again, gestational age, birth weight and maternal age showed much higher mean decrease in the Gini Index than the other covariates (Figure B.11). While gestational age and birth weight were only moderately correlated and clustered together (Figure B.12), maternal age clustered with qPMD 99.05 and 41.91, albeit with low correlation (r = 0.16 and 0.1, respectively). Therefore, maternal age may confound relationships between these serum qPMDs and ALL in infants.

Figure 3.11a shows hierarchical clustering on spearman correlation values of log abundances for both selected neonatal qPMD's along with linoleic and linolenic acid. Both linoleic and linolenic acid are moderately correlated with 72.02 (r = 0.54 and 0.58 respectively). The correlation strengthens slightly when linoleic and linolenic adbundances are adjusted for the previously mentioned quality control variables (Figure 3.11b, r = 0.56 and 0.62 respectively).

3.4 Discussion

Our study has elements of both hypothesis testing and generation. Though the results from the linolenic and linoleic acid analysis are not significant, potentially due to sample size, it is encouraging that the effect is in the same direction as previously reported for the same sub-population, showing a reproducible trend. Future studies powered to the adequate level for late-diagnosis cases are needed to offer more evidence. For hypothesis generation, we took a new approach to analyzing untargeted metabolomics data to understand the role of the exposome in development of pediatric ALL. Rather than looking at the metabolite level (where interpretation relies solely on imprecise annotations), we transformed our features to qPMDs to reflect broader biological signal at the enzyme level.

Selected qPMDs	*m/z* Involved	Elemental Composition of Reaction	Potential Enzyme Involved via KEGG	KEGG Reaction Class ID
		Neonatal DBS		
98.00	481.1866; 383.1888	+4C2H3O	6.3.2.46; 3.7.1.2	RC00065; RC000957; RC00326; RC00446
136.04	519.2271; 383.1888	+4C8H50		RC02065
30.01	511.1969; 481.1866; and 493.3885; 463.3772	+C2HO	2.2.1.1; 2.2.1.2; 2.2.1.5 1.14.13.0; 1.14.14.42	RC00001; RC02295; RC00014; RC02295
149.08	164.0354; 313.1193	+9C11HON		RC01859; RC02353
206.08	329.1137; 123.0323	+9C10H2O4N	6.1.2.2	RC00041
72.02	311.2223; 383.2415	+3C4H2O	1.14.11.43; 1.14.11.44	RC03192; RC03193
41.06	662.3637; 621.3033	+3C7HN/-O	3.5.4.42	RC00458; RC01422
64.02	323.2585; 259.2430	+C4H3O		RC02682
5.99	335.1008; 329.1137; and 529.4246; 523.4347	+3N/-4H2O		RC02945
29.97*	603.1687; 573.1961; 509.3834; 479.4093	+2O/-2H	1.7.1.0	RC00001; RC01333
		Maternal Serum		
118.08	317.2505; 199.1702	+9C1OH	1.13.11.70	RC01329; RC01330
90.05	385.2794; 295.2274	+7C6H	1.3.0.0; 2.9.3.15; 3.1.1.70	RC01576; RC01643; RC00014; RC00137; RC00020; RC00041
236.05	825.4777; 589.4314	+8C12H8O		RC02202
0.08	335.2949; 335.2196	+9C2H/-3O2P	2.5.1.138	RC01869
147.07	414.3073; 267.2327	+9C9HON	2.3.2.0	RC0004; RC00055
52.07	183.1026; 131.0347	+5C8H/-O	1.13.11.59	RC00912; RC01690
129.06	729.4411; 600.3806; and	+9C7HN	6.3.2.0	
84.12	344.2155; 260.0994; 624.2864; 540.3659	+4C16H5O/-2NS	2.1.1.0	RC03407
149.08	409.1837; 260.0994;	+9C11HON		RC01859; RC02353
127.10	258.1392; 131.0347; and 387.2002; 260.0994	+7C13HON		RC01908
89.00	336.1164; 247.1184	+3C3H4O/-N	4.1.3.27	RC02148; RC02414
21.00	336.1164; 315.1123	+8CH3S/-6ON2P	2.4.1.0	RC00005; RC00049
64.03	425.1785; 361.1499	+5C4H	2.3.1.246	RC00004; RC02933
99.05	699.4321; 600.3806	+8C5HN/-O	4.2.1.20; 4.2.1.122	RC00209; RC00210
41.91	489.1018; 447.1948; 527.2169; 485.3111	+6O2P/-5C4N	2.4.2.8	RC00063
266.26	926.6089; 660.3488	+18C34HO	2.3.1.288; 6.2.1.57	RC00004; RC00055

Table 3.3: Selected qPMDs, the pairs of *m/z* used to create them, element composition of reaction, potential enzyme involved, and KEGG reaction class ID.

* qPMD 29.97 was selected in the joint analysis combining qPMDs from the maternal and neonatal and performing variable selection.

In addition, as it has been posited that there are likely risk factors present in early-stage development, joint analysis of both child DBS samples as well as maternal serum samples from during gestation allowed for detection of earlier and more complicated relationships than with either set individually.

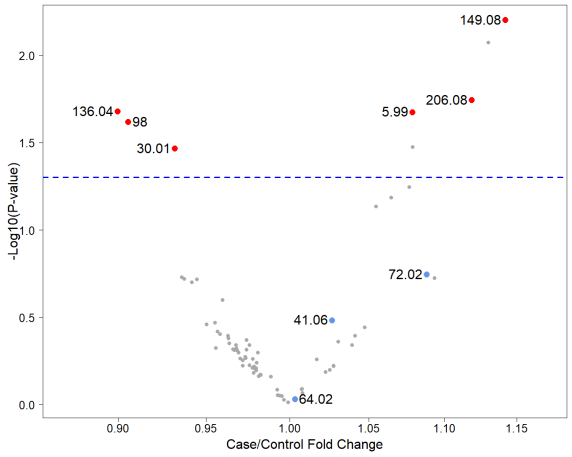
Several patterns emerged in the individual and joint analyses. In particular, weakening of correlations for highly correlated features in cases from DBS data may indicate disregulation of otherwise tightly regulated homeostatic processes in cases versus controls. For the selected qPMDs from each analysis, the potential enzymes involved are broad and include oxidoreductases, transferase, hydrolases, lyases, and ligases. In particular, ligases may be important enzyme reactions in DBS associated with pediatric leukemia, while transferases may be particularly important enzyme reactions in maternal serum associated with leukemia in offspring. Transferases are involved in many biological processes including glutathione reactions involved in cellular detoxification and glutaminolysis, which may initiate cancer; and early altering central carbon metabolism [111, 6]. Similarly, ligases are involved with cancer initiation and progression [24]. An apparent enrichment of 8 and 9 carbon reactions points to their potential importance in development of the disease and garners future examination.

Performing the joint analysis offered additional insights related to the relative rankings between features as well as an additional selected feature (29.97 from the DBS qPMDs) which was not selected in the individual analysis. The selection of 149.84 in both individual analyses (one of which retained in the joint analysis) highlights its potential importance in identifying risk factors. Similarly, annotating common metabolites involved in the selected reactions provides additional direction for future study. The robustness of the joint analysis to scaling highlights a potential benefit of this feature selection approach as both linear regression and Random Forests would not be impacted by differences in scaling. The relative rankings of features selected in the joint analysis as well as the correlation patterns among all qPMDs suggests that qPMDs from the maternal serum provide less information related to leukemia, albeit unique information, compared to the qPMDs from the DBS. In addition to reflecting the differences in from whom the sample was taken, this may also reflect differences in time-points as well as sample type (serum is a component of whole blood so will have a subset of all the metabolites present in the latter).

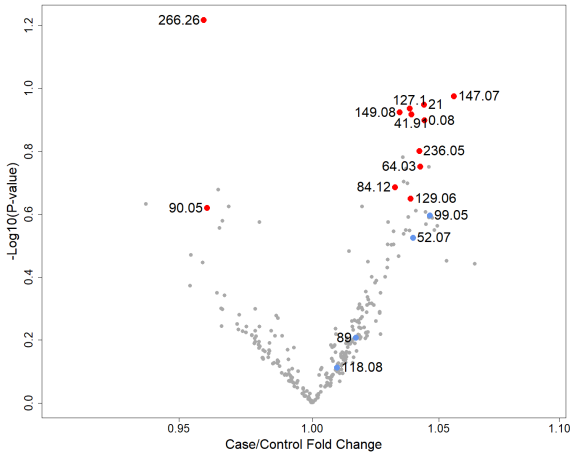
Our study has several strengths. As mentioned previously, childhood leukemia is a rare disease and having unique access to DBS samples pre-diagnosis as well as matched maternal serum samples from during gestation gives us a rare comprehensive insight into potential early-stage risk factors. Cases and controls were sex and age-matched to reduce the confounding effect of sample storage time, and storage time was additionally adjusted for during normalization. Metabolomic features were transformed to reactomic ones to circumvent the need for metabolite annotations, and thus provided novel targets at the reaction/enzyme level. This does mean, however, that included qPMD features were limited by the existing set in the database. As this area of research continues, the inclusion of more candidate reactions will be possible. Reactomics is a relatively new

area of exposomic study and we also showed the continuing importance of adequate pre-processing and normalization in this context in the quality of the resulting data. Failure to do so would likely mask any signal present. The feature selection method we used is based on several methods designed to detect both linear and non-linear associations and does not rely on arbitrary *p*-value cutoffs associated with traditional hypothesis testing frameworks. In addition, by jointly analyzing both maternal and child data, the resulting feature list reflected the combination of information stored at both time points, allowing for the detection of more complicated associations. That being said, the feature selection approach used has yet to be fully bench-marked and relies on manual application (potentially incorporating user bias). In addition, the joint analysis did not incorporate the natural time-ordering of the two feature sets. Specific enzymes of interest cannot be elucidated from this data, but our discovery study provides evidence in support of additional proteomics studies to further characterize these reactions.

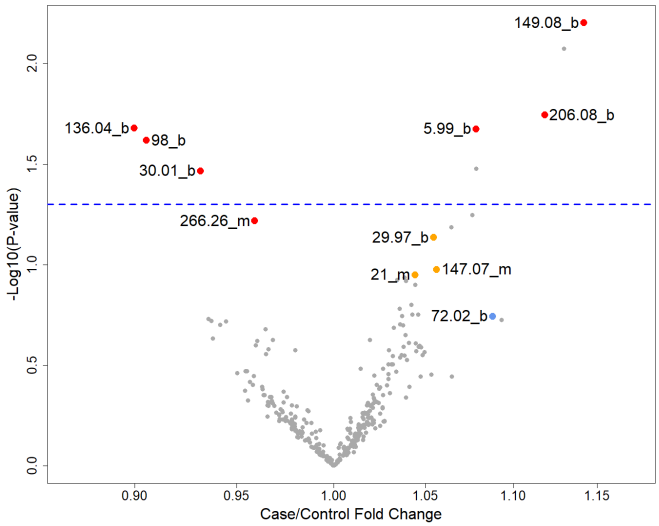
The first two chapters of this dissertation focus on elucidating potential exposomic risk factors of pediatric ALL. Chapter 2 focused on examining a known exposure while Chapter 3 broadened the search in both data source and targets to potentially find novel areas of study, relying on a previously discussed method for feature selection. The last chapter will focus on this feature selection methodology, offering automated options for its use and evidence for its utility over more traditional approaches.



(a) Neonatal

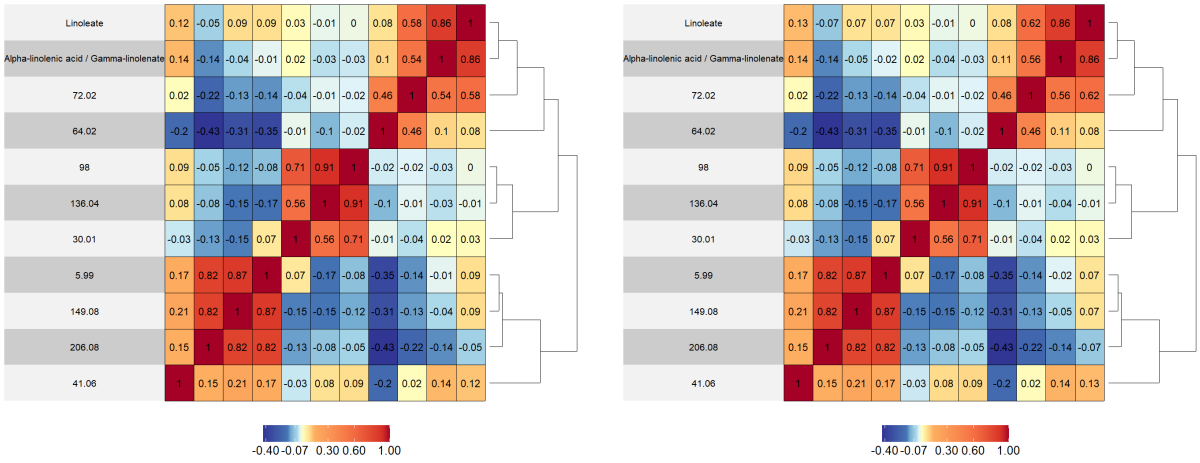


(b) Maternal



(c) Joint

Figure 3.10: Volcano plot of (a) neonatal, (b) maternal, and (c) neonatal and maternal qPMDs associated with leukemia. Numbered qPMDs were selected by the ensemble feature selection method. Estimated case-control fold-change from the linear regression is plotted on the x-axis and the negative log of the associated nominal p-value of the t-test is plotted on the y-axis. Red points indicate the qPMD was selected through the concordance of the regression methods and blue points are qPMDs that were selected only by Random Forests. Orange points (in (c)) were also selected through the concordance of regression methods using a more lenient cutoff. The dashed horizontal line corresponds to a nominal p-value of 0.05 for reference.



(a) Unadjusted

(b) Adjusted

Figure 3.11: a) Heatmap of hierarchical clustering results of Spearman correlation values of selected neonatal qPMDs with linolenic and linoleic acid. b) Heatmap of hierarchical clustering results of Spearman correlation values of selected neonatal qPMDs with linoleic and linolenic acid after adjustment.

Chapter 4

Untargeted Adductomics: Feature Selection in Lung Cancer Case-control Study and Simulation

4.1 Introduction

We now turn our attention to a new type of exposomic data: a subset of proteomics called adductomics. Adductomics, similarly to metabolomics, is designed to detect features (or adducts) which can act as proxies for various exposures of interest [26]. In the case of adductomics, these are specifically adducts of reactive electrophiles which have interacted with available nucleophiles in their environment [104]. Due to their reactivity, these electrophilic species are difficult to measure and are relatively short-lived. However, it is this same reactivity that makes them such an important target for study, as they are known to damage DNA and proteins. There are several types of adductomic analyses which focus on these proxy electrophiles gathered from different locations, e.g., DNA, Hemoglobin (Hb), Human Serum Albumin (HSA), glutathione, etc. Rappaport et al. argue that adducts purified from residues on proteins circulating in the blood (like Hb and HSA) are better candidates for study as their damage is not repaired (like DNA adducts) and they are longer-lived than other alternatives (like glutathione adducts) [88]. For the purposes of our discussion, we will focus on HSA adductomics.

In HSA, the Cys34 residue is the largest nucleophilic target for these species and was the first set of adducts to be studied in an untargeted manner [34, 35]. This pipeline was then extended to the Lys525 residue in order to broaden the potential species detected [36]. In either case, the goal of these analyses is to cleave whatever electrophiles are bound at these sites and thereby acquire an agnostic set of adducts that are representative of the species present. Though the measured adducts are relatively short-lived and dynamic, having a residence time of one month in humans, they have been shown to hold information regarding various exposures. By comparing their abundance in diseased subjects and

controls in an untargeted manner, we can hope to better understand the connections between exposures and their impact on disease risk and etiology [34].

Similarly to untargeted metabolomics analyses, MS analyses of the samples provide *m/z* and retention times, which are used to glean putative annotations in the absence of in-house standards, and the peak areas are integrated to acquire measures of feature abundance. This similarly introduces many of the considerations discussed previously regarding a high noise-to-signal ratio and the importance of appropriate pre-processing and normalization to remove unwanted technical effects before downstream analyses can be performed [92]. In addition, as relatively few in-house standards are available, most annotations are putative, being based on accurate mass and elemental composition, which can pose a challenge in interpretation of findings.

Due to the untargeted nature of this adductomics pipeline, which detects any species present, the questions often investigated relate to hypothesis generation for diseases or exposures of interest, utilizing feature selection. In the previous chapter, we discussed feature selection in the context of mother-child paired reactomics data for hypothesis generation in the context of pediatric ALL. We will now apply similar methodology in the context of an untargeted adductomics analysis. Originally, methods used to select a set of associated adducts with an outcome or exposure of interest were limited to a more traditional hypothesis testing framework (i.e., p-values using logistic or linear regression). This was limited both in modeling assumptions as well as enforcing potentially arbitrarily stringent cutoffs used in hypothesis testing frameworks (i.e., p-value cutoffs). These become particularly difficult with data, such as adductomics, where measurements are continuous (and have error), which dampens the signal-to-noise ratio. This is compounded when multiple testing issues are introduced. To address some of these limitations more recently, the ensemble method described previously and in Chapter 3 has been used to this end [35, 37]. However, while the ensemble method makes intuitive sense, there is yet to be any attempt to validate it and compare its performance with other common practices in this setting, namely logistic regression. Answering this question raises another issue: that the process is not yet automated. Therefore, in this chapter we seek to both present several options for automating the ensemble feature selection pipeline as well as compare its performance under various DGPs. The dataset we will use as our motivating example is a lung cancer matched case-control untargeted adductomics dataset, which we describe in detail below.

4.2 Lung Cancer Analysis

Background

Lung cancer is the leading cause of cancer mortality worldwide, accounting for 1.82 million deaths in 2022 [17]. Although most lung cancers are caused by tobacco smoking, nearly 25% occur among never-smokers and are predominately lung adenocarcinomas (LUADs)

that are distinct from smoking-related cancers both histologically and epidemiologically [64]. In never-smokers, LUADs often contain oncogenic EGFR mutations, particularly among East Asian women (including those residing in the United States), and differ from those of never-smoking women in other geographic regions [64, 27]. This implicates interactions across genetic factors, EFGR mutations, and environmental factors as potential causes of lung cancer in never-smokers. Among never-smoking Chinese women, environmental risk factors for lung cancer include environmental tobacco smoke (ETS, secondhand smoke) [140], exposure to indoor/outdoor air pollutants [64, 44], including cooking fumes [30, 131], and dietary variables [125, 101].

The International Agency for Research on Cancer (IARC) has declared air pollution to be carcinogenic based on evidence that combinations of pollutants and genetic factors cause lung cancers [63]. Across the plethora of air pollutants, fine particulate matter less than 2.5 μm in diameter ($PM_{2.5}$) is regarded as a primary risk factor for LUADs in nonsmokers [41]. Current exposures to $PM_{2.5}$ in urban China are among the highest in the world with average $PM_{2.5}$ concentrations in Shanghai exceeding 80 $\mu g/m^3$, or roughly 20 times the World Health Organization (WHO) health guideline of 5 $\mu g/m^3$ [64]. Following deposition in the deep lung, $PM_{2.5}$ mobilizes inflammatory cells (alveolar macrophages and polymorphonuclear neutrophils) that generate reactive oxygen species (ROS) as a defense to kill invading microorganisms [43, 107]. The same inflammatory response to nonviable $PM_{2.5}$ - including soot, fibers, transition metals and cooking fumes - damages pulmonary epithelial cells and releases an array of cytokines including interleukin-1β (IL-1β) [43].

Based on an elegant set of experiments in which genetically engineered mice were induced with oncogenic human EGFRL858R and then exposed to relevant doses of $PM_{2.5}$, Hill et al. demonstrated that macrophages are a primary source of IL-1β and that signaling of this cytokine promotes EGFR-driven LUADs [41]. Then, using aggregate data from a set of epidemiological databases, including never-smokers, the authors reported an association between "... the frequency of EGFR mutant lung cancer incidence and increasing $PM_{2.5}$ levels". Based on these results and the literature, they concluded that the "... data suggest a mechanistic and causative link between air pollutants and lung cancer ..." with $PM_{2.5}$ being an important tumor promoter. However, because tumors in mice exposed to $PM_{2.5}$ did not contain increased numbers of EGFR mutant cells, they inferred that LUADs must be initiated by mutagenic co-exposures.

Prominent mutagenic exposures to air pollutants involve the ubiquitous mixture of polycyclic aromatic hydrocarbons (PAHs) adsorbed on soot particles arising from combustion processes as well as aerosols generated from pyrolysis of cooking oils (both subsets of $PM_{2.5}$) [100, 19]. PAHs can cause cancer by pathways involving metabolism to mutagenic diol epoxides that form bulky DNA adducts and initiate tumors, and to pairs of PAH catechols and quinones, metabolites that undergo redox cycling with generation of tumor-promoting ROS [74, 81]. Diesel-engine exhaust (DEE) contains nanoparticles (diameters $< 0.1\mu$m, also a subset of $PM_{2.5}$) with adsorbed PAHs and mutagenic nitro-PAHs [47].

The hypothesis that exposure to $PM_{2.5}$ and PAHs can cause lung cancer in never-smokers is supported by studies of inhabitants in Xuanwei, China who are exposed to effluents from indoor combustion of bituminous ("smoky") coal that contain massive concentrations of $PM_{2.5}$ and carcinogenic PAHs [66]. Although women in Xuanwei are primarily nonsmokers, they nonetheless have very high incidence of lung cancer [9] with LUAD tumors that harbor *EGFR* mutations [44]. Furthermore, Ho et al. conducted *in vitro* experiments in which human LUAD cells, with and without *EGFR* mutations, were exposed to samples of $PM_{2.5}$ from indoor coal combustion in Xuanwei, and found a dose-related reduction in viability of cells containing *EGFR-TKI* mutations [42, 138]. Finally, Lan et al. reported that a single-nucleotide polymorphism in the *AKR1C3-Gln/Gln* genotype, which regulates production of PAH catechols/quinones that generate ROS, was associated with increased incidence of LUADs in Xuanwei, particularly in never-smoking females [57].

In addition to production of ROS by inflammatory cells, ROS are generated by non-phagocytic cells as signaling molecules for oxidation and regulation of target proteins that promote critical cellular functions [53]. Under homeostasis, ROS are removed by reactions with cellular and extracellular antioxidants, of which low-molecular-weight thiols and protein thiols form the basis of redox biology [53, 135]. When ROS production exceeds the body's homeostatic defenses a state of oxidative stress exists, which can be mitigated by antioxidants derived from cellular processes and the diet. Indeed, there is evidence in both smokers and never-smokers for inverse associations between lung cancer and increased fruit and vegetable intake [51, 115]. Increased dietary soy, which contains isoflavones having strong anti-inflammatory and antioxidant properties (26), was associated with reduced risk of lung cancer in never-smokers from the Shanghai Women's Health Study (SWHS) [125]. An untargeted urinary metabolomics study of never-smokers in this population identified biological pathways associated with lung cancer, including oxidative stress, inflammation and one-carbon metabolism (OCM) [47].

One-carbon metabolism represents a network of pathways involving folate, several B vitamins, homocysteine (hCys), and methionine and its derivative S-adenosylmethionine (SAM), a universal methyl donor for DNA and other molecules [25]. Methylation of DNA drives gene expression and the methionine–hCys cycle leads to production of antioxidants including Cys and glutathione (GSH) [25]. Given the dynamic nature of OCM, various co-dependent molecules can become imbalanced, thereby altering redox homeostasis and expression of genes that modulate suppression and progression of cancers [58]. Pathways involving folate and methionine-hCys cycles were reported to reduce risks in lung-cancer cases from the European Prospective Investigation into Cancer and Nutrition (EPIC) [8] and meta-analyses found that increased levels of folate and methionine in urine or blood were associated with reduced lung-cancer risks [58, 126].

The above evidence points to ROS as key intermediates in lung-cancer etiology among never-smokers that involve the interplay between indoor/outdoor air pollutants, antioxidants and redox biology. This motivates research to compare levels of ROS between nonsmoking lung-cancer cases and controls in prospective cohorts. Although ROS cannot

be directly measured in vivo, blood levels can be inferred from their stable adducts in human serum albumin (HSA) [88], which is the primary scavenger of ROS in the interstitial space. In particular, the Cys residue (Cys34) of the third largest tryptic (T3) peptide of HSA (^{21}ALVLIAFAQYLQQC^{34}PFEDHVK^{41}, *m/z* 811.7593) is a nucleophilic locus that accounts for most of the antioxidant capacity of serum [3], and two Cys34 oxidation products (the sulfinic and sulfonic acids) are recognized biomarkers of oxidative stress [53, 23, 113]. Furthermore, measuring adducts of the $\mathcal{E}$-amino group of Lys525 in another HSA peptide (^{525}KQTALVELVK^{534}, *m/z* 564.8529) expands coverage to different classes of adducts including Schiff bases and advanced glycation end products [36].

The aim of this study is to determine whether alterations to the Cys34/Lys525 adductome is associated with lung-cancer risk in never-smokers from the SWHS, using plasma collected up to 14 years prior to diagnosis. By considering correlations among Cys34/Lys525 adduct features (hereafter simply "features") and covariates including biomarkers of $PM_{2.5}$ and PAH exposure and dietary factors associated with lung cancer, we seek insight from this unique vantage point into roles played by air pollution, oxidative-stress and OCM pathways in lung carcinogenesis.

Methods

Study Population

Our study population of interest is women with lung cancer who are never-smokers. The Shanghai Women's Health Study (SWHS) is a prospective population-based cohort study of 74,942 women living in urban Shanghai, China who were enrolled between the ages of 40 and 70 years in December 1996-May 2000 [139]. Within this larger prospective-cohort study, we conducted a nested case-control study of 573 never-smoking women with available plasma samples including 278 incident lung cancer cases and 293 cancer-free controls. Controls were individually matched to cases on date of birth (±2 years) and date of sample collection (±3 months). Procedures for biospecimen collection and storage have been previously described in detail [139]. Participants completed a baseline survey, which collected information on demographics, anthropometric and lifestyle factors; dietary habits were gleaned from a food frequency questionnaire with 7-day recall preceding blood collection [139].

Lung Cancer Ascertainment

The study participants were followed using in-person surveys every 2-3 years and periodic linkage to the population-based Shanghai Cancer Registry, Shanghai Vital Statistics Registry, and Shanghai Resident Registry was conducted [139]. In-home visits and a review of medical records by a panel of oncologists were used to verify all possible lung cancer diagnoses. Lung cancer was defined as code 162 under the International Classification of Diseases, Ninth Revision (ICD-9). All female lung cancer patients were captured upon

diagnosis after recruitment through December 31, 2009. Lung cancer cases diagnosed before recruitment were excluded from the study. Data on lung cancer histological subtypes were extracted from medical records and classified according to code 8140/3 from the International Classification of Diseases for Oncology, Second Edition (ICD-O-2). Of the 278 incident lung cancer cases, we excluded from statistical analyses 72 cases that had not been histologically confirmed (HC), leaving 206 HC cases of which 148 were either LUADs (133) or adenosquamous carcinomas (18) (combined hereafter as 'LUADs'). The remaining HC cases included small cell carcinomas (2), 'other' (3), 'unclassified' (22) and 'missing ICD-O code' (22).

Mass Spectrometry and Data Acquisition

Plasma samples were maintained at −80°C for up to 21 years prior to analysis. Laboratory investigators were masked with regard to disease status except that case-control pairs were analyzed in the same daily analytical batch and in random order to reduce technical variation. Sample processing and analysis by nano-liquid chromatography tandem-mass spectrometry (nLC-HRMSMS) were performed with duplicate injections as previously described [36, 37, 34] and elaborated in Appendix C.1. Briefly, after digesting HSA with trypsin, an isotopically labeled T3 peptide that had been modified with iodoacetamide (IAA-iT3) was added as an internal standard to normalize data for instrument performance. One microliter of each digest was then analyzed by nLC-HRMSMS to locate spectra from the Cys34 and Lys525 peptides. The corresponding precursor ions were extracted to obtain a monoisotopic mass (MIM) for each peptide feature and assign accurate masses. To normalize peak areas for the amount of HSA in each sample, the MIM was also extracted for the robust doubly-charged HSA peptide ^{42}LVNEVTEFAK^{51} (m/z=575.3111, referred to as the 'housekeeping peptide'), adjacent to the T3 peptide. Peak picking and integration were performed with the average MIMs and retention times. Added masses relative to the Cys34 thiolate ion and the Lys525 peptide were estimated as described in Appendix C.1. Putative features were annotated based on a combination of retention time, accurate mass, elemental composition, database searches and available reference standards.

Several urinary metabolites of exogenous PAHs and nitro-PAHs were also measured by LC-MS/MS, as described in Appendix C.1, for use as covariates in statistical analyses. Two composite sets of analytes were generated, namely 'hydroxy-PAHs' (sum of 2-, 3-, and 4-hydroxyphenanthrene, 1,9-dihydroxyphenanthrene, and 10-hydroxypyrene) that are metabolites of unmodified PAHs, and 'amino-PAHs' (sum of 9-aminophenanthrene, 3-aminobenzanthracene, 1- and 2-aminonaphthalene, 2-aminofluorene, and 1-aminopyrene) that are metabolites of nitro-PAHs, arising primarily from diesel emissions [47, 127].

Statistical Analysis

We assume our data are n identically distributed random draws of $\{X_i, W_i, Y_i\}_{i=0}^{n} \sim P_0 \in \mathcal{M}$, where X is our $n \times J$ matrix of continuous adduct abundances, W consists of the measured baseline covariates which are a mix of continuous, categorical, and binary variables, and Y is the observed binary outcome variable which indicates lung cancer status. The relationship between (X, W) and Y is unknown and our goal is to select the full subset of adducts X^S which are truly associated with the outcome Y in the underlying data generating distribution P_0 while not selecting any that are not. The biological question at hand, along with the observational nature of the study design as well as the noisy nature of the data, lends itself to framing this question in the context of hypothesis generation.

Our adduct data (X) consists of nLC-HRMSMS peak areas for 34 Cys34-containing features and 13 Lys525-containing features, measured on more than 20% of 573 never-smoking subjects, with duplicate injections for all but two individuals ($n = 1,144$). Statistical analysis was conducted in the R environment (version 4.2.2) and follows methods described previously [1, 35, 37]. A random effects model (Equation 2.1), with the log feature abundance as the outcome and the subject-specific random effect and random error term was used to estimate the intraclass correlation coefficient (ICC) for each feature, representing the proportion of variance that was not attributed to technical variation. In this case, the duplicate injections were for the biological samples. As none of the ICC values were below an a priori cutoff of 0.2, no features were excluded. Duplicate injections were averaged, ignoring missing values and features were filtered based on percent of missing values across subjects. Three features with more than 30% missing values were removed after checking for differentially missing values by case-control status using Fisher's exact test; this left 44 features for subsequent analysis. Any values that were still missing were imputed using the k-Nearest Neighbor algorithm with k equal to 4. The scone package was then used to evaluate various candidate normalization schemes based on several criteria measuring how well technical noise was removed while maintaining the signal of interest [21]. The inputs to scone were the feature abundance matrix, a batch variable, the QC matrix (comprised of the housekeeping peptide, internal standard, sample age, and run order), case-control status, and several candidate scaling methods [DESeq, Trimmed Mean of M-values, upper-quartile, and no scaling]. Based on scone results, the selected normalization procedure adjusted for all QC measures as well as batch and case-control status.

Following normalization, separate analyses were performed for comparing all HC cases to all controls and only LUAD cases (a subset of the HC cases) to all controls. A previously described ensemble method was applied to select the features that appeared to have the strongest association with the outcome for either HC cases or LUAD cases [62, 92, 37]. First, linear regression was used where each logged feature abundance was regressed on case-control status, batch, the QC measures, and age, and the estimated regression coefficient corresponding to the case-control variable as well as the associated nominal p-value from the t-test were retained. Next, regularized logistic regression (LASSO)

was used with case-control status as the outcome, and all normalized adducts and age as predictors. This was performed over 500 bootstrapped iterations and the features were ranked based on the number of times their corresponding coefficients were non-zero. Features that ranked highly according to both nominal p-value from the linear regression and the bootstrapped LASSO were selected. Finally, the same inputs were given to Random Forests and any additional features that appeared important based on the mean decrease in Gini Index, were selected, thereby allowing for detection of non-linear relationships. Hierarchical clusterings of features using pairwise Spearman correlation were visualized with the R package 'Superheat' [11]. Heatmaps were similarly used to visualize feature correlations with various dietary factors and urinary PAH biomarkers as well as age, body-mass index (BMI), and education level (defined below), and these variables were given to Random Forests to examine relative importance as predictors of case-control status.

We used permutation tests to check for associations between relative feature abundance (case to matched control) and years from recruitment to diagnosis (ttd) to differentiate between potentially causal effects or reactive effects of disease progression (reverse causality) [35, 37, 85, 86]. For each of 10,000 random permutations of the vector of ttd (thereby breaking any association with relative case-control feature abundance), the estimated regression coefficient of case-control log-fold-change (LFC) on ttd was recorded. The p-value reported is the proportion of the 10,000 random permutations in which the absolute value of the estimated regression coefficient was greater than or equal to the absolute value estimated using the observed data. If a significant linear trend (either positive or negative) in LFC for a given feature was detected with increasing ttd, the feature was considered as potentially reactive. The permutation test for each feature used its respective set of HC or LUAD cases and controls.

Results

Table 4.1 lists summary statistics for the lung cancer cases and controls covariates selected based on previous associations with lung-cancer risk. Across our subjects, evidence of a significant inverse association between lung cancer and BMI was observed, consistent with meta-analyses [130, 141]. Soy consumption, age, and education level were also lower in cases than controls (p-values < 0.05). Table 4.1 also provides statistics for ttd (years from recruitment).

Covariate		**HC cases** (n = 206)	**LUAD cases** (n = 148)	**Controls** (n = 293)	***p*-value (HC)**[1]	***p*-value (LUAD)**[1]
BMI (kg/m^2)	Min	16.51	16.59	16.45	0.001*	0.005*
	Mean	23.82	23.82	24.90		
	Median	23.82	23.90	24.44		
	Max	35.42	35.42	36.18		
	Not Available	0	0	1		
Amino-PAHs[2] (pg/mL)	Min	555.8	591.6	288.1	0.226	0.449
	Mean	4783	4949	3916		
	Median	3544	3508	2951		
	Max	57121	57121	16024		
	Not Available	116	85	178		
Hydroxy-PAHs[3] (pg/mL)	Min	95.87	95.87	105.5	0.820	0.840
	Mean	1040	1010	978.0		
	Median	872.0	854.4	799.9		
	Max	3914	3914	3578		
	Not Available	126	95	189		
Education	College +	30	21	34	0.024*	0.003*
	High School	60	50	60		
	Middle School	57	39	80		
	Elementary -	59	38	119		
Age (years)	Min	40	40	40	0.002*	0.0003*
	Mean	56.2	55.5	58.5		
	Median	57	56.5	62		
	Max	70	70	70		
All Soy Consumption (g/d)	Min	6.29	14.42	2.28	0.014*	0.270
	Mean	211.3	236.8	249.6		
	Median	89.10	100.3	151.3		
	Max	1099	1099	1277		
All Vegetable Consumption (g/d)	Min	52.2	59.5	23.8	0.487	0.075
	Mean	308.7	320.7	298.1		
	Median	1.393	1.440	1.454		
	Max	1050	983.6	878.6		
Methionine (g/d)	Min	0.414	0.414	0.630	0.781	0.659
	Mean	1.501	1.522	1.505		
	Median	1.393	1.440	1.454		
	Max	3.986	3.826	3.167		
Folate (μg/d)	Min	45.3	45.3	68.4	0.675	0.182
	Mean	236.6	269.8	264.3		
	Median	250.0	262.8	244.0		
	Max	692.2	658.3	1155		

continued . . .

... continued

Covariate		**HC cases** (n = 206)	**LUAD cases** (n = 148)	**Controls** (n = 293)	***p*-value (HC)**[1]	***p*-value (LUAD)**[1]
ETS (Ever Exposed?)	No	44	34	75	0.522	0.624
	Yes	141	133	204		
	Not Available	21	15	14		
Time to Diagnosis (years)	Min	0.183	0.183			
	Mean	7.533	7.568			
	Median	8.355	8.475			
	Max	13.503	13.336			

* p-value < 0.05

[1] Nominal p-values for testing for differences in covariate distributions between cases (HC or LUAD) and controls were computed using a χ^2 test for categorical variables and a Wilcoxon-rank sum test for continuous variables.

[2] Sum of 9-aminophenanthrene, 3-aminobenzanthracene, 1-, and 2-aminonapthalene, 2-aminofluorene and 1-aminopyrene.

[3] Sum of urinary 2-,3-, and 4-hydroxyphenanthrene, 1,9-dihydroxyphenanthrene and 10-hydroxypyrene.

Table 4.1: Summary statistics for covariates measured on histologically-confirmed (HC) cases, LUAD (lung adenocarcinoma) cases, and controls.

The 34 Cys34 and 13 Lys525 features are listed in Figures C.1 and C.2, respectively. For each feature, the nLC retention time, observed and theoretical MIMs and the corresponding mass deviation (Δmass), added mass and elemental composition are shown along with the annotation and the mean nLC-HRMSMS peak area for each feature. Table C.1 shows the values of the ICC (median=0.55, range: 0.25-0.93) and coefficient of variation (CV, median=0.37, range:0.18-1.4) based on the random effects model of duplicate LC-HRMSMS analyses of feature abundances. The impact of technical variation on feature selection was reduced by averaging duplicate observations for each subject in the analyses. Tables C.2 and C.3 show FCs and the nominal p-values for testing the null hypothesis that the case-control coefficient from the multivariate regression is zero using a t-test for HC and LUAD cases, respectively, and all controls. Thirty four of these 44 features have been reported previously in our laboratory, as footnotes in Figures C.1 and C.2.

Figures C.3 (A-C) and Figures C.4 (A-C) show the results of feature selection for HC cases-controls and LUAD cases-controls, respectively, and Table 4.2 combines the results to list the nine selected features. These included seven Cys34 features, i.e., 822.42 (Cys34 sulfinic acid), 827.088 (S-methanethiol), 845.42 (S-Cys-H2O), 849.07 (S-S-sulfonic acid trisulfide), 856.10 (S-hCys), 858.75 (Na adduct of S-Cys), and 914.83 (unknown), and two Lys525 features, i.e., 571.84 (oxidation product, -H2+O) and 587.31 (unknown). Most of the selected features were common to both HC and LUAD analyses, exceptions being 822.42 (only for HC) and 849.07, 858.75, and 827.088 (only for LUADs). Also, two of the five features from the overlapping list (587.31 and 849.07) had stronger signals in the LUAD analysis despite the smaller sample size (148 cases as opposed to 206).

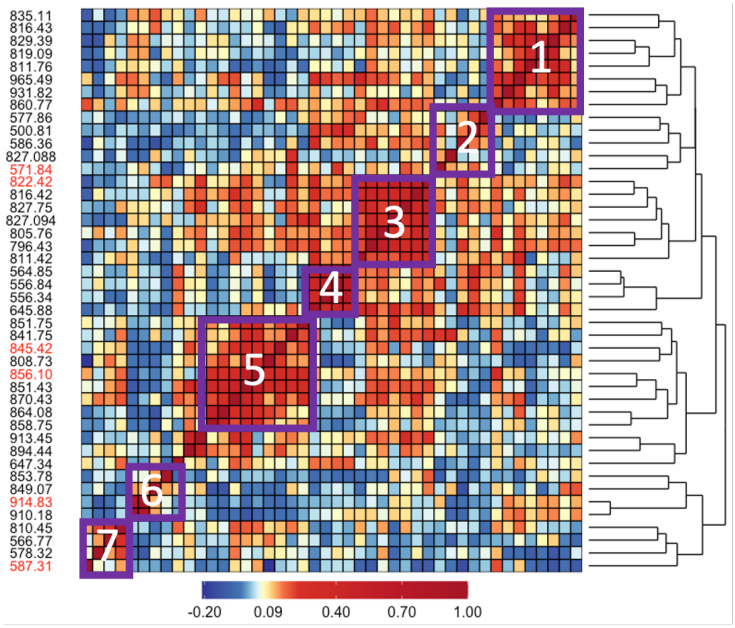
		Histologically-confirmed				LUAD			
Feature	**Annotation**	**FC**	***p*-value**	**LASSO**	RF	**FC**	***p*-value**	**LASSO**	**RF**
571.84	Lys525 oxidation (-H_2+O)	0.842	**0.040**	**Yes**	**Yes**	0.823	**0.095**	**Yes**	**Yes**
587.31	Unknown	0.921	**0.081**	**Yes**	No	0.896	**0.040**	**Yes**	No
822.42	Cys34 sulfinic acid	1.042	**0.096**	**Yes**	**Yes**	1.018	0.504	No	No
827.088	S-Methanethiol	0.963	0.453	No	No	0.999	0.922	No	**Yes**
845.42	S-Cys(-H_2O)	0.932	**0.014**	**Yes**	**Yes**	0.937	**0.047**	**Yes**	**Yes**
849.07	S-Sulfonic acid trisulfide	0.935	**0.079**	No	No	0.912	**0.028**	**Yes**	No
856.10	s-hCys	0.942	**0.044**	**Yes**	**Yes**	0.946	**0.087**	**Yes**	**Yes**
858.75	Na adduct of S-Cys	1.029	0.400	No	No	1.070	**0.087**	**Yes**	No
914.83	Unknown	1.114	**0.047**	**Yes**	**Yes**	1.113	**0.073**	**Yes**	**Yes**

Table 4.2: Features selected by the ensemble pipeline for association with lung cancer. Selection criteria satisfied for each feature are shown in boldface. [Fold-change based on normalized and imputed data (FC); linear regression nominal *p*-value; bootstrap LASSO; Random Forests (RF)]

The left panel of Figures 4.1 and 4.2 show heatmaps of pairwise Spearman correlation values for all features in the two analyses with selected features shown in red. Clusters labeled in the adjoining heatmaps on the right are annotated in the embedded tables. The seven clusters in the heatmaps in Figures 4.1 and 4.2 between HC features and LUAD features, respectively, contain the same sets, although the relative locations differ in the heatmaps. This concordance of clusters again reflects the fact that LUADs represent 72 percent of HC cases.

Figures 4.3 and 4.4 show variable importance plots by Random Forests of the selected features and covariates from HC and LUAD analyses (on the left) side-by-side with heatmaps of pairwise Spearman correlation values (on the right). Based on the Random Forests variable importance plot in Figure 4.3, BMI and methionine intake were the most important predictors of HC lung cancers followed closely by features 587.31, 914.83, and 822.42. The adjoining heatmap shows that all of the selected features except 914.83 were weakly-to-moderately correlated with urinary hydroxy-PAHs and/or amino-PAHs and that these correlations were negative in all cases. Feature 914.83 was negatively correlated with 'all soy' while 822.42 was negatively correlated with the entire dietary cluster (methionine, all soy, all vegetables, and folate) as well as the cluster representing education, age, and BMI. Turning now to the Random Forests plot for the features and covariates across LUADs in Figure 4.4, BMI was the most predictive followed by 587.31, 849.07, and 571.84. As shown in the heatmap in Figure 4.4, all selected features except 914.93 and 849.07 were weakly-to-moderately negatively correlated with one or both of the PAH classes. Also, the dietary cluster was positively correlated with 845.42, while 914.83 was negatively correlated with soy intake.

To determine whether abundances of selected features may have resulted from disease progression (reverse causality) rather than their potentially causal role in lung cancer, we examined the relationships between LFC of abundances of lung-cancer case-control pairs



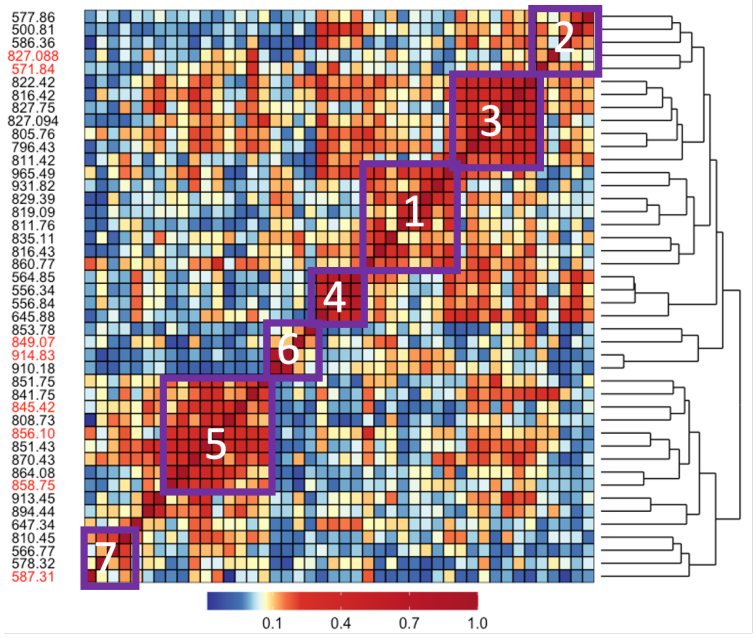
Cluster	Feature	Annotation	Fold change
1	835.11	Crotonaldehyde	1.04
	816.43	Methylation (at Glu37)	0.99
	829.39	Unknown	0.97
	819.09	Na adduct of T3 peptide	0.98
	811.76	T3 Peptide	0.99
	965.49	Unknown	1.02
	931.82	Unknown	0.98
	860.77	S-hCys (+CH3)	1.02
2	577.86	Acetaldehyde adduct	1.00
	500.81	Loss of Lys525	1.01
	586.36	Carbamylation	1.06
	827.088	S-Methanethiol	0.96
	571.84	Lys525 Oxidation (-H2+O)	0.84
3	822.42	Cys34 Sulfinic acid (+O2)	1.04
	816.42	Cys34-Gln crosslink (+O)	1.02
	827.75	Cys34 Sulfonic acid (+O3)	1.02
	827.094	S-O-O-CH3	1.04
	805.76	Cys34→OA/FG	1.00
	796.43	Cys34→Gly	0.98
	811.42	T3 Peptide dimer	1.04
4	564.85	Lys525 Peptide	1.01
	556.84	Unknown	1.01
	556.34	Unknown	1.01
	645.88	Fructosyl lysine (glycation)	1.01
5	851.75	S-Cys(NH2→OH)	1.00
	841.75	S-Mercaptoacetic acid	0.99
	845.42	S-Cys(-H2O)	0.93
	808.73	Unknown	0.94
	856.10	S-hCys	0.94
	851.43	S-Cys	1.00
	870.43	S-CysGly	0.98
	864.08	Unknown	1.01
	858.75	Na Adduct of S-Cys	1.03
6	853.78	Unknown	1.02
	849.07	S-S-Sulfonic acid trisulfide	0.94
	914.83	Unknown	1.11
	910.18	Unknown	1.10
7	810.45	Unknown	0.98
	566.77	Unknown	0.97
	578.32	Unknown	1.08
	587.31	Unknown	0.92

Figure 4.1: Heatmap of pairwise Spearman correlation matrix for all features in HC cases and controls (left) with annotations and case-control fold-changes shown (right). Features selected for association with lung cancer are highlighted in red.

and ttd. Results are presented in Figures C.5 and C.6 as individual plots for the 6 selected features in HC cases-controls and 8 selected features in LUAD cases-controls, respectively. There was no apparent linear trend by ttd for any of the selected features in their respective subsets, indicating little evidence of reverse causality for these features.

Conclusions

An increasing body of observational, mechanistic, and experimental evidence points to air pollution as a potential cause of lung cancer in never-smokers. A key element linking these lines of inquiry is production of ROS by $PM_{2.5}$ and PAH, two ubiquitous classes of air pollutants that collectively initiate and promote lung cancers. A missing component of this amalgamation of findings is direct evidence that the production and removal of ROS in lung cancer cases is imbalanced in never-smokers. The Cys34/Lys525 adductome is an emerging biomarker of exogenous exposures and endogenous processes that provides a unique vantage point for linking adducts generated by ROS *in vivo* to lung cancer. Cys34 adducts of *N*-acetylcysteine, cysteinyl glycine (CysGly), and hCys were associated with lung cancer and smoking histories in the EPIC cohort [22] and in patients with chronic



Cluster	Feature	Annotation	Fold change
2	577.86	Acetaldehyde adduct	0.98
	500.81	Loss of Lys525	1.01
	586.36	Carbamylation	1.03
	827.088	S-Methanethiol	1.00
	571.84	Lys525 Oxidation (-H2+O)	0.82
3	822.42	Cys34 Sulfinic acid (+O2)	1.02
	816.42	Cys34-Gln crosslink (+O)	1.01
	827.75	Cys34 Sulfonic acid (+O3)	1.02
	827.094	S-O-O-CH3	1.04
	805.76	Cys34→OA/FG	1.01
	796.43	Cys34→Gly	0.99
	811.42	T3 Peptide dimer	1.04
1	965.49	Unknown	1.00
	931.82	Unknown	1.00
	829.39	Unknown	1.00
	819.09	Na Adduct of T3 peptide	0.98
	811.76	T3 peptide	0.99
	835.11	Crotonaldehyde	1.03
	816.43	Methylation (at Glu37)	0.98
	860.77	S-hCys(+CH3)	0.99
4	564.85	Lys525 Peptide	1.01
	556.84	Unknown	1.00
	556.34	Unknown	1.00
	645.88	Fructosyl lysine (glycation)	1.00
6	853.78	Unknown	1.02
	849.07	S-S-sulfonic acid trisulfide	0.91
	914.83	Unknown	1.11
	910.18	Unknown	1.10
5	851.75	S-Cys(NH2→OH)	1.03
	841.75	S-Mercaptoacetic acid	1.02
	845.42	S-Cys(-H2O)	0.94
	808.73	Unknown	0.95
	856.10	S-hCys	0.95
	851.43	S-Cys	1.02
	870.43	S-CysGly	1.01
	864.08	Unknown	1.02
	858.75	Na Adduct of S-Cys	1.07
7	810.45	Unknown	0.98
	566.77	Unknown	0.96
	578.32	Unknown	1.06
	587.31	Unknown	0.90

Figure 4.2: Heatmap of pairwise Spearman correlation matrix for all features in LUAD cases and controls (left) with annotations and case-control fold-changes shown (right). Features selected for association with lung cancer are highlighted in red.

obstructive pulmonary disease (COPD, a risk factor for lung cancer) [61], but have not been investigated heretofore in never-smoking lung cancer cases. In this study we provide evidence that Cys34 adduct features do indeed link lung cancer in never-smokers with altered redox biology and further implicates OCM as an important mechanistic pathway.

Our study focused on Chinese women because they have remarkably high lung-cancer incidence despite being predominately never-smokers. This has led to speculation that the high lung-cancer risk across Chinese women reflects the confluence of high exposures to $PM_{2.5}$ and PAHs – arising from indoor/outdoor combustion of fossil fuels and household frying with cooking oils — and their interactions with genetic factors, somatic events, and tumor biology [64]. A key component to this conjecture is chronic inflammation resulting from the imbalance between production of ROS by phagocytic cells (in response to $PM_{2.5}$) [41, 43] plus redox cycling of PAH catechols/quinones [74, 81] and removal of ROS by cellular and extracellular antioxidants [135]. To investigate such imbalance of ROS production and removal in the SWHS, we monitored modifications to the Cys34/Lys525 adductome in 573 never-smoking women including 278 incident lung cancer cases. Nine features were selected for association with lung cancer (Table 4.2) using a set of regression

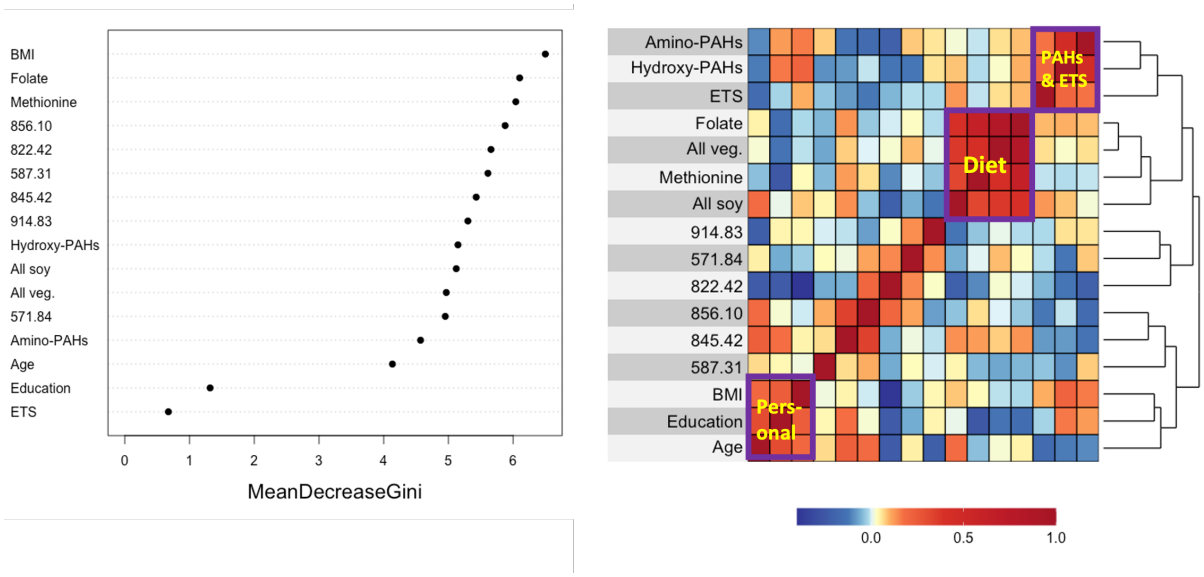


Figure 4.3: Variable importance of covariates and selected features as ranked by Random Forests in HC cases and controls (left), and heatmap showing the corresponding pairwise Spearman correlations (right).

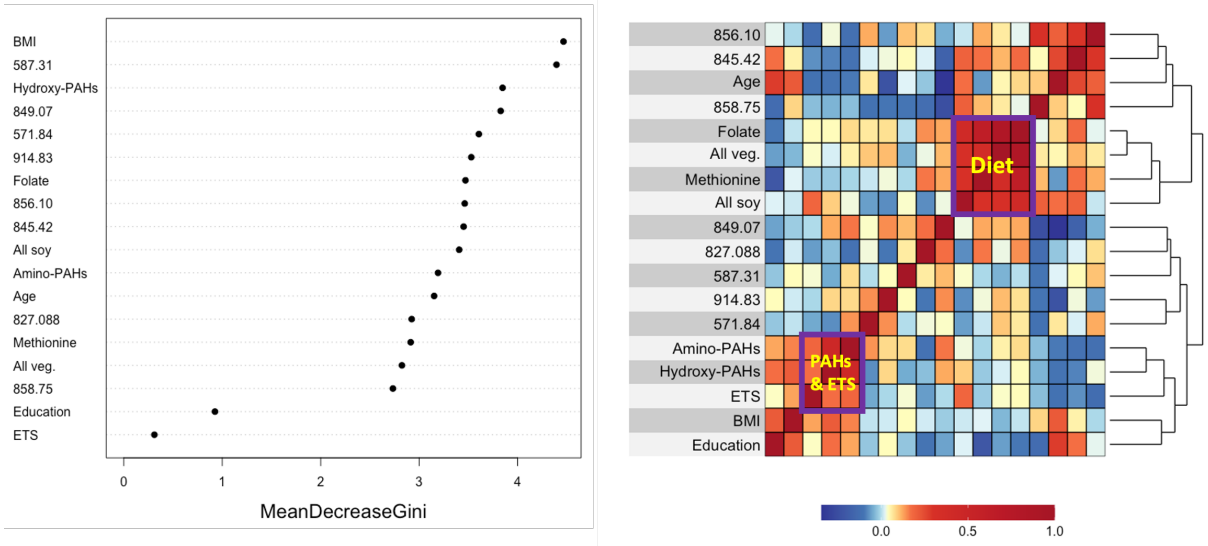


Figure 4.4: Variable importance of covariates and selected features as ranked by Random Forests in LUAD cases and controls (left), and heatmap showing the corresponding Spearman correlations (right).

and classification methods, as applied previously with incident cases and controls of CRC and NHL from the EPIC cohort [36, 37]. Interestingly, abundances of 7 of these 9 features were lower in cases than controls ($0.823 \leq FC \leq 0.946$), suggesting that enhanced ROS homeostasis in control subjects was protective for lung cancer.

Analyses of covariates by Random Forests, summarized in Figures 4.3 and 4.4, indicate that BMI was ranked highest in lung cancer cases and controls. This reflects the significantly lower values of BMI in cases here (Table 4.1) and in the literature [130, 141]. Dietary methionine was also highly ranked in HC cases and controls (Figure 4.3) consistent with elevated plasma methionine in lung-cancer cases across observational studies [58] and with dysregulated methionine and Cys metabolism in never-smoking lung-cancer cases in the SWHS [94]. Also, the high rank of the Cys34 sulfinic acid (822.42) in HC cases points to oxidative stress as a potentially causal factor and is reinforced in LUAD cases by the high ranks of the Cys34 S-sulfonic-acid trisulfide (849.07) and hydroxy-PAHs (Figure 4.4).

The heatmaps in Figures 4.3 and 4.4 show correlations of selected features with covariates for HC and LUAD cases, respectively. All of the dietary variables (folate, vegetables, methionine and soy) are correlated with each other as are the two sets of PAH biomarkers and ETS (which includes $PM_{2.5}$ and adsorbed PAHs). Several selected features were negatively correlated with the PAH/ETS cluster, namely the Cys34 disulfides of Cys (845.42) and hCys (856.10) and the Cys34 sulfinic acid (822.42) with HC cases (Figure 4.3), suggesting that ROS generated by inhaled $PM_{2.5}$ and PAH metabolism perturbed redox homeostasis. Again, of particular interest is 856.10 (hCys, FC=0.94), which is negatively correlated with ETS and the amino-PAHs (metabolites of nitro-PAHs) that are biomarkers of DEE [47], but not the hydroxy-PAHs derived from other combustion processes (Figures 4.3 and 4.4). Since we had previously found levels of 856.10 to be reduced in workers and urban residents exposed to DEE [61, 122], this suggests that DEE may have a particular connection with lung cancer in urban nonsmokers, possibly due to inflammation arising from nano-particles and nitro-PAHs that are produced primarily by diesel engines [47]. The DEE connection with lung cancer is also likely to be explained by the strong mutagenicity of nitro-PAHs, as observed in our previous work [136]. Of the dietary variables, the most noteworthy finding is the strong negative correlation between the Cys34 sulfinic acid (822.42) and consumption of soy and methionine in HC cases (Figure 4.3), which links dietary antioxidants with reduced systemic oxidative stress. Dietary methionine may be especially important because it plays a role in formation of GSH via OCM and is itself a scavenger of ROS [16]. The personal cluster (BMI, age and education) in HC cases also shows strong negative correlations with the Cys34 sulfinic acid (822.42) (Figure 4.4) suggesting inverse associations with systemic oxidative stress as subjects get older and heavier and are better educated.

This investigation has several strengths. First, by focusing on never-smokers in the SWHS we were able to investigate lung cancer risks in a particularly vulnerable population without the confounding effects of smoking. Second, because the serum samples were collected up to 14 years prior to diagnosis the potential for reverse causality was reduced. Third, the untargeted adductomics assay provides an unbiased and unique vantage point

with which to view the disposition of ROS in the interstitial space during the month preceding phlebotomy and thus is highly relevant to ROS-promotion of lung cancer. Fourth, our hypothesis-free design and ensemble of regression/classification methods permitted unbiased selection of features for associations with lung cancer, including both linear and nonlinear relationships. Fourth, expanding the Cys34 adductomics pipeline to Lys525 extended the range of adducts and their underlying chemistries, and included two selected features (571.84, Lys oxidation and 587.32, unknown).

We also recognize some weaknesses. Although storage of biological specimens for decades can produce artifacts, specimens were stored at -80°C (for up to 21 years), and cases and controls were matched by year of enrollment to minimize potential storage effects. While four selected features were confirmed with synthetic standards (822.42, 827.088, 856.10 and 858.75), annotations of the others were based on elemental composition and should be regarded as putative. Also, two seemingly important features (587.31 and 914.83) were unannotated, limiting our ability to address their implications. Since HSA adducts have a residence time of one month in humans, adduct levels measured in blood collected at recruitment may not accurately reflect those in following months or years. For our feature selection methodology, the ensemble approach we utilize is as of yet fully benchmarked and depends on manual implementation, which may be sensitive to tuning parameters or user bias. A separate test study would be necessary to determine whether features selected are predictive or if we are just selecting the top of noise. In addition, the analysis did not utilize the matched design. Matching variables, however, were still adjusted for and this had the benefit of including many more control samples. The apparent predictiveness of selected features based on the secondary analyses performed by using RF to predict case-control status based on selected features and covariates should be interpreted with caution as RF was used to partially select the features in the first place. Therefore, it would not be surprising that features that were selected by RF would be ranked highly. Relative variable importance of selected features and covariates based on a separate classification method could be beneficial as a validation. As many of the features were selected by the regression methods, however, overall relative patterns between the set of selected features and covariates are potentially informative.

In summary, this study selected 9 HSA adduct features that provided a unique vantage point for considering associations with lung cancer up to 14 years prior to diagnosis in never-smoking females (206 cases and 293 controls). Annotations of 7 of these features point to both direct and indirect links among $PM_{2.5}$/PAH exposure, oxidative stress, OCM and dietary antioxidants consistent with the literature. Several of these features have previously been selected with similar effects in adductomic investigations of lung cancer and COPD and with environmental exposures to $PM_{2.5}$, PAHs and DEE.

4.3 Simulation and Automated Analysis

In the previous section, we have shown an application of the ensemble feature selection method in the context of hypothesis generation in selecting adductomic features associated with lung cancer status. However, in this case as well as previous analyses using this method [35, 37], the current implementation requires active interaction and decision making by the user, which poses issues for reproducibility. In addition, there have been no studies showing relative performance of this method compared to alternatives used to benchmark its performance. We will propose several options for automating the procedure and compare its performance to that of the common hypothesis testing approach used in this context: logistic regression.

The goals of our simulation are to 1) compare the performance of the ensemble feature selection method to the standard practice of using logistic regression, and 2) examine how the methods' performances vary under different underlying distributional assumptions. In addition, we will compare results using the different automated implementations of the ensemble approach we discuss below on the real lung cancer data, to determine sensitivity of the results to these variations.

Data Structure

We have an $n \times J$ matrix X with n observations for J features, where the J features in X are continuous. We have a binary outcome vector Y which has length n and indicates case-control status. We define X^S as the mean-centered and standard deviation-scaled version of X. X^* and X^*_{int} denote the subset of X or interaction terms created from X, respectively, which are associated with Y according to its model. K_{int} indicates the set of k_{int} features included as interaction terms in the model for Y and K indicates the set of $k - k_{int}$ features included as linear terms.

For ease of simulation, we will limit our scope to simulating X and $Y|X$, and not including any baseline covariates W, though these can be adjusted for if necessary. The outline of the simulation is shown below and each section is described in further detail:

1. Data Generating Processes (DPGs)
    - Simulating X
    - Simulating $Y|X$
2. Feature Selection Methods
3. Visualization and Evaluation

The user inputs to the simulation are as follows:

- *data*: the real-world dataset that the DPG parameters will be learned from

- n: sample size
- k: the number of features in X that will be used in the model for generating $(Y|X^*, X^*_{int})$ in the main effects model, the interaction model, and the nonlinear model
- k_{int}: the number of features (out of k associated features) that will be used in interaction terms in the model for simulating $(Y|X^*, X^*_{int})$ in the model with both main effect and interaction terms.
- β: the possible values for the regression coefficients in the main effects model as well as the nonlinear model
- β_{int}: the possible values for the regression coefficients in the interaction model
- β_2: the possible values for the regression coefficients for the main effects terms in the model with both main effect and interaction terms
- β_{int_2}: the possible values for the regression coefficients for the interaction terms in the model with both main effect and interaction terms

Data Generating Processes

The performance of the various feature selection methods can vary depending on both the marginal distribution of X as well as the conditional distribution of $Y|X$. We therefore sought to provide several different scenarios to examine if there are some that seem to favor one method over another. An important aspect of the simulated distribution of X is that the necessary parameters were learned from the real data, in order to recreate as best as possible the data we actually observe in the real world. So we have $\hat{\mu}$ and $\hat{Cov}$ learned from the real data; $\hat{Var}$ denotes the diagonal variance matrix.

Note that, as our question of interest is related to feature selection rather than inference, we will use a scaled version of X (as shown below denoted as X^S). This aids in avoiding issues of instability in the simulation of Y and does not interfere with the conclusions we draw from results.

$$X_j^S = \frac{X_j - \bar{X}_j}{\sqrt{\frac{1}{n-1}\Sigma_i^n (X_j - \bar{X}_j)^2}},$$

where each feature has been mean-centered and scaled by its sample standard deviation.

Simulating X

In simulating the matrix of features X, we would like to best replicate our real data, as well as examine how the performance of the methods varies when using the distributions our methods are designed to perform well under versus less ideal circumstances. To that end,

we compare three multivariate distributions: normal, skew-normal (sn), and log-normal, as well as allowing for the absence of (simpler setting) or presence of (more complex and realistic setting) correlation between features. For each of the distribution families, X_1 denotes when X was simulated with a diagonal covariance matrix (no correlation between features), and X_2 denotes when X was simulated with a non-diagonal covariance matrix (correlation between features). In each case, the mean vector ($\hat{\mu}_{data}$), diagonal covariance matrix ($\hat{Var}_{data}$), and non-diagonal covariance matrix ($\hat{Cov}_{data}$) are learned from the real data. Note that in our implementation, we end up mean-centering and standard deviation-scaling each feature to increase the stability of the distribution of Y. For simulating X for the log-normal distribution, the analog mean vector and covariance matrices are computed on the logged data and are denoted by $\hat{\mu}_{log-data}$, $\hat{Var}_{log-data}$, and $\hat{Cov}_{log-data}$, respectively. In the case of the skew-normal distribution, a standard skew-normal distribution was first simulated from and then the mean shifting and covariance structure was introduced. The following describes how $X1$ and $X2$ were simulated in each of the three settings:

1. Multivariate Normal
$$X_1 \sim MVN(\mu = \hat{\mu}_{data}, \Sigma = \hat{Var}_{data})$$
$$X_2 \sim MVN(\mu = \hat{\mu}_{data}, \Sigma = \hat{Cov}_{data})$$

2. Multivariate Skew-Normal
$$X_1 = \hat{\mu}_{data} + L_1^* Z$$
$$X_2 = \hat{\mu}_{data} + L_2^* Z$$
where
$$Z \sim 2\phi(x)\Phi(\alpha x)$$
$$L_1^* L_1 = Cholesky(\hat{Var}_{data})$$
$$L_2^* L_2 = Cholesky(\hat{Cov}_{data})$$
$$\alpha = 1$$
and ϕ and Φ are the standard normal density and standard normal cumulative density functions, respectively.

3. Multivariate Log-Normal
$$X_1 \sim MVLN(\mu = \hat{\mu}_{log-data}, \Sigma = \hat{Var}_{log-data})$$
$$X_2 \sim MVLN(\mu = \hat{\mu}_{log-data}, \Sigma = \hat{Cov}_{log-data})$$

Note that in each case $5 \times n$ number of observations were simulated for each X ($X1$ and $X2$) and then once $Y|X$ was simulated, $n/2$ cases and $n/2$ controls were randomly selected. This was to ensure that each dataset would have an addequate number of cases and controls similar to the ratio seen in a 1:1 matched case-control study.

Simulating $Y|X$

We next introduce the four ways that we simulated $Y|X$, ranging from the simplest logistic model with only main effect terms (Y_1), a logistic model with only interaction effect terms (Y_2), a logistic model with a mix of main and interaction effect terms (Y_3), and a non-linear logistic model (Y_4). In each case, k and k_{int} control how many main and interaction terms respectively are used when simulating $Y|X$. As mentioned previously, in order to maintain the stability of the distribution of Y, X was mean-centered and standard-deviation scaled to get X^S. We also define $\mathcal{K}$ as the set of unique combinations of c_1 features from J and $\mathcal{K}_{int}$ as the set of unique combinations of c_2 features from J and have that:

$$P(K = \kappa) = \begin{cases} \frac{c_1!(J-(c_1))!}{J!}, & \text{if } \kappa \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}$$

$$P(K_{int} = \kappa) = \begin{cases} \frac{c_2!(J-c_2)!}{J!}, & \text{if } \kappa \in \mathcal{K}_{int} \\ 0, & \text{otherwise} \end{cases}$$

1. **Logistic model with only main terms**

$$P(K = \kappa) = \begin{cases} \frac{k!(J-k)!}{J!}, & \text{if } \kappa \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}$$

which randomly selects the k features associated with Y without replacement.

$$X^* = X^S[, K]$$

$$Y|X \sim Bernoulli(logit^{-1}(0.5 + (X^*)^T * \beta))$$

2. **Logistic model with only interaction terms**

$$P(K = \kappa) = \begin{cases} \frac{k!(J-k)!}{J!}, & \text{if } \kappa \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}$$

which randomly selects the k features associated with Y without replacement.

$$X^* = X^S[, K]$$

which randomly selects the k features associated with Y. We define X^*_{int} as the set of multiplied interaction terms from the partition of X truly associated with Y.

$$X^*_{int} = X^S_{\cdot, K[1:\frac{k}{2}]} \circ X^S_{\cdot, K[\frac{k}{2}+1:k]}$$

where $\circ$ is the Hadamard product and simulate Y according to the following logistic model.

$$Y|X \sim Bernoulli(logit^{-1}(0.5 + (X^*_{int})^T * \beta_{int}))$$

Note that the sampling process of K means that each feature can be included in at most one interaction term in the model for Y.

3. **Logistic model with both main and interaction terms**

$$P(K = \kappa) = \begin{cases} \frac{(k-k_{int})!(J-(k-k_{int})!}{J!}, & \text{if } \kappa \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}$$

which randomly selects the $k - k_{int}$ features associated with Y without replacement and

$$P(K_{int} = \kappa) = \begin{cases} \frac{k_{int}!(J-k_{int})!}{J!}, & \text{if } \kappa \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}$$

which randomly selects the k_{int} features included in interaction terms in the model for Y.

We define X^* as the partition of X truly associated with Y and X^*_{int} as the set of multiplied interaction terms from the partition of X truly associated with Y.

$$X^* = X^S[, K]$$

which creates the subset of X with the main effect terms.

$$X^*_{int} = X^S[, K_{int}[1 : \frac{k_{int}}{2}]] \circ X^S[, K_{int}[\frac{k_{int}}{2} + 1 : k_{int}]]$$

which creates the randomly paired interaction terms (where $\circ$ is the Hadamard product) and simulate Y according to the following logistic model:

$$Y|X \sim Bernoulli(logit^{-1}(0.75 + (X^*)^T * \beta_2 + (X^*_{int})^T * \beta_{int_2}))$$

Note that while each feature can be selected at most one time in each of the two lists (main effects and interaction terms), it is possible for a feature to be selected in both lists and to, therefore, be included as a main term as well as in an interaction term.

4. **Logistic model with non-linearity in β**

$$P(K = \kappa) = \begin{cases} \frac{k!(J-k)!}{J!}, & \text{if } \kappa \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}$$

which randomly selects the k features associated with Y without replacement. The subset of X associated with Y is created:

$$X^* = X^s[, K]$$

and $Y|X$ is simulated using the following logistic model with non-linearity in β:

$$Y|X \sim Bernoulli(logit^{-1}(-0.5 + |(X^*)^T * \beta)|)$$

Simulated Joint Distributions of X and $Y|X$

Table 4.3 below summarizes the five simulation combinations given simulations of X and $Y|X$ described above:

X	Y	Description
X_1	Y_1	Main terms logistic model with uncorrelated X
X_2	Y_1	Main terms logistic model with correlated X
X_2	Y_2	Interaction terms logistic model with correlated X
X_2	Y_3	Main and interaction terms logistic model with correlated X
X_2	Y_4	Nonlinear in β logistic model with correlated X

Table 4.3: DGP Combinations of X and Y

Feature Selection Methods

In this section, we describe the different methods applied to the simulated data to select candidate lists of features deemed to be associated with the outcome. The simplest option, logistic regression, is applied using two decision rules: BH adjusted p-values at the α = 5% level and unadjusted p-values at the α = 5% level. For the ensemble method, there are various decisions that can affect feature selection results, often similarly related to the trade-off between sensitivity and specificity. The ones we examined will be described and compared in detail below but generally relate to:

1. Which λ value to use as the penalization parameter in the LASSO.
2. How to determine the cutoff point for the top features according to the concordance of the regression methods.
3. How to determine the set of selected features by Random Forests.

Ensemble Feature Selection Methods

As a reminder, broadly the ensemble feature selection approach involves fitting three models and then combining the lists of features to get a final list of selected features.

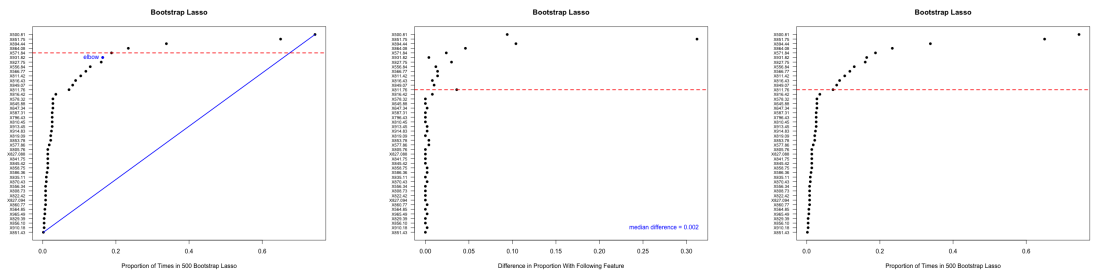
- First, a univariate linear regression model is fit where, for each feature j, the log abundance of X_j is regressed on the case-control status (Y) and features are ranked based on their nominal p-values from the t-test for the case-control regression coefficient.
- Second, a LASSO model is fit where the case-control outcome Y is regressed on all J features over 500 bootstrapped iterations. Features are then ranked by the proportion of the 500 iterations that their associated regression coefficient was non-zero.
- The intersection of the top J^* features from the two regression methods are selected.
- Finally, a Random Forests model is trained on X to predict Y and the top R^* features according to the mean decrease in the Gini index variable importance are added to the previous list to get the final list of selected features S.

Within this broad framework, there are several decisions that must be made. We begin by considering options for the first of the three decisions listed above: which λ value to use for the bootstrap LASSO. There are, in theory, an infinite number of options for the value of the tuning parameter, which can seem unwieldy. To help us make our decision, we will apply the *cv.glmnet*() function from the glmnet package, which, using v-fold cross-validation, fits the LASSO model for a range of λ values and then calculates the cross-validated mean-squared error on the hold-out set [29]. We then use two of the λ values that it returns: λ_{min}, which is the λ value which minimized the cross-validated mean-squared error across the training and testing splits, and λ_{1se}, which is the largest λ

value assessed that had a mean-squared error within one standard error of that achieved by λ_{min}. Note that $\lambda_{1se} \geq \lambda_{min}$, by definition, and so λ_{1se} penalizes $|\beta|_1$ at least as much as λ_{min}, resulting in a model with at most as many features as the model fit with λ_{min}. In this way, we have an automated way to select a couple candidate λ values, one option more conservative in terms of over-fitting than the other.

The second consideration is how to determine the cutoff value, J^*, for taking the intersection of the top J^* features according to the linear regression rankings and the top J^* features according to the bootstrap LASSO rankings. As a reminder, for linear regression, the features are ranked according to the nominal p-value associated with the t-test for the regression coefficient associated with the case-control variables, and for the bootstrap LASSO, the features are ranked according to the proportion of the 500 bootstrap iterations that their respective regression coefficients were non-zero (the proportion of times they were kept in the model). Until now, there have been various ad hoc implementations including taking an a priori decided top percentage of the features (say 25%), finding local maxima in the concordance plot, or using the shape of the lasso variable importance plot. The first option is sub-optimal because is not data adaptive and may result in an arbitrary cutoff point. In an automated application, the second option is difficult to implement as creating a robust rule that would mimic user decision making to pick a local maxima is very difficult given the range of contours the plot can have. So we focus our attention on comparing different ways of using the LASSO variable importance plot to also help in deciding this cutoff point. This seems a natural extension of its use, as a property and useful quality of the LASSO is in selecting a parsimonious set of predictive features (according to its criteria), and therefore results in clearer separation between the set it deems to be predictive and not than other possible options. We will notate the two options for selecting J^* as J^*_E using the elbow method and J^*_D using the difference method (as described below).

The first option for using the LASSO variable importance plot is by taking all the features that are above the elbow in the plot. The elbow of the plot is calculated in the following manner. For the rank-sorted LASSO variable importance plot (which is monotonically decreasing), one first draws a line segment connecting the first and last points. The elbow of the plot is then the observed point which maximizes the distance to that line segment. An example of this is shown in Figure 4.5a. Here, the elbow is calculated to be at the 6^{th} feature and therefore we would set $J^*_E = 5$ and take the intersection of top 5 features from the regression lists (cutoff shown with the red line). While this seems straightforward, in practice, we see that it sometimes results in lists that are too restrictive. In particular, it can be the case that while there are several features that are truly associated with the outcome and stand out according to the LASSO variable importance plot, if the plot shows multiple breaks between these features, then the elbow may be computed too high up the list. To try to allow for the option of having multiple groups of predictive features at different levels, we also implemented a decision rule based on the differences in the proportions of times a features was included between adjacent features after ranking. We can define the proportion of the 500 bootstrap iterations that



(a) Calculating Elbow Cutoff (b) Calculating Differences (c) Difference Cutoff

Figure 4.5: a) Example variable importance plot from the bootstrapped LASSO model fits. The adduct features are sorted by the proportion of iterations that they were included in the model and shown on the y-axis. The x-axis shows the corresponding proportion of the 500 iterations that the feature was included in the LASSO model. Blue line shows the line segment used to determine the elbow of the plot (blue point). Red line indicates the cutoff for the top features selected according to the elbow method. b) Calculated adjacent differences from plot a). Red line shows lowest ranked adduct value where adjacent difference > 4× the median difference. c) Same variable importance plot from a) but this time red line indicates cutoff using difference method (from b).

feature j was included in the LASSO model as p_j and the rank sorted list of p_j values as p^r, where r is the rank of the p_j value and a rank of 1 is the most highly ranked feature. We then denote $diff_r$ = as the calculate the pairwise differences of the sorted values so that $diff_1 = p^1 - p^2, diff_2 = p^2 - p^3, ..., diff_J = 0$. We then define our cutoff J_D^* as the largest rank r for which $diff_r$ is greater than four times the median of all $diff_r$ (med_{diff}). Namely,

$$J_D^* = \arg\max_r r \times \mathbb{1}(diff_r \geq 4 \times med_{diff})$$

A factor of four was selected as it was a large enough scaling that occurred reliably across simulations. As this method runs the risk of returning a large number of features, a limit of $\frac{1}{3} \times J$ was enforced.

To illustrate this, we can use the same variable importance plot example from Figure 4.5a and calculate the differences (shown in 4.5b). In Figure 4.5b, for each adduct on the y-axis, the difference in p to the next highest ranked adduct ($diff_r$) is shown on the x-axis. As we can see, most of these values are very very small. We compute the median of these differences (med_{diff}) and set our cutoff J_D^* as the largest ranking for which the adjacent difference $diff_r$ is ≥ 4× the median value of these differences.

Figure 4.5b shows this cutoff (J_D^*) with the red line and Figure 4.5c shows the cutoff applied to the original variable importance plot from Figure 4.5a. In this way, we would end up selecting not only the features originally selected with the elbow method, but the additional set that may also be associated, albeit not as strongly as the first group. There is the possibility that there is little to no signal in the data (or the relationships are

too complex) and therefore LASSO selects the features relatively equally resulting in a potentially very long list of selected features according to either of the graphical methods described. To try to avoid selecting a very long list of features, we set a cutoff of 50% of the total features. This tuning parameter may need to be adjusted to a reasonable number based on the number of total features in the dataset.

The third and final option that we discuss is the method for selecting features according to Random Forests. In the past, this has been perhaps the most ad hoc component of the ensemble method. It relied on visually inspecting Random Forests graphs and making a decision based on how important features seemed across different random seeds. This was done by looking for separations in the rank sorted values in the variable importance plot according to the mean decrease in the Gini index. These breaks create a separation between a group of more predictive and less predictive features. The first issue is that often, especially if the signal to noise ratio in the data is relatively weak or you have correlated features, you can have different results depending on the random seeds you use, necessitating evaluation on a larger number of seeds. The second issue is that people can be prone to different biases in making decisions based on visual graphics, especially when there are multiple breaks to different degrees in the plot. So, similar to the reasoning in creating an automated decision rule in the regression case, we would want the same for Random Forests. The first option was to try to automate the process in the spirit of what a visually inspecting user would be trying to do: make a decision for each of the Random Forests fits across different seeds and then combine the information together in some type of voting system. We therefore took the list of selected features using the elbow of the variable importance plot in each Random Forests fit across 25 random seeds, and then kept features that were selected in at least 24 of the iterations. The number of iterations was set to 25 as this was large enough to produce robust results. Increasing the number of iterations further did not seem to change results at the cost of computation time, though the user is welcome to do so if desired. This method assumes that predictive features will correctly be identified in the vast majority of the Random Forests variable importance plots and we will hereafter refer to as the "plots" option for Random Forests. To compare this to a potentially more lenient approach, we also looked at taking the set of features using the elbow on the mean of the variable importance metric (mean decrease in the Gini index) across the 25 random seeds. In this implementation, consistently highly ranked features (even if they were not high enough to be selected with the elbow method) could still be chosen. It also used the continuous nature of the variable importance metric rather mapping it to a binary variable, as in the first case. This approach we refer to as the "mean" option for Random Forests.

For ease of visualization, we will first compare performance of options for the first and third decision points to each other as well as logistic regression and then compare the performance of the ensemble method varying the second decision point (how to select the cutoff for the top set of features according to the regression methods). As the results from a logistic regression model fit are continuous (i.e., p-values), we needed to have a decision rule to map the p-values to a binary decision rule. Standard practice in hypothesis testing

would be to adjust for multiple testing and then use a 5% type I error rate cutoff. To that end, the first logistic method we evaluated was to adjust for the false discovery rate (FDR) using Benjamini-Hochberg (BH) correction and then use an adjusted p-value cutoff of 0.05. In addition, we also tried a more lenient approach where the p-values were not adjusted for multiple testing and the same 0.05 cutoff was used. A detailed summary of the implemented methods and various options can be found in the section below.

Methods Summary

Given a matrix X with n observations of J continuous features, as well as a binary outcome vector Y, we then have:

1. Univariate Logistic Regression
    - For each feature X_j, a logistic regression model is fit where the outcome Y is regressed on the log feature value, i.e., $P(Y = 1|X_j) = (1 + e^{-(\beta * X_j)})^{-1}$
    - Nominal p-values associated with the t-test performed for the X_j β coefficients are returned and another list is created with them adjusted for the FDR (BH correction)
    - **Logistic 1**: Features with adjusted p-values <= 0.05 are selected
    - **Logistic 2**: Features with unadjusted p-values <= 0.05 are selected
2. Ensemble Method
    a) Linear Regression
        - For each feature X_j, a linear regression model is fit where the feature abundance is regressed on the outcome Y, i.e., $E[X_j|Y] = \beta_0 + \beta_1 * Y$
        - The features are ranked according to the nominal p-value associated with the t-test for the β_1 coefficient and the ranked list of features we denote as $ranked_{LR}$, where the top feature is first.
    b) Bootstrapped Penalized Regression (LASSO)
        - For each of 500 bootstrapped iterations, a LASSO model is fit where the outcome is regressed on all the J logged features: $P(Y = 1|X) = (1+e^{-(X^T * \boldsymbol{\beta})})^{-1}$, subject to the constraint $\|\boldsymbol{\beta}\|_1 < \lambda$
        - Options for choosing λ using cross-validated glmnet:
            - $\lambda_{\mathbf{min}}$: From the range of λ values evaluated by glmnet, the selected λ is the value which minimized the mean squared error.
            - $\lambda_{\mathbf{1se}}$: From the range of λ values evaluated by glmnet, the selected λ is the largest value such that the cross-validated error is within one standard error of that which is achieved with the λ_{min}.

- The features are ranked in decreasing order according to the proportion of the 500 bootstrapped iterations that their corresponding β coefficient in the model was non-zero and these values are plotted against an index vector that ranges from 1 to J. The ranked list of feature we denote as $ranked_{LASSO}$
- Options for selecting the cutoff point for the regression features J^*
  - **Elbow**: Calculate the elbow of the curve and take the rank of the feature above it. (Using J_E^*)
  - **Difference**: Calculate the pairwise differences of adjacent values in the LASSO variable importance plot and set the cutoff to J_D^* as discussed above.

c) Concordance of Regression Methods

- If the cutoff value J^* determined by the bootstrapped LASSO $\leq \frac{1}{2}J$, then the set of selected features from the regression methods S_{reg} is defined as the intersection of the top J^* features according to linear regression and the top J^* features according to the bootstrapped LASSO.
- If $J^* > \frac{1}{2}J$, then an empty set is returned (no features are selected according to the regression methods. This is only really relevant for the elbow case, since a maximum number of features was already enforced on the difference approach.

d) Random Forests

- For each of 25 seeds selected at random, a Random Forests model is trained on all the features in X to predict the outcome Y.
- Options to select features based on Random Forests:
  - **Mean**: For each feature, the average mean decrease in the Gini index is calculated across the 25 iterations and plotted according to their ranking. The elbow of the plot is determined and all features above that elbow are selected as the set S_{RF}.
  - **Elbow**: For each of the 25 Random Forests models, the elbow is calculated on a rank sorted variable importance plot using the mean decrease in the Gini index and features above the elbow are selected for each of the 25 plots. Features that are selected in at least 24 of the iterations are returned as the final list S_{RF}.
- In either case, if the number of features selected by Random Forests $> 0.5 \times J$, then no features are selected.

e) Final list selection

- The union of the selected features lists (one list from the concordance of the corresponding regression methods and one list from Random Forests) is taken as the final list of selected features S, i.e., $S = S_{reg} \cup S_{RF}$

For each of 500 iterations, the relevant X and $Y|X$ combinations were simulated (from Table 4.3) and each of the ten methods was used to select features (the two logistic methods and the ensemble method with eight specified option combinations). Across simulation runs, the magnitude of the regression coefficients (β values), number of features associated with the outcome (k values), sample size (n), and marginal distribution family of X were varied to examine their effect on method performance.

Evaluation and Visualization

True positive and true negative rates grouped by the different DPGs and feature selection methods are calculated and summarized graphically. Scatterplots are used to compare mean TNR and mean TPR rates of different methods under a given DGP, as well as compare how this performance changes across different DGPs.

Simulation Results

Figures 4.7 and 4.8 show the simulation results for a sample size of $n = 200$ (100 cases and 100 controls) and $n = 400$ (200 cases and 200 controls), respectively. Each of the panels (a through i) shows results for running the simulation with k features associated with the outcome and a certain scaling of the regression coefficients (β). Within a panel, there are 15 scatterplots showing TPR (y-axis) versus TNR (x-axis) values for the 15 combinations of the marginal distributions of X on the rows (labeled along the right side) and distributions of X and $Y|X$ along the columns (labeled along the top), as summarized in Table 4.3.

Figure 4.6 shows results for the simulation run with sample size $n = 200$, $\beta = 100\% \times \beta$, $k = 12$, and $k_{int} = 6$. Each of the fifteen plot shows mean TPR (y-axis) versus mean TNR (x-axis) for each of the feature selection methods. The marginal distribution of X is indicated along the right side and the joint distribution of X and $Y|X$ is indicated along the top. The feature selection method and/or options used are indicated by the legend. Top performing features would have both high TPR and high TNR (top right of the plot). In this instance, for example, we can see that results seem relatively stable across marginal disitrbutions of X and that the ensemble methods outperform the logistic for $X2, Y1$ and $X2, Y3$ joint distributions. For $X2, Y2$ and $X2, Y4$, the logistic methods are hardly able to detect anything, whereas the ensemble methods are at a cost of their mean TNR. This plot is an example shown for clarity of one of the fifteen panels in Figures 4.7 and 4.8, which each show performance for a given sample size ($n = 200$ and $n = 400$, respectively), when varying the magnitude of β values, k, and k_{int}. We now compare performance across these various settings.

In the case of $n = 200$ (Figure 4.7), we can see that generally, and perhaps unsurprisingly, the various ensemble method implementations perform more similarly to each other than the logistic methods. In addition, in particular in the $X2, Y1$ and $X3, Y3$ scenarios, the ensemble methods seem to perform at least as well as the logistic methods. Across the board, the ensemble methods select a higher proportion of the truly associated features,

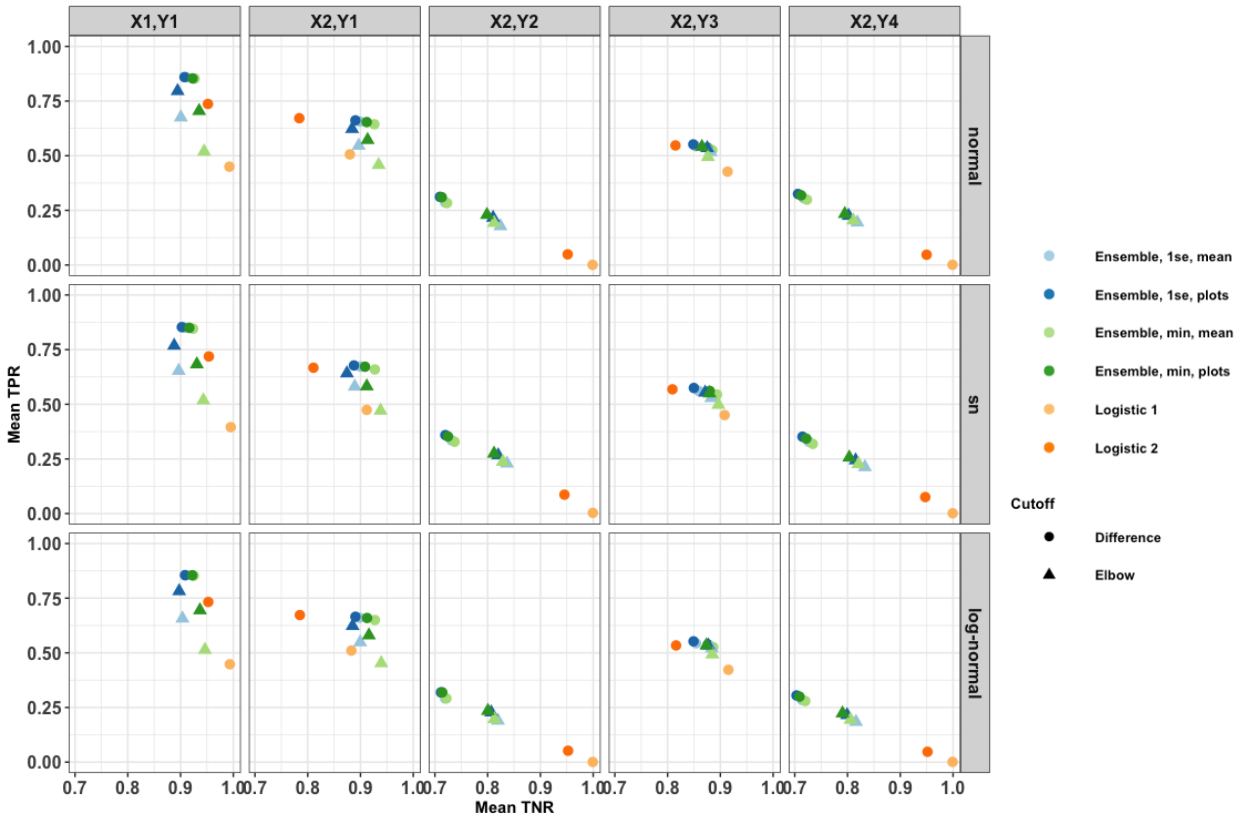
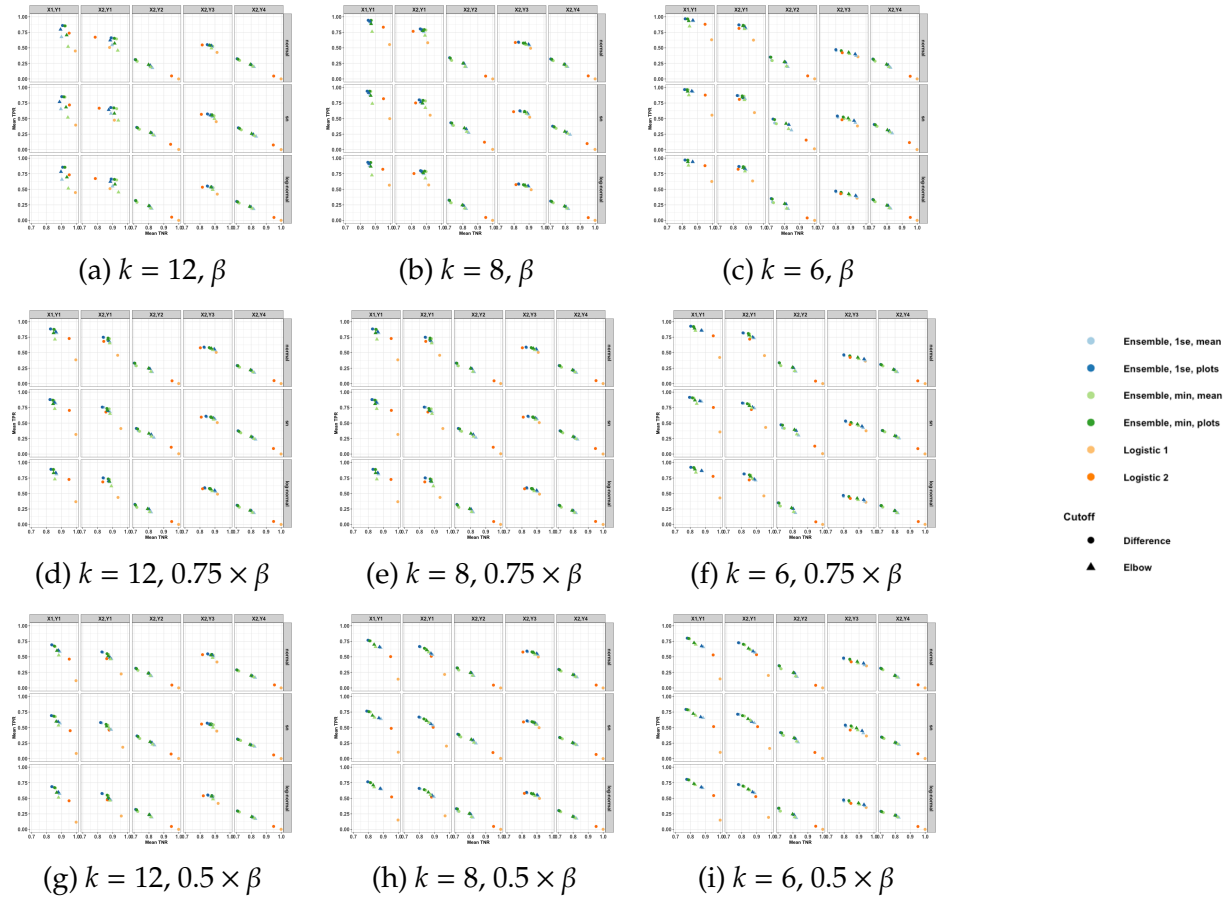


Figure 4.6: Simulation results with sample size $n = 200$, $\beta = 100\% \times \beta$, $k = 12$, and $k_{int} = 6$. Each of the fifteen panels corresponds to a given combination of the marginal distribution of X and distributions of X and $Y|X$, as indicated by the titles above the top row and along the right side (see Table 4.3). Mean TPR (y-axis) is plotted against mean TNR (x-axis) grouped by the method used for feature selection. Points are colored by method (BH adjusted logistic regression (light orange), unadjusted logistic regression (dark orange), ensemble method using λ_{1se} vs. λ_{min} (blue and green respectively) and mean RF vs. elbow RF (light and dark respectively) and the different cutoff point methods used for the regression features in the ensemble method is indicated by a circle for the difference method and a triangle for the elbow method.



(a) $k = 12, \beta$ (b) $k = 8, \beta$ (c) $k = 6, \beta$

(d) $k = 12, 0.75 \times \beta$ (e) $k = 8, 0.75 \times \beta$ (f) $k = 6, 0.75 \times \beta$

(g) $k = 12, 0.5 \times \beta$ (h) $k = 8, 0.5 \times \beta$ (i) $k = 6, 0.5 \times \beta$

Figure 4.7: Simulation results with sample size $n = 200$. In each figure, each of the fifteen panels corresponds to a given combination of the marginal distribution of X and distributions of X and $Y|X$, as indicated by the titles above the top row and along the right side (see Table 4.3). Across the columns, the number of features k and k_{int} associated with the outcome was varied (left: $k = 12, k_{int} = 6$, middle: $k = 8, k_{int} = 4$, right: $k = 6, k_{int} = 4$). Across the rows, the magnitude of the regression coefficients (β) was varied (top: 100% β, middle: 75% β, bottom: 50% β). So the panel with the most associated features with the strongest signal is in the top left (a) and the panel with the fewest associated features with the weakest signal is in the bottom right (i). Within each panel, mean TPR (y-axis) is plotted against mean TNR (x-axis) grouped by the method used for feature selection. Points are colored by method (BH adjusted logistic regression (light orange), unadjusted logistic regression (dark orange), ensemble method using λ_{1se} vs. λ_{min} (blue and green respectively) and mean RF vs. elbow RF (light and dark respectively) and the different cutoff point methods used for the regression features in the ensemble method is indicated by a circle for the difference method and a triangle for the elbow method.

most noticeably for $X2, Y3$ and $X2, Y4$, though this comes with the trade-off of a lower mean TNR. Interestingly, we can see that changing the marginal distribution of X between the three options, conditional on $Y|X$, k and k_{int} and the scaling factor in front of β, does not seem to considerably change the results (looking down the rows within a panel). In simulations using 100% of β (panels a, b, and c), TPR values for ensemble methods and the logistic methods are best when $k = 8$, particularly for $X2, Y1$. In general, lowering the amount of signal (fewer features and smaller magnitude of β values) seems to hurt the ability of the logistic methods to detect truly associated features more than the ensemble methods.

While the ensemble methods perform relatively similarly to each other, there do seem to be general patterns across the simulations. The biggest difference among them seems to be due to the cutoff method for selecting the top regression features (circles versus triangles). This can be seen most clearly in the $X2, Y2$ and $X2, Y4$ scenarios, where the triangles have lower mean TPR but higher mean TNR values. While not as extreme, the circles and triangles still seem to cluster with each other in the other DGPs as well. In addition, with the exception of the simplest DGP ($X1, Y1$), the blue circle and blue triangle seem to bookend the spectrum in terms of the trade-off of higher TPR with lower TNR (blue circle) versus higher TNR with lower TPR (blue triangle). The difference in performance profiles between the different ensemble methods becomes more noticeable as the k becomes smaller (looking across a row).

In comparing the previous discussed results with those from the case where $n = 400$ (Figure 4.8), we can see that the increased sample size seemed to improve mean TPR values for both logistic methods and ensemble methods. In contrast, while mean TNR rates seem to improve or at least stay consistent for the ensemble methods with the increased sample size, they actually noticeably worsened in many cases for the logistic methods, creating a larger separation between them and the ensemble methods. Another pattern that emerged with the larger sample size amongst the ensemble methods is the separation between the darker shapes and the lighter ones, in particular for the $X2, Y2$ and $X2, Y3$ DGPs. There, we can clearly see that the darker shapes (using the "plots" option for RF) have higher mean TPR values compared to their lighter (using the "mean" option for RF) counterparts.

These results show that the ensemble methods are less sensitive to perturbations in the underlying marginal distributions of X studied than the logistic methods, both in terms of the correlation structure of X as well as its multivariate distribution. In addition, the ensemble methods were able to select more associated features when $Y|X$ contained interaction terms or was non-linear. These differences widened in many cases when sample size was increased, as the mean TNR for the logistic methods noticeably worsened. While the ensemble methods seemed to perform relatively similarly in many cases, patterns that emerged (particularly in the larger sample size case) indicate that results may be sensitive to which automated method was chosen.

Automated Methods Applied to Lung Cancer Data

We now turn our attention to examining how the various automated feature selection results compare to the previously discussed manual implementation in the lung cancer data. For each of the two sub-populations previously described (histologically confirmed (HC) vs. all controls and adenocarcenoma (LUAD) vs. all controls), each of the eight ensemble feature selection method variations was applied to each subset and the resulting feature lists are compared in Tables 4.4 and 4.5, respectively.

LASSO λ	λ_{1se}				λ_{min}			
LASSO cutoff	difference		elbow		difference		elbow	
RF method	plots	mean	plots	mean	plots	mean	plots	mean
571.84	571.84	571.84	571.84	571.84	571.84	571.84	571.84	571.84
587.31	587.31	587.31	587.31	587.31	587.31	587.31	587.31	587.31
822.42	822.42	822.42	822.42	822.42	822.42	822.42	822.42	822.42
845.42	845.42	845.42	845.42	845.42	845.42	845.42	845.42	845.42
856.10	856.10	856.10	856.10	856.10	856.10	856.10	856.10	856.10
914.83	914.83	914.83	914.83	914.83	914.83	914.83	914.83	914.83
features		566.77		566.77		566.77		566.77
added	578.32	578.32			578.32	578.32	578.32	578.32
		586.36		586.36		586.36		586.36
	647.34	647.34	647.34	647.34	647.34	647.34	647.34	647.34
	808.73	808.73	808.73	808.73	808.73	808.73	808.73	808.73
		827.088		827.088		827.088		827.088
	827.094	827.094	827.094	827.094	827.094	827.094	827.094	827.094
	835.11	835.11		835.11		835.11		835.11
					849.07	849.07	849.07	849.07
	853.78	853.78		853.78		853.78		853.78
		870.43		870.43		870.43		870.43
	910.18	910.18	910.18	910.18		910.18		910.18
Total #								
6	13	17	10	16	11	18	11	18

Table 4.4: Feature selection results on the HC subset of the lung cancer data using each of the eight previously described automated ensemble feature selection methods. A given list is read vertically, corresponding to the options listed above it. For example, the first column corresponds to using the λ_{1se} and the "difference" options for the LASSO and the "plots" option for RF. Features added by each automated approach compared to the manual implementation in Section 4.2 are listed below the double line.

In both the adenocarcenoma and histologically confirmed analyses, the original set of features selected during the manual implementation of the ensemble feature selection

LASSO λ	λ_{1se}				λ_{min}			
LASSO cutoff	difference		elbow		difference		elbow	
RF method	plots	mean	plots	mean	plots	mean	plots	mean
571.84	571.84	571.84	571.84	571.84	571.84	571.84	571.84	571.84
587.31	587.31	587.31	587.31	587.31	587.31	587.31	587.31	587.31
827.088	827.088	827.088	827.088	827.088	827.088	827.088	827.088	827.088
845.42	845.42	845.42	845.42	845.42	845.42	845.42	845.42	845.42
849.07	849.07	849.07	849.07	849.07	849.07	849.07	849.07	849.07
856.10	856.10	856.10	856.10	856.10	856.10	856.10	856.10	856.10
858.75	858.75	858.75	858.75	858.75	858.75	858.75	858.75	858.75
914.83	914.83	914.83	914.83	914.83	914.83	914.83	914.83	914.83
features	566.77	566.77	566.77	566.77		566.77		566.77
added	647.34	647.34	647.34	647.34	647.34	647.34	647.34	647.34
	808.73	808.73		808.73	808.73	808.73	808.73	808.73
		811.42		811.42		811.42		811.42
		811.76		811.76		811.76		811.76
		816.43		816.43		816.43		816.43
	827.094	827.094	827.094	827.094	827.094	827.094	827.094	827.094
		835.11		835.11		835.11		835.11
	853.78	853.78	853.78	853.78	853.78	853.78	853.78	853.78
		870.43		870.43		870.43		870.43
	910.18	910.18	910.18	910.18		910.18		910.18
Total #								
8	14	19	13	19	12	19	12	19

Table 4.5: Feature selection results on the LUAD subset of the lung cancer data using each of the eight previously described automated ensemble feature selection methods. A given list is read vertically, corresponding to the options listed above it. For example, the first column of features correspond to using the λ_{1se} and the "difference" options for the LASSO and the "plots" option for RF. Features added by each automated approach compared to the manual implementation in Section 4.2 are listed below the double line.

process were selected using all eight automated processes. Past that, we can see that the decision that makes the biggest difference in the resulting feature list is the implementation of Random Forests. Using the “mean” implementation results in a much longer list of selected features irrespective of the other factors. In the case of the adenocarcenoma data, it resulted in the same 19 selected features across all four variants. When comparing the relative increase in the number of features selected using the “mean” option versus the “plots” option in Random Forests, we see about a 50% increase in the number of features selected (sometimes more).

In terms of the decisions regarding the regression methods, we see that using λ_{min}

tended to result in selecting the fewest features (assuming the "plots" option of Random Forests was specified) and therefore seems most similar to the original data analysis (three out of four times, it had the fewest additional features selected). In addition, in the case of the histologically controlled analysis (4.4), one feature (849.07) was selected for all the methods using λ_{min} but none of the ones using λ_{1se}. Upon investigation, we saw that 849.07 was added from the list of concordant features using the regression methods and though was always highly ranked by linear regression, only became highly ranked enough in the LASSO ranking to be selected using λ_{min}.

Overall we can see that in the case of the lung cancer data, the resulting selected feature lists vary greatly depending on which automated options were implemented. These results indicate that the "mean" option for RF seems overly sensitive and should be used with caution. There also seems to be evidence in support of using λ_{min} in implementing LASSO, though that difference is much smaller. The "difference" versus "elbow" cutoff for LASSO does not result in a large difference (just one feature) so the results were less sensitive to that option. As the "elbow" option does not require a tuning parameter, though, it would appear preferable given similar performance. One option would be to select features that are shared across all selected feature lists using the eight automated approaches. This would help ensure that selected features are not spurious due to having picked one set of tuning parameters. Or at the least, by implementing a range of options, one can examine how sensitive the resulting feature list is to decisions made in the implementation.

4.4 Discussion

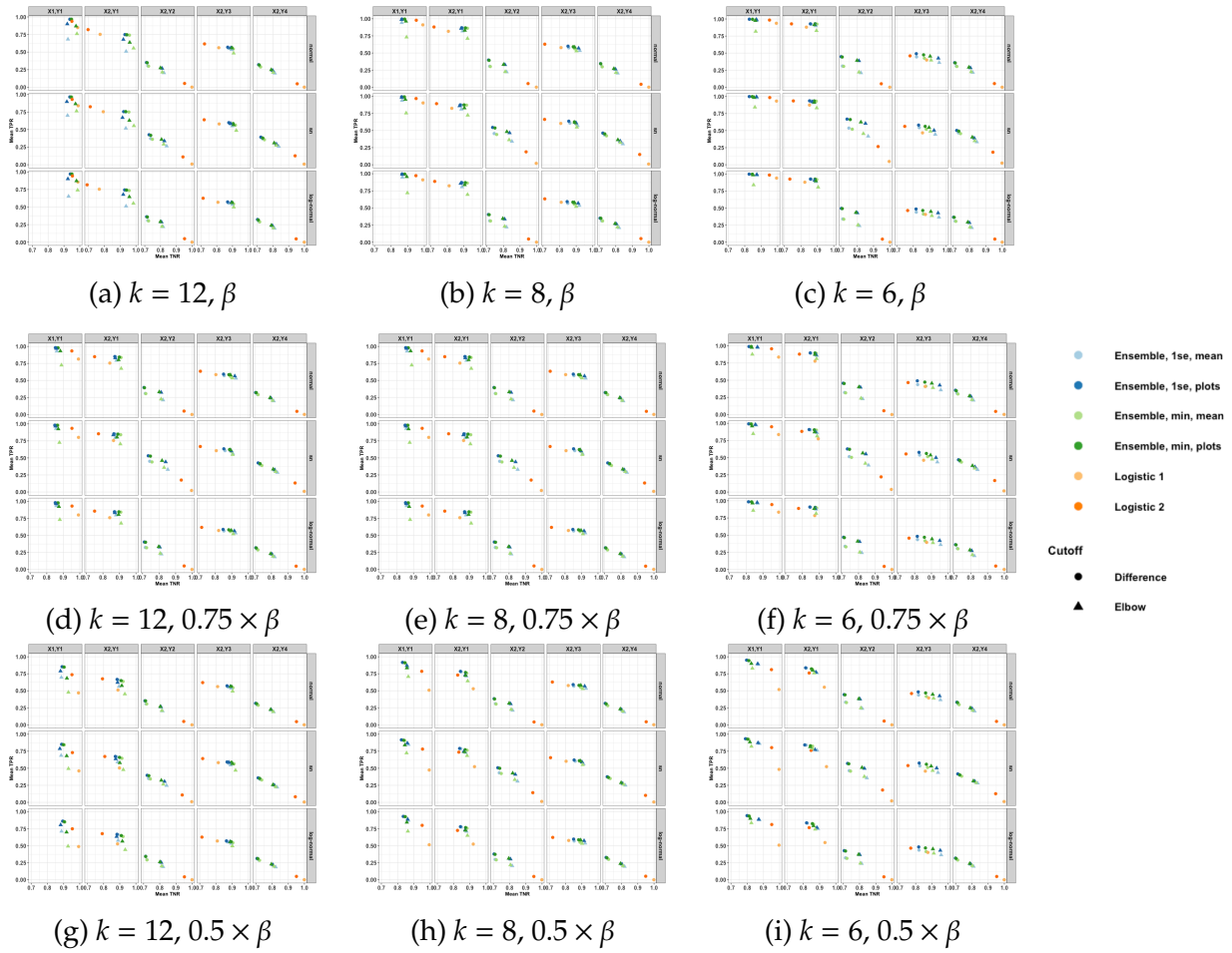
The results from analysis of the lung cancer data highlight interesting associations and potential risk factors for this disease, including exposure to $PM_{2.5}$/PAH, oxidative stress, OCM, and a range of dietary antioxidants. Our ensemble feature selection approach allowed for detection of a broader range of associations, which can help direct future studies seeking to better understand risks for development of lung cancer in never-smokers. We then sought to automate and benchmark this ensemble feature selection approach using simulations and examine the sensitivity of results to the various options studied.

The results from our simulation study were informative, if not completely conclusive as to which automated version is optimal. The poor predictive performance of the logistic regression methods under certain DGPs reinforces the intuition that limiting the analysis to logistic regression will fail to allow for detection of a broader range of associations between predictors and the outcome. Though increasing sample size helped logistic regression in the setting of uncorrelated features (an empirically confirmed unrealistic assumption), the smaller estimated standard errors for the t-tests of the regression coefficients with the larger sample size came at a large cost in TNR. The ensemble methods performed better in both metrics in DGPs that included main effect terms or a combination of main effect and interaction terms (under the more realistic setting of having correlated features). In

addition, our current simulation only had 44 features (a relatively small number in the context of exposomic datasets). One can imagine that applying logistic regression in feature selection analyses with untargeted metabolomics datasets for example (where J could be 2,000 or more features), would make selecting features using a multiple-testing adjusted p-value even more difficult.

As there was a trade-off between TNR and TPR in the different implementations of the ensemble method, it is not easy to give a definitive recommendation on which option to use. One consideration is that using the "elbow" method for determining the cutoff for the regression methods is less sensitive to tuning parameters than the "difference" method (as it does not have any) and therefore may be simpler to apply. In the case of applying the automated methods to the lung cancer data, it had a large difference in the resulting feature list, though that may vary depending on the dataset. Potentially surprisingly, using λ_{min} rather than λ_{1se} tended to result in fewer selected features, which may be a consideration when deciding between the two. In addition, in particular based on the applications to the real data, using the "mean" implementation of Random Forests may be too sensitive, and one should consider the risk of introducing a large number of false negatives. Implementations in the simulations with larger sample size support this conclusion, as the "plots" option seemed to consistently outperform the "mean" option. At the least comparing resulting selected feature lists from implementation of the different automated methods can give the user an idea of the level of sensitivity of their results to these decisions, with aggregating over the lists in some way an option to increase stability.

While varying several options, both in methods as well as the underlying DPG, were examined during our simulation, there are many more factors or variations that someone interested in feature selection in this context may want to consider or perturb. To that end, the effects of different distributions of X or $Y|X$ that may be particularly relevant to the question at hand may be worth examining, as well as implementing variations on the ensemble method using different criteria or safeguards. The options described were meant to show a range of reasonable options but are by no means exhaustive. In addition, it would be worth repeating this simulation learning parameters from different datasets to see if results change. Finally, though applications in epidemiology and related fields are often focused on more traditional approaches, this simulation framework can be utilized to compare more complicated (potentially at the cost of interpretability) methods for feature selection. The hope is that this offers additional insight and a framework from which to continue discussion of examining and validating candidate methods in the context of biomarker discovery.



(a) $k = 12, \beta$ (b) $k = 8, \beta$ (c) $k = 6, \beta$

(d) $k = 12, 0.75 \times \beta$ (e) $k = 8, 0.75 \times \beta$ (f) $k = 6, 0.75 \times \beta$

(g) $k = 12, 0.5 \times \beta$ (h) $k = 8, 0.5 \times \beta$ (i) $k = 6, 0.5 \times \beta$

Figure 4.8: Simulation results with sample size $n = 400$. In each figure, each of the fifteen panels corresponds to a given combination of the marginal distribution of X and distributions of X and $Y|X$, as indicated by the titles above the top row and along the right side (see Table 4.3). Across the columns, the number of features k and k_{int} associated with the outcome was varied (left: $k = 12, k_{int} = 6$, middle: $k = 8, k_{int} = 4$, right: $k = 6, k_{int} = 4$). Across the rows, the magnitude of the regression coefficients (β) was varied (top: 100% β, middle: 75% β, bottom: 50% β). So the panel with the most associated features with the strongest signal is in the top left (a) and the panel with the fewest associated features with the weakest signal is in the bottom right (i). Within each panel, mean TPR (y-axis) is plotted against mean TNR (x-axis) grouped by the method used for feature selection. Points are colored by method (BH adjusted logistic regression (light orange), unadjusted logistic regression (dark orange), ensemble method using λ_{1se} vs. λ_{min} (blue and green respectively) and mean RF vs. elbow RF (light and dark respectively) and the different cutoff point methods used for the regression features in the ensemble method is indicated by a circle for the difference method and a triangle for the elbow method.

Chapter 5

Conclusion and Future Directions

In this dissertation, we discussed various analyses of three different kinds of exposomic data. We performed exploratory data analysis and adapted pre-processing protocols for each dataset, particularly important steps when analyzing this type of data. This helped us understand the variability in each dataset and make more thoughtful decisions, particularly for the integration of multiple datasets. Only after removing as much of the technical variability as possible, were we able to turn our attention to the statistical analysis implemented to try to answer the biological question at hand.

In Chapter 2, we used a prediction framework to evaluate the importance of disregulation of the folate pathway, as measured by a subset of related metabolites in neonatal DBS, in increasing risk for pediatric ALL. This afforded flexibility in the nature of the detectable associations and did not rely on traditional hypothesis testing approaches that were not appropriate for the data at hand. We emphasized the importance of complete splitting of learning and test data from the onset to maintain independence and unbiased performance metrics. Though we reported null results, this is likely due to the lack of biological signal, as evident during EDA, and this general framework can be built on to evaluate various hypotheses in future works.

In Chapter 3, we expanded our study of potential risk factors for our case-control subjects from Chapter 2 by now examining untargeted metabolomics data for both the child DBS and the paired maternal serum samples from during gestation. The untargeted metabolomics data required more complex filtering schemes applied independently and adaptively to each dataset separately. We showed how differences can arise between features in two datasets due to source type, method of collection, storage, and pre-processing and how these need to be considered in results of downstream analyses. In addition, because of challenges in biological interpretability of results due to unreliability of annotations in untargeted metabolomics, we first transformed our features to more stable reaction level features for hypothesis generation. We applied a more flexible ensemble feature selection pipeline to marginal and joint datasets, allowing for the detection of a broader range of biological signals than more traditional approaches.

Chapter 4 applies EDA and pre-processing in the context of untargeted adductomics

data and elaborates on the ensemble feature selection method utilized in Chapter 3 for generating hypotheses related to lung cancer risk in never-smokers. Due to potential effect heterogeneity between sub-types, stratified analyses were performed to further annotate selected features. The second portion of the chapter addresses two gaps regarding application and benchmarking of the ensemble method described. Performance of several automated options of the method are compared to standard hypothesis testing methods and provide additional guidance and support for its merit in this application.

Due to the complex nature of the diseases under study, it is of increasing interest to incorporate information from a variety of biological sources and data types. This allows for integration of biological signals that reflect different aspects of both genetic and exposomic associations, providing a more comprehensive view into the underlying network of biological processes creating risk profiles. Even more so than discussed in Chapter 3, where there was at least a common data type (untargeted metabolomics), integrating different data types will further increase the need for careful examination and pre-processing before joint analyses can be performed. In addition, moving away from traditional hypothesis testing approaches unsuited to the question and the nature of the data generated by these high-throughput technologies can broaden ones ability to detect the signal of interest. In an era where the bottleneck for scientific discovery has becoming decreasingly due to lack of data, but rather the ability to sift through it, these measures can help ensure that these studies achieve a greater potential for answering the important biological questions asked of them.

Bibliography

[1] {R Core Team}. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[2] Behnaz Abiri et al. "Effects of Maternal Diet During Pregnancy on the Risk of Childhood Acute Lymphoblastic Leukemia: A Systematic Review". eng. In: *Nutrition and Cancer* 68.7 (Oct. 2016), pp. 1065–1072. ISSN: 1532-7914. DOI: 10.1080/01635581.2016.1206581.

[3] Giancarlo Aldini et al. "Albumin is the main nucleophilic target of human plasma: a protective role against pro-atherogenic electrophilic reactive carbonyl species?" eng. In: *Chemical Research in Toxicology* 21.4 (Apr. 2008), pp. 824–835. ISSN: 0893-228X. DOI: 10.1021/tx700349r.

[4] Krasimira Aleksandrova et al. "A Prospective Study of the Immune System Activation Biomarker Neopterin and Colorectal Cancer Risk". In: *JNCI Journal of the National Cancer Institute* 107.4 (Apr. 2015), djv010. ISSN: 0027-8874. DOI: 10.1093/jnci/djv010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402364/ (visited on 03/08/2023).

[5] F. E. Alexander et al. "Transplacental chemical exposure and risk of infant leukemia with MLL gene fusion". eng. In: *Cancer Research* 61.6 (Mar. 2001), pp. 2542–2546. ISSN: 0008-5472.

[6] Nerino Allocati et al. "Glutathione transferases: substrates, inihibitors and prodrugs in cancer and neurodegenerative diseases". In: *Oncogenesis* 7.1 (Jan. 2018), p. 8. ISSN: 2157-9024. DOI: 10.1038/s41389-017-0025-3. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5833873/ (visited on 04/17/2024).

[7] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data". eng. In: *Genome Biology* 11.10 (2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106.

[8] Valéria Troncoso Baltar et al. "A structural equation modelling approach to explore the role of B vitamins and immune markers in lung cancer risk". In: *European Journal of Epidemiology* 28.8 (2013). Publisher: Springer, pp. 677–688. ISSN: 0393-2990. URL: https://www.jstor.org/stable/43774879 (visited on 04/04/2024).

[9] Francesco Barone-Adesi et al. "Risk of lung cancer associated with domestic use of coal in Xuanwei, China: retrospective cohort study". eng. In: *BMJ (Clinical research ed.)* 345 (Aug. 2012), e5414. ISSN: 1756-1833. DOI: 10.1136/bmj.e5414.

[10] Jessica L. Barrington-Trimis et al. "Trends in Childhood Leukemia Incidence Over Two Decades from 1992–2013". In: *International journal of cancer* 140.5 (Mar. 2017), pp. 1000–1008. ISSN: 0020-7136. DOI: 10.1002/ijc.30487. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5550103/ (visited on 03/08/2023).

[11] Rebecca Barter and Bin Yu. *superheat: A Graphical Tool for Exploring Complex Datasets Using Heatmaps}*. 2023.

[12] Subit Barua and Mohammed A. Junaid. "Lifestyle, pregnancy and epigenetic effects". eng. In: *Epigenomics* 7.1 (2015), pp. 85–102. ISSN: 1750-192X. DOI: 10.2217/epi.14.71.

[13] B. C. Blount et al. "Folate deficiency causes uracil misincorporation into human DNA and chromosome breakage: implications for cancer and neuronal damage". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 94.7 (Apr. 1997), pp. 3290–3295. ISSN: 0027-8424. DOI: 10.1073/pnas.94.7.3290.

[14] Leo Breiman. "Random Forests". en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324 (visited on 03/27/2024).

[15] James P. Burkett and Gary W. Miller. "Using the exposome to understand environmental contributors to psychiatric disorders". en. In: *Neuropsychopharmacology* 46.1 (Jan. 2021). Publisher: Nature Publishing Group, pp. 263–264. ISSN: 1740-634X. DOI: 10.1038/s41386-020-00851-0. URL: https://www.nature.com/articles/s41386-020-00851-0 (visited on 03/27/2024).

[16] Kate Campbell et al. "Methionine Metabolism Alters Oxidative Stress Resistance via the Pentose Phosphate Pathway". In: *Antioxidants & Redox Signaling* 24.10 (Apr. 2016), pp. 543–547. ISSN: 1523-0864. DOI: 10.1089/ars.2015.6516. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4827311/ (visited on 04/04/2024).

[17] *Cancer Today*. URL: https://gco.iarc.who.int/today/ (visited on 04/04/2024).

[18] Caroline Carrico et al. "Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting". eng. In: *Journal of Agricultural, Biological, and Environmental Statistics* 20.1 (Mar. 2015), pp. 100–120. ISSN: 1085-7117. DOI: 10.1007/s13253-014-0180-3.

[19] T. A. Chiang, P. F. Wu, and Y. C. Ko. "Identification of carcinogens in cooking oil fumes". eng. In: *Environmental Research* 81.1 (July 1999), pp. 18–22. ISSN: 0013-9351. DOI: 10.1006/enrs.1998.3876.

[20] Anand P. Chokkalingam et al. "Blood levels of folate at birth and risk of childhood leukemia". eng. In: *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 22.6 (June 2013), pp. 1088–1094. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-12-1438.

[21] Michael Cole and Davide Risso. *scone: Single Cell Overview of Normalized Expression data version 1.14.0 from Bioconductor*. en. 2022. URL: https://rdrr.io/bioc/scone/ (visited on 03/08/2023).

[22] Sonia Dagnino et al. "Agnostic Cys34-albumin adductomics and DNA methylation: Implication of N-acetylcysteine in lung carcinogenesis years before diagnosis". eng. In: *International Journal of Cancer* 146.12 (June 2020), pp. 3294–3303. ISSN: 1097-0215. DOI: 10.1002/ijc.32680.

[23] Isabella Dalle-Donne et al. "Biomarkers of oxidative damage in human disease". eng. In: *Clinical Chemistry* 52.4 (Apr. 2006), pp. 601–623. ISSN: 0009-9147. DOI: 10.1373/clinchem.2005.061408.

[24] Shanshan Duan and Michele Pagano. "Ubiquitin ligases in cancer: functions and clinical potentials". In: *Cell chemical biology* 28.7 (July 2021), pp. 918–933. ISSN: 2451-9456. DOI: 10.1016/j.chembiol.2021.04.008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8286310/ (visited on 04/17/2024).

[25] Gregory S. Ducker and Joshua D. Rabinowitz. "One-Carbon Metabolism in Health and Disease". eng. In: *Cell Metabolism* 25.1 (Jan. 2017), pp. 27–42. ISSN: 1932-7420. DOI: 10.1016/j.cmet.2016.08.009.

[26] L. Ehrenberg, F. Granath, and M. Törnqvist. "Macromolecule adducts as biomarkers of exposure to environmental mutagens in human populations". eng. In: *Environmental Health Perspectives* 104 Suppl 3.Suppl 3 (May 1996), pp. 423–428. ISSN: 0091-6765. DOI: 10.1289/ehp.96104s3423.

[27] Meira Epplein et al. "Smoking-adjusted lung cancer incidence among Asian-Americans (United States)". eng. In: *Cancer causes & control: CCC* 16.9 (Nov. 2005), pp. 1085–1090. ISSN: 0957-5243. DOI: 10.1007/s10552-005-0330-6.

[28] *Exposome and Exposomics | NIOSH | CDC*. en-us. Dec. 2022. URL: https://archive.cdc.gov/www_cdc_gov/niosh/topics/exposome/default.html (visited on 03/26/2024).

[29] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". en. In: *Journal of Statistical Software* 33 (Feb. 2010), pp. 1–22. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: https://doi.org/10.18637/jss.v033.i01 (visited on 04/08/2024).

[30] Y. T. Gao et al. "Lung cancer among Chinese women". eng. In: *International Journal of Cancer* 40.5 (Nov. 1987), pp. 604–609. ISSN: 0020-7136. DOI: 10.1002/ijc.2910400505.

[31] *Genetic Disorders*. en. URL: https://www.genome.gov/For-Patients-and-Families/Genetic-Disorders (visited on 03/27/2024).

[32] Tusharkanti Ghosh et al. "Predictive Modeling for Metabolomics Data". In: *Methods in molecular biology (Clifton, N.J.)* 2104 (2020), pp. 313–336. ISSN: 1064-3745. DOI: 10.1007/978-1-0716-0239-3_16. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7423323/ (visited on 03/25/2024).

[33] Mel F. Greaves et al. "Leukemia in twins: lessons in natural history". eng. In: *Blood* 102.7 (Oct. 2003), pp. 2321–2333. ISSN: 0006-4971. DOI: 10.1182/blood-2002-12-3817.

[34] Hasmik Grigoryan et al. "An adductomics pipeline for untargeted analysis of modifications to Cys34 of human serum albumin". In: *Analytical chemistry* 88.21 (Nov. 2016), pp. 10504–10512. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.6b02553. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5555296/ (visited on 04/04/2024).

[35] Hasmik Grigoryan et al. "Cys34 Adductomics Links Colorectal Cancer with the Gut Microbiota and Redox Biology". In: *Cancer research* 79.23 (Dec. 2019), pp. 6024–6031. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-19-1529. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6891211/ (visited on 04/04/2024).

[36] Hasmik Grigoryan et al. "Extending the HSA-Cys34-Adductomics Pipeline to Modifications at Lys525". eng. In: *Chemical Research in Toxicology* 34.12 (Dec. 2021), pp. 2549–2557. ISSN: 1520-5010. DOI: 10.1021/acs.chemrestox.1c00311.

[37] Hasmik Grigoryan et al. "HSA Adductomics Reveals Sex Differences in NHL Incidence and Possible Involvement of Microbial Translocation". In: *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 32.9 (Sept. 2023), pp. 1217–1226. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-23-0231. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10529301/ (visited on 04/04/2024).

[38] B. Gruhn et al. "Prenatal origin of childhood acute lymphoblastic leukemia, association with birth weight and hyperdiploidy". eng. In: *Leukemia* 22.9 (Sept. 2008), pp. 1692–1697. ISSN: 1476-5551. DOI: 10.1038/leu.2008.152.

[39] Khalid M Hamed et al. "Overview of Methotrexate Toxicity: A Comprehensive Literature Review". In: *Cureus* 14.9 (), e29518. ISSN: 2168-8184. DOI: 10.7759/cureus.29518. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9595261/ (visited on 03/08/2023).

[40] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer, 2009. DOI: 10.1007/978-0-387-84858-7. URL: http://link.springer.com/10.1007/978-0-387-84858-7 (visited on 03/21/2024).

[41] William Hill et al. "Lung adenocarcinoma promotion by air pollutants". en. In: *Nature* 616.7955 (Apr. 2023). Publisher: Nature Publishing Group, pp. 159–167. ISSN: 1476-4687. DOI: 10.1038/s41586-023-05874-3. URL: https://www.nature.com/articles/s41586-023-05874-3 (visited on 04/04/2024).

[42] Kin-Fai Ho et al. "Effects of polycyclic aromatic compounds in fine particulate matter generated from household coal combustion on response to EGFR mutations in vitro". eng. In: *Environmental Pollution (Barking, Essex: 1987)* 218 (Nov. 2016), pp. 1262–1269. ISSN: 1873-6424. DOI: 10.1016/j.envpol.2016.08.084.

[43] James C. Hogg and Stephen van Eeden. "Pulmonary and systemic response to atmospheric pollution". eng. In: *Respirology (Carlton, Vic.)* 14.3 (Apr. 2009), pp. 336–346. ISSN: 1440-1843. DOI: 10.1111/j.1440-1843.2009.01497.x.

[44] H. Dean Hosgood et al. "Driver mutations among never smoking female lung cancer tissues in China identify unique EGFR and KRAS mutation pattern associated with household coal burning". eng. In: *Respiratory Medicine* 107.11 (Nov. 2013), pp. 1755–1762. ISSN: 1532-3064. DOI: 10.1016/j.rmed.2013.08.018.

[45] Mengna Huang et al. "Maternal Metabolome in Pregnancy and Childhood Asthma or Recurrent Wheeze in the Vitamin D Antenatal Asthma Reduction Trial". In: *Metabolites* 11.2 (Jan. 2021), p. 65. ISSN: 2218-1989. DOI: 10.3390/metabo11020065. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7910853/ (visited on 04/19/2024).

[46] Wolfgang Huber et al. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". eng. In: *Bioinformatics (Oxford, England)* 18 Suppl 1 (2002), S96–104. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/18.suppl_1.s96.

[47] IARC. *Diesel and Gasoline Engine Exhausts and Some Nitroarenes*. Vol. 105. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. IARC Monogr Eval Carcinog Risks Hum, 2014. ISBN: 978-92-832-1328-4.

[48] Hiroto Inaba and Charles G. Mullighan. "Pediatric acute lymphoblastic leukemia". eng. In: *Haematologica* 105.11 (Nov. 2020), pp. 2524–2539. ISSN: 1592-8721. DOI: 10.3324/haematol.2020.247031.

[49] Maria Jackson et al. "The genetic basis of disease". In: *Essays in Biochemistry* 62.5 (Dec. 2018), pp. 643–723. ISSN: 0071-1365. DOI: 10.1042/EBC20170053. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6279436/ (visited on 03/27/2024).

[50] Bernd O. Keller et al. "Interferences and contaminants encountered in modern mass spectrometry". eng. In: *Analytica Chimica Acta* 627.1 (Oct. 2008), pp. 71–81. ISSN: 1873-4324. DOI: 10.1016/j.aca.2008.04.043.

[51] Timothy J. Key et al. "The effect of diet on risk of cancer". eng. In: *Lancet (London, England)* 360.9336 (Sept. 2002), pp. 861–868. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(02)09958-0.

[52] Marin Kharrazi et al. "California Very Preterm Birth Study: design and characteristics of the population- and biospecimen bank-based nested case-control study". In: *Paediatric and perinatal epidemiology* 26.3 (May 2012), pp. 250–263. ISSN: 0269-5022. DOI: 10.1111/j.1365-3016.2011.01252.x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3536480/ (visited on 04/12/2024).

[53] Hee-Jung Kim et al. "ROSICS: CHEMISTRY AND PROTEOMICS OF CYSTEINE MODIFICATIONS IN REDOX BIOLOGY". In: *Mass Spectrometry Reviews* 34.2 (Apr. 2015), pp. 184–208. ISSN: 0277-7037. DOI: 10.1002/mas.21430. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340047/ (visited on 04/04/2024).

[54] Ilan J. N. Koppen, Frederik J. R. Hermans, and Gertjan J. L. Kaspers. "Folate related gene polymorphisms and susceptibility to develop childhood acute lymphoblastic leukaemia". eng. In: *British Journal of Haematology* 148.1 (Jan. 2010), pp. 3–14. ISSN: 1365-2141. DOI: 10.1111/j.1365-2141.2009.07898.x.

[55] Henry Kurniawan, Takumi Kobayashi, and Dirk Brenner. "The emerging role of one-carbon metabolism in T cells". In: *Current Opinion in Biotechnology*. Systems Biology Nanobiotechnology 68 (Apr. 2021), pp. 193–201. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2020.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0958166920301841 (visited on 03/28/2024).

[56] Yvonne Lamers. "Folate recommendations for pregnancy, lactation, and infancy". eng. In: *Annals of Nutrition & Metabolism* 59.1 (2011), pp. 32–37. ISSN: 1421-9697. DOI: 10.1159/000332073.

[57] Qing Lan et al. "Oxidative damage-related genes AKR1C3 and OGG1 modulate risks for lung cancer due to exposure to PAH-rich coal combustion emissions". eng. In: *Carcinogenesis* 25.11 (Nov. 2004), pp. 2177–2181. ISSN: 0143-3334. DOI: 10.1093/carcin/bgh240.

[58] Kian Boon Lee et al. "Association between Metabolites and the Risk of Lung Cancer: A Systematic Literature Review and Meta-Analysis of Observational Studies". In: *Metabolites* 10.9 (Sept. 2020), p. 362. ISSN: 2218-1989. DOI: 10.3390/metabo10090362. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7570231/ (visited on 04/04/2024).

[59] Gunnar Libiseller et al. "IPO: a tool for automated optimization of XCMS parameters". In: *BMC Bioinformatics* 16.1 (Apr. 2015), p. 118. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0562-8. URL: https://doi.org/10.1186/s12859-015-0562-8 (visited on 04/15/2024).

[60] Martha S. Linet et al. "International long-term trends and recent patterns in the incidence of leukemias and lymphomas among children and adolescents ages 0-19 years". eng. In: *International Journal of Cancer* 138.8 (Apr. 2016), pp. 1862–1874. ISSN: 1097-0215. DOI: 10.1002/ijc.29924.

[61] Sa Liu et al. "Cys34 Adductomes Differ between Patients with Chronic Lung or Heart Disease and Healthy Controls in Central London". eng. In: *Environmental Science & Technology* 52.4 (Feb. 2018), pp. 2307–2313. ISSN: 1520-5851. DOI: 10.1021/acs.est.7b05554.

[62] Petrick Lm et al. "Metabolomics of neonatal blood spots reveal distinct phenotypes of pediatric acute lymphoblastic leukemia and potential effects of early-life nutrition". en. In: *Cancer letters* 452 (June 2019). Publisher: Cancer Lett. ISSN: 1872-7980. DOI: 10.1016/j.canlet.2019.03.007. URL: https://pubmed.ncbi.nlm.nih.gov/30904619/ (visited on 03/08/2023).

[63] Dana Loomis et al. "The carcinogenicity of outdoor air pollution". eng. In: *The Lancet. Oncology* 14.13 (Dec. 2013), pp. 1262–1263. ISSN: 1474-5488. DOI: 10.1016/s1470-2045(13)70487-x.

[64] Jaclyn LoPiccolo et al. "Lung cancer in patients who have never smoked - an emerging disease". eng. In: *Nature Reviews. Clinical Oncology* 21.2 (Feb. 2024), pp. 121–146. ISSN: 1759-4782. DOI: 10.1038/s41571-023-00844-0.

[65] Mario Lovrić et al. "Vertical Metabolome Transfer from Mother to Child: An Explainable Machine Learning Method for Detecting Metabolomic Heritability". In: *Metabolites* 14.3 (Feb. 2024), p. 136. ISSN: 2218-1989. DOI: 10.3390/metabo14030136. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10972480/ (visited on 04/19/2024).

[66] K. H. Lui et al. "Cancer risk from polycyclic aromatic compounds in fine particulate matter generated from household coal combustion in Xuanwei, China". eng. In: *Chemosphere* 169 (Feb. 2017), pp. 660–668. ISSN: 1879-1298. DOI: 10.1016/j.chemosphere.2016.11.112.

[67] Léa Maitre et al. "Human Early Life Exposome (HELIX) study: a European population-based exposome cohort". en. In: *BMJ Open* 8.9 (Sept. 2018). Publisher: British Medical Journal Publishing Group Section: Epidemiology, e021311. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2017-021311. URL: https://bmjopen.bmj.com/content/8/9/e021311 (visited on 03/27/2024).

[68] *Metabolomics studies - NCI*. en. cgvArticle. Archive Location: nciglobal,ncienterprise. May 2012. URL: https://dceg.cancer.gov/research/how-we-study/metabolomics (visited on 03/27/2024).

[69] Catherine Metayer et al. "Genetic variants in the folate pathway and risk of childhood acute lymphoblastic leukemia". eng. In: *Cancer causes & control: CCC* 22.9 (Sept. 2011), pp. 1243–1258. ISSN: 1573-7225. DOI: 10.1007/s10552-011-9795-7.

[70] Catherine Metayer et al. "Maternal supplementation with folic acid and other vitamins and risk of leukemia in offspring: a Childhood Leukemia International Consortium study". eng. In: *Epidemiology (Cambridge, Mass.)* 25.6 (Nov. 2014), pp. 811–822. ISSN: 1531-5487. DOI: 10.1097/EDE.0000000000000141.

[71] Catherine Metayer et al. "One-Carbon (Folate) Metabolism Pathway at Birth and Risk of Childhood Acute Lymphoblastic Leukemia: A Biomarker Study in Newborns". In: *Cancers* 15.4 (Feb. 2023), p. 1011. ISSN: 2072-6694. DOI: 10.3390/cancers15041011. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9953980/ (visited on 04/25/2024).

[72] Catherine Metayer et al. "Tobacco smoke exposure and the risk of childhood acute lymphoblastic and myeloid leukemias by cytogenetic subtype". eng. In: *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 22.9 (Sept. 2013), pp. 1600–1611. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-13-0350.

[73] Tingting Mo et al. "Genetic susceptibility, homocysteine levels, and risk of all-cause and cause-specific mortality: A prospective cohort study". eng. In: *Clinica Chimica Acta; International Journal of Clinical Chemistry* 538 (Jan. 2023), pp. 1–8. ISSN: 1873-3492. DOI: 10.1016/j.cca.2022.11.001.

[74] Bhagavatula Moorthy, Chun Chu, and Danielle J. Carlin. "Polycyclic aromatic hydrocarbons: from metabolism to lung cancer". eng. In: *Toxicological Sciences: An Official Journal of the Society of Toxicology* 145.1 (May 2015), pp. 5–15. ISSN: 1096-0929. DOI: 10.1093/toxsci/kfv040.

[75] Shama Naz et al. "Development of a Liquid Chromatography-High Resolution Mass Spectrometry Metabolomics Method with High Specificity for Metabolite Identification Using All Ion Fragmentation Acquisition". eng. In: *Analytical Chemistry* 89.15 (Aug. 2017), pp. 7933–7942. ISSN: 1520-6882. DOI: 10.1021/acs.analchem.7b00925.

[76] Eric M. Nickels et al. "DNA methylation at birth in monozygotic twins discordant for pediatric acute lymphoblastic leukemia". eng. In: *Nature Communications* 13.1 (Oct. 2022), p. 6077. ISSN: 2041-1723. DOI: 10.1038/s41467-022-33677-z.

[77] Eric M. Nickels et al. "Periconceptional folate intake influences DNA methylation at birth based on dietary source in an analysis of pediatric acute lymphoblastic leukemia cases and controls". eng. In: *The American Journal of Clinical Nutrition* 116.6 (Dec. 2022), pp. 1553–1564. ISSN: 1938-3207. DOI: 10.1093/ajcn/nqac283.

[78] Henriette O'Geen, Lorigail Echipare, and Peggy J. Farnham. "Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications". In: *Methods in molecular biology (Clifton, N.J.)* 791 (2011), pp. 265–286. ISSN: 1064-3745.

DOI: 10.1007/978-1-61779-316-5_20. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151291/ (visited on 04/28/2024).

[79] Kevin C. Oeffinger et al. "Chronic health conditions in adult survivors of childhood cancer". eng. In: *The New England Journal of Medicine* 355.15 (Oct. 2006), pp. 1572–1582. ISSN: 1533-4406. DOI: 10.1056/NEJMsa060185.

[80] Alessandra Olarini et al. "Vertical Transfer of Metabolites Detectable from Newborn's Dried Blood Spot Samples Using UPLC-MS: A Chemometric Study". In: *Metabolites* 12.2 (Jan. 2022), p. 94. ISSN: 2218-1989. DOI: 10.3390/metabo12020094. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8879569/ (visited on 04/19/2024).

[81] Nisha T. Palackal et al. "Activation of polycyclic aromatic hydrocarbon trans-dihydrodiol proximate carcinogens by human aldo-keto reductase (AKR1C) enzymes and their functional overexpression in human lung carcinoma (A549) cells". eng. In: *The Journal of Biological Chemistry* 277.27 (July 2002), pp. 24799–24808. ISSN: 0021-9258. DOI: 10.1074/jbc.M112424200.

[82] Aris T. Papageorghiou et al. "The INTERGROWTH-21st fetal growth standards: toward the global integration of pregnancy and pediatric care". eng. In: *American Journal of Obstetrics and Gynecology* 218.2S (Feb. 2018), S630–S640. ISSN: 1097-6868. DOI: 10.1016/j.ajog.2018.01.011.

[83] Irene Peral-Sanchez et al. "Epigenetics in the Uterine Environment: How Maternal Diet and ART May Influence the Epigenome in the Offspring with Long-Term Health Consequences". eng. In: *Genes* 13.1 (Dec. 2021), p. 31. ISSN: 2073-4425. DOI: 10.3390/genes13010031.

[84] Serafina Perrone et al. "Newborn metabolomic profile mirrors that of mother in pregnancy". In: *Medical Hypotheses* 137 (Apr. 2020), p. 109543. ISSN: 0306-9877. DOI: 10.1016/j.mehy.2019.109543. URL: https://www.sciencedirect.com/science/article/pii/S0306987719313465 (visited on 04/19/2024).

[85] Kelsi Perttula et al. "Evaluating Ultra-long-Chain Fatty Acids as Biomarkers of Colorectal Cancer Risk". eng. In: *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 25.8 (Aug. 2016), pp. 1216–1223. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-16-0204.

[86] Kelsi Perttula et al. "Untargeted lipidomic features associated with colorectal cancer in a prospective cohort". In: *BMC Cancer* 18 (Oct. 2018), p. 996. ISSN: 1471-2407. DOI: 10.1186/s12885-018-4894-4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6194742/ (visited on 04/04/2024).

[87] Lauren Petrick et al. "An untargeted metabolomics method for archived newborn dried blood spots in epidemiologic studies". eng. In: *Metabolomics: Official Journal of the Metabolomic Society* 13.3 (Mar. 2017), p. 27. ISSN: 1573-3882. DOI: 10.1007/s11306-016-1153-z.

[88] Stephen M. Rappaport et al. "Adductomics: Characterizing exposures to reactive electrophiles". In: *Toxicology letters* 213.1 (Aug. 2012), pp. 83–90. ISSN: 0378-4274. DOI: 10.1016/j.toxlet.2011.04.002. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4758449/ (visited on 04/04/2024).

[89] Stefano Renzetti, Chris Gennings, and Stefano Calza. "A weighted quantile sum regression with penalized weights and two indices". In: *Frontiers in Public Health* 11 (July 2023), p. 1151821. ISSN: 2296-2565. DOI: 10.3389/fpubh.2023.1151821. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10392701/ (visited on 03/26/2024).

[90] Lee D. Roberts et al. "Targeted Metabolomics". In: *Current Protocols in Molecular Biology* CHAPTER (Apr. 2012), Unit30.2. ISSN: 1934-3639. DOI: 10.1002/0471142727.mb3002s98. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334318/ (visited on 03/27/2024).

[91] Mark D. Robinson and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". eng. In: *Genome Biology* 11.3 (2010), R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25.

[92] Courtney Schiffman. "The Role of Exploratory Data Analysis and Pre-processing in Omics Studies". en. PhD thesis. UC Berkeley, 2019. URL: https://escholarship.org/uc/item/7xd95799 (visited on 03/27/2024).

[93] Courtney Schiffman et al. "Filtering procedures for untargeted LC-MS metabolomics data". eng. In: *BMC bioinformatics* 20.1 (June 2019), p. 334. ISSN: 1471-2105. DOI: 10.1186/s12859-019-2871-9.

[94] Wei Jie Seow et al. "Association of Untargeted Urinary Metabolomics and Lung Cancer Risk Among Never-Smoking Women in China". In: *JAMA Network Open* 2.9 (Sept. 2019), e1911970. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2019.11970. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6755532/ (visited on 04/04/2024).

[95] Amanda W. Singer et al. "Maternal diet quality before pregnancy and risk of childhood leukaemia". eng. In: *The British Journal of Nutrition* 116.8 (Oct. 2016), pp. 1469–1478. ISSN: 1475-2662. DOI: 10.1017/S0007114516003469.

[96] Amanda W. Singer et al. "Maternal prenatal intake of one-carbon metabolism nutrients and risk of childhood leukemia". In: *Cancer causes & control : CCC* 27.7 (July 2016), pp. 929–940. ISSN: 0957-5243. DOI: 10.1007/s10552-016-0773-y. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4939072/ (visited on 03/08/2023).

[97] Colin A. Smith et al. "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification". eng. In: *Analytical Chemistry* 78.3 (Feb. 2006), pp. 779–787. ISSN: 0003-2700. DOI: 10.1021/ac051437y.

[98] Amanda L. Souza and Gary J. Patti. "A Protocol for Untargeted Metabolomic Analysis: From Sample Preparation to Data Processing". eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 2276 (2021), pp. 357–382. ISSN: 1940-6029. DOI: 10.1007/978-1-0716-1266-8_27.

[99] Massimo Stafoggia et al. "Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science". en. In: *Current Environmental Health Reports* 4.4 (Dec. 2017), pp. 481–490. ISSN: 2196-5412. DOI: 10.1007/s40572-017-0162-z. URL: https://doi.org/10.1007/s40572-017-0162-z (visited on 03/27/2024).

[100] Kurt Straif et al. "Carcinogenicity of polycyclic aromatic hydrocarbons". eng. In: *The Lancet. Oncology* 6.12 (Dec. 2005), pp. 931–932. ISSN: 1470-2045. DOI: 10.1016/s1470-2045(05)70458-7.

[101] Yumie Takata et al. "Calcium intake and lung cancer risk among female nonsmokers: a report from the Shanghai Women's Health Study". eng. In: *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 22.1 (Jan. 2013), pp. 50–57. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-12-0915-T.

[102] *The Human Genome Project*. en. URL: https://www.genome.gov/human-genome-project (visited on 03/27/2024).

[103] Jessica A Timms et al. "Exploring a potential mechanistic role of DNA methylation in the relationship between in utero and post-natal environmental exposures and risk of childhood acute lymphoblastic leukaemia". In: *International journal of cancer* 145.11 (Feb. 2019), pp. 2933–2943. ISSN: 0020-7136. DOI: 10.1002/ijc.32203. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6790139/ (visited on 03/08/2023).

[104] M Törnqvist and A Kautiainen. "Adducted proteins for identification of endogenous electrophiles." In: *Environmental Health Perspectives* 99 (Mar. 1993), pp. 39–44. ISSN: 0091-6765. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1567046/ (visited on 04/12/2024).

[105] O. Troyanskaya et al. "Missing value estimation methods for DNA microarrays". eng. In: *Bioinformatics (Oxford, England)* 17.6 (June 2001), pp. 520–525. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.6.520.

[106] Dimitris Tsoukalas et al. "Targeted Metabolomic Analysis of Serum Fatty Acids for the Prediction of Autoimmune Diseases". English. In: *Frontiers in Molecular Biosciences* 6 (Nov. 2019). Publisher: Frontiers. ISSN: 2296-889X. DOI: 10.3389/fmolb.2019.00120. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00120 (visited on 03/25/2024).

[107] Athanasios Valavanidis et al. "Pulmonary oxidative stress, inflammation and cancer: respirable particulate matter, fibrous dusts and ozone as major causes of lung carcinogenesis through reactive oxygen species mechanisms". eng. In: *International Journal of Environmental Research and Public Health* 10.9 (Aug. 2013), pp. 3886–3907. ISSN: 1660-4601. DOI: 10.3390/ijerph10093886.

[108] Catalina A. Vallejos et al. "Normalizing single-cell RNA sequencing data: challenges and opportunities". eng. In: *Nature Methods* 14.6 (June 2017), pp. 565–571. ISSN: 1548-7105. DOI: 10.1038/nmeth.4292.

[109] Wan Rosmawati Wan Ismail et al. "The Protective Effect of Maternal Folic Acid Supplementation on Childhood Cancer: A Systematic Review and Meta-analysis of Case-control Studies". eng. In: *Journal of Preventive Medicine and Public Health = Yebang Uihakhoe Chi* 52.4 (July 2019), pp. 205–213. ISSN: 2233-4521. DOI: 10.3961/jpmph.19.020.

[110] Rong Wang et al. "Cesarean Section and Risk of Childhood Acute Lymphoblastic Leukemia in a Population-Based, Record-Linkage Study in California". eng. In: *American Journal of Epidemiology* 185.2 (Jan. 2017), pp. 96–105. ISSN: 1476-6256. DOI: 10.1093/aje/kww153.

[111] Zhefang Wang et al. "Targeting Glutaminolysis: New Perspectives to Understand Cancer Development and Novel Strategies for Potential Target Therapies". In: *Frontiers in Oncology* 10 (Oct. 2020), p. 589508. ISSN: 2234-943X. DOI: 10.3389/fonc.2020.589508. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7649373/ (visited on 04/17/2024).

[112] Elizabeth J. Want et al. "Global metabolic profiling of animal and human tissues via UPLC-MS". eng. In: *Nature Protocols* 8.1 (Jan. 2013), pp. 17–32. ISSN: 1750-2799. DOI: 10.1038/nprot.2012.135.

[113] Hiroshi Watanabe et al. "Clinical Implications Associated With the Posttranslational Modification-Induced Functional Impairment of Albumin in Oxidative Stress-Related Diseases". eng. In: *Journal of Pharmaceutical Sciences* 106.9 (Sept. 2017), pp. 2195–2203. ISSN: 1520-6017. DOI: 10.1016/j.xphs.2017.03.002.

[114] Runmin Wei et al. "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data". en. In: *Scientific Reports* 8.1 (Jan. 2018). Publisher: Nature Publishing Group, p. 663. ISSN: 2045-2322. DOI: 10.1038/s41598-017-19120-0. URL: https://www.nature.com/articles/s41598-017-19120-0 (visited on 04/28/2024).

[115] Xiaoxia Wei et al. "Diet and Risk of Incident Lung Cancer: A Large Prospective Cohort Study in UK Biobank". eng. In: *The American Journal of Clinical Nutrition* 114.6 (Dec. 2021), pp. 2043–2051. ISSN: 1938-3207. DOI: 10.1093/ajcn/nqab298.

[116] Todd P. Whitehead et al. "Childhood Leukemia and Primary Prevention". eng. In: *Current Problems in Pediatric and Adolescent Health Care* 46.10 (Oct. 2016), pp. 317–352. ISSN: 1538-3199. DOI: 10.1016/j.cppeds.2016.08.004.

[117] J. L. Wiemels and M. Greaves. "Structure and possible mechanisms of TEL-AML1 gene fusions in childhood acute lymphoblastic leukemia". eng. In: *Cancer Research* 59.16 (Aug. 1999), pp. 4075–4082. ISSN: 0008-5472.

[118] J. L. Wiemels et al. "Prenatal origin of acute lymphoblastic leukaemia in children". eng. In: *Lancet (London, England)* 354.9189 (Oct. 1999), pp. 1499–1503. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(99)09403-9.

[119] Joseph L. Wiemels et al. "In utero origin of t(8;21) AML1-ETO translocations in childhood acute myeloid leukemia". eng. In: *Blood* 99.10 (May 2002), pp. 3801–3805. ISSN: 0006-4971. DOI: 10.1182/blood.v99.10.3801.

[120] Christopher Paul Wild. "The exposome: from concept to utility". In: *International Journal of Epidemiology* 41.1 (Feb. 2012), pp. 24–32. ISSN: 0300-5771. DOI: 10.1093/ije/dyr236. URL: https://doi.org/10.1093/ije/dyr236 (visited on 03/27/2024).

[121] Lindsay A. Williams et al. "Is There Etiologic Heterogeneity between Subtypes of Childhood Acute Lymphoblastic Leukemia? A Review of Variation in Risk by Subtype". eng. In: *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 28.5 (May 2019), pp. 846–856. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-18-0801.

[122] Jason Y. Y. Wong et al. "Exposure to diesel engine exhaust and alterations to the Cys34/Lys525 adductome of human serum albumin". eng. In: *Environmental Toxicology and Pharmacology* 95 (Oct. 2022), p. 103966. ISSN: 1872-7077. DOI: 10.1016/j.etap.2022.103966.

[123] Xu Xu et al. "Selective detection of monohydroxy metabolites of polycyclic aromatic hydrocarbons in urine using liquid chromatography/triple quadrupole tandem mass spectrometry". eng. In: *Rapid communications in mass spectrometry: RCM* 18.19 (2004), pp. 2299–2308. ISSN: 0951-4198. DOI: 10.1002/rcm.1625.

[124] Jingrong Yan et al. "A meta-analysis of MTHFR C677T and A1298C polymorphisms and risk of acute lymphoblastic leukemia in children". eng. In: *Pediatric Blood & Cancer* 58.4 (Apr. 2012), pp. 513–518. ISSN: 1545-5017. DOI: 10.1002/pbc.23137.

[125] Gong Yang et al. "Soy Food Intake and Risk of Lung Cancer: Evidence From the Shanghai Women's Health Study and a Meta-Analysis". In: *American Journal of Epidemiology* 176.10 (Nov. 2012), pp. 846–855. ISSN: 0002-9262. DOI: 10.1093/aje/kws168. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3626060/ (visited on 04/04/2024).

[126] Jia Yang et al. "Association of One-Carbon Metabolism-Related Vitamins (Folate, B6, B12), Homocysteine and Methionine With the Risk of Lung Cancer: Systematic Review and Meta-Analysis". eng. In: *Frontiers in Oncology* 8 (2018), p. 493. ISSN: 2234-943X. DOI: 10.3389/fonc.2018.00493.

[127] Zhenchun Yang et al. "Urinary Amino-Polycyclic Aromatic Hydrocarbons in Urban Residents: Finding a Biomarker for Residential Exposure to Diesel Traffic". eng. In: *Environmental Science & Technology* 55.15 (Aug. 2021), pp. 10569–10577. ISSN: 1520-5851. DOI: 10.1021/acs.est.1c01549.

[128] Yukiko Yano et al. "Untargeted adductomics of Cys34 modifications to human serum albumin in newborn dried blood spots". eng. In: *Analytical and Bioanalytical Chemistry* 411.11 (Apr. 2019), pp. 2351–2362. ISSN: 1618-2650. DOI: 10.1007/s00216-019-01675-8.

[129] Lina Youssef et al. "Paired maternal and fetal metabolomics reveal a differential fingerprint in preeclampsia versus fetal growth restriction". en. In: *Scientific Reports* 11.1 (July 2021). Publisher: Nature Publishing Group, p. 14422. ISSN: 2045-2322. DOI: 10.1038/s41598-021-93936-9. URL: https://www.nature.com/articles/s41598-021-93936-9 (visited on 04/19/2024).

[130] Danxia Yu et al. "Overall and Central Obesity and Risk of Lung Cancer: A Pooled Analysis". In: *JNCI Journal of the National Cancer Institute* 110.8 (Mar. 2018), pp. 831–842. ISSN: 0027-8874. DOI: 10.1093/jnci/djx286. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6093439/ (visited on 04/04/2024).

[131] Ignatius T. S. Yu et al. "Dose-response relationship between cooking fumes exposures and lung cancer among Chinese nonsmoking women". eng. In: *Cancer Research* 66.9 (May 2006), pp. 4961–4967. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-05-2932.

[132] Miao Yu, Mariola Olkowicz, and Janusz Pawliszyn. "Structure/reaction directed analysis for LC-MS based untargeted analysis". In: *Analytica Chimica Acta* 1050 (Mar. 2019), pp. 16–24. ISSN: 0003-2670. DOI: 10.1016/j.aca.2018.10.062. URL: https://www.sciencedirect.com/science/article/pii/S0003267018313047 (visited on 04/15/2024).

[133] Miao Yu and Lauren Petrick. "Untargeted high-resolution paired mass distance data mining for retrieving general chemical relationships". en. In: *Communications Chemistry* 3.1 (Nov. 2020). Publisher: Nature Publishing Group, pp. 1–6. ISSN: 2399-

3669. DOI: 10.1038/s42004-020-00403-z. URL: https://www.nature.com/articles/s42004-020-00403-z (visited on 03/27/2024).

[134] Miao Yu et al. "Untargeted metabolomics profiling and hemoglobin normalization for archived newborn dried blood spots from a refrigerated biorepository". eng. In: *Journal of Pharmaceutical and Biomedical Analysis* 191 (Nov. 2020), p. 113574. ISSN: 1873-264X. DOI: 10.1016/j.jpba.2020.113574.

[135] Kefei Yuan et al. "Thiol-based redox proteomics in cancer research". eng. In: *Proteomics* 15.2-3 (Jan. 2015), pp. 287–299. ISSN: 1615-9861. DOI: 10.1002/pmic.201400164.

[136] Junfeng Jim Zhang et al. "Urinary Amino-PAHs in relation to diesel engine emissions and urinary mutagenicity". eng. In: *International Journal of Hygiene and Environmental Health* 253 (Aug. 2023), p. 114223. ISSN: 1618-131X. DOI: 10.1016/j.ijheh.2023.114223.

[137] Pei Zhang et al. "Defining the Scope of Exposome Studies and Research Needs from a Multidisciplinary Perspective". In: *Environmental Science & Technology Letters* 8.10 (Oct. 2021), pp. 839–852. ISSN: 2328-8930. DOI: 10.1021/acs.estlett.1c00648. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8515788/ (visited on 04/28/2024).

[138] Tongwu Zhang et al. "Distinct Genomic Landscape of Lung Adenocarcinoma from Household Use of Smoky Coal". eng. In: *American Journal of Respiratory and Critical Care Medicine* 208.6 (Sept. 2023), pp. 733–736. ISSN: 1535-4970. DOI: 10.1164/rccm.202302-0340LE.

[139] Wei Zheng et al. "The Shanghai Women's Health Study: Rationale, Study Design, and Baseline Characteristics". In: *American Journal of Epidemiology* 162.11 (Dec. 2005), pp. 1123–1131. ISSN: 0002-9262. DOI: 10.1093/aje/kwi322. URL: https://doi.org/10.1093/aje/kwi322 (visited on 04/04/2024).

[140] L. Zhong et al. "A case-control study of lung cancer and environmental tobacco smoke among nonsmoking women living in Shanghai, China". eng. In: *Cancer causes & control: CCC* 10.6 (Dec. 1999), pp. 607–616. ISSN: 0957-5243. DOI: 10.1023/a:1008962025001.

[141] Hongjun Zhu and Shuanglin Zhang. "Body mass index and lung cancer risk in never smokers: a meta-analysis". In: *BMC Cancer* 18 (June 2018), p. 635. ISSN: 1471-2407. DOI: 10.1186/s12885-018-4543-y. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5987408/ (visited on 04/04/2024).

[142] Jun Zhu et al. "Acute Lymphoblastic Leukemia in Combined Methylmalonic Acidemia and Homocysteinemia (cblC Type): A Case Report and Literature Review". eng. In: *Frontiers in Genetics* 13 (2022), p. 856552. ISSN: 1664-8021. DOI: 10.3389/fgene.2022.856552.

Appendix A

A.1 Supplementary Figures and Tables

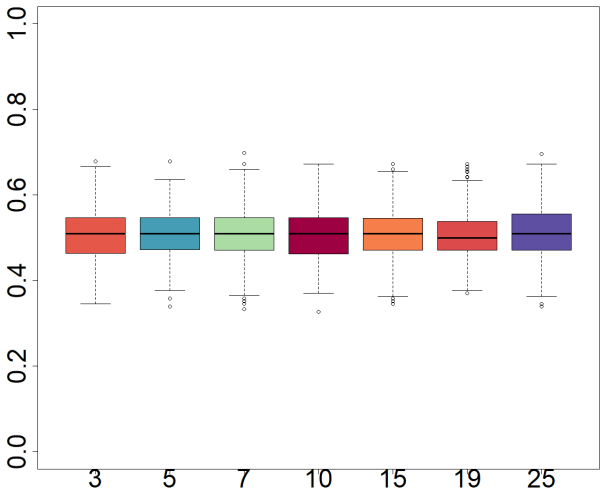


Figure A.1: Boxplots of classification error rates using k-NN on the Monte Carlo sample partitions of the full learning dataset for a range of k values (shown on x-axis). Boxplots are colored by the mean classification error rate by k. Darker red indicates a lower (better) mean classification error rate and dark purple indicates the highest.

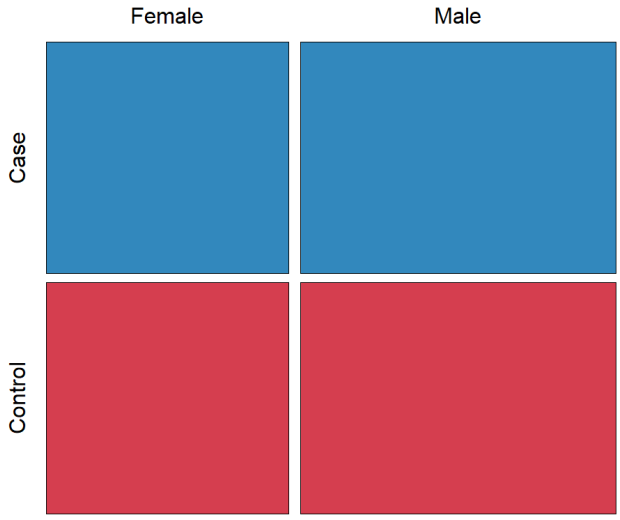


Figure A.2: Mosaic plot of 2×2 contingency table for joint distribution of case-control status and sex.

Metabolites	Spearman Correlation Coefficient
L-Methionine	0.057
Betaine / L-Valine	-0.002
Homocysteine	-0.016
Choline	0.043
Taurine	0.013
Dimethylglycine / 2-Aminoisobutyrate	0.003
L-Cysteine	0.027
Glycine	-0.021
L-Serine	0.009
S-Adenosylhomocysteine	-0.022
Cystathionine	-0.031

Table A.1: Spearman correlation coefficients between blood specimen age (defined as the numbers of years between the sample collection and processing at the laboratory) and measured metabolites in the folate pathway.

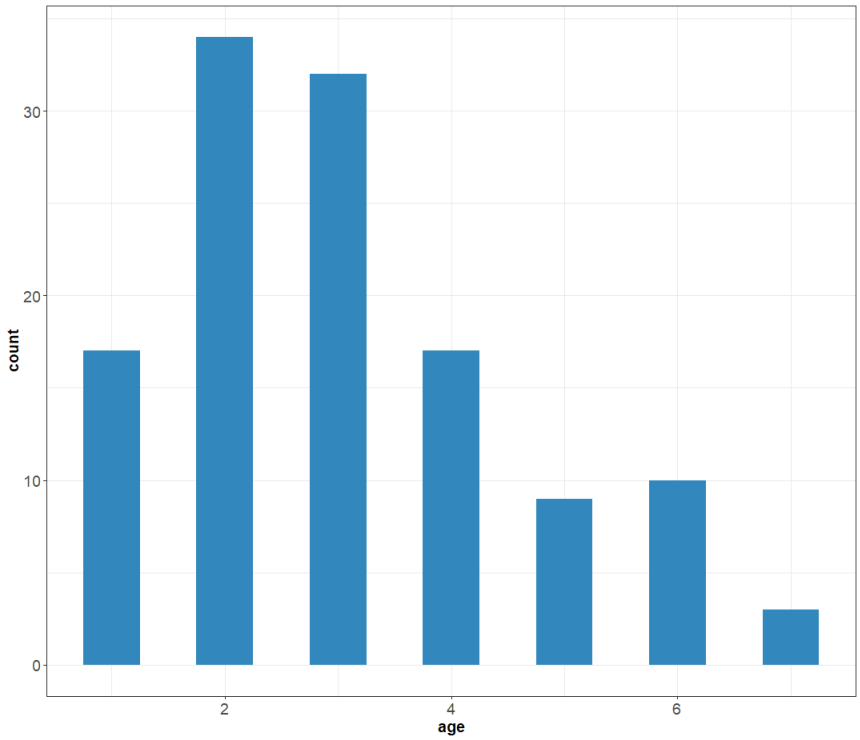


Figure A.3: Barplot for distribution of age at diagnosis (in years) for cases.

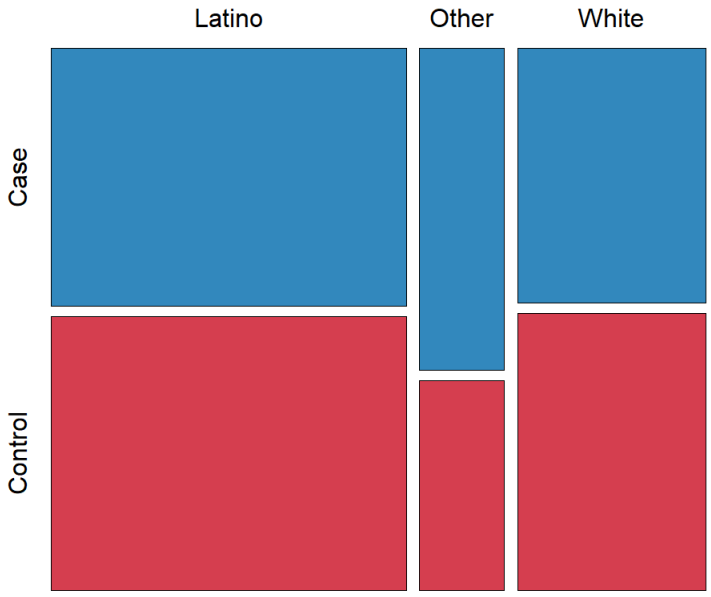


Figure A.4: Mosaic plot of 2×3 contingency table for joint distribution of case-control status and race.

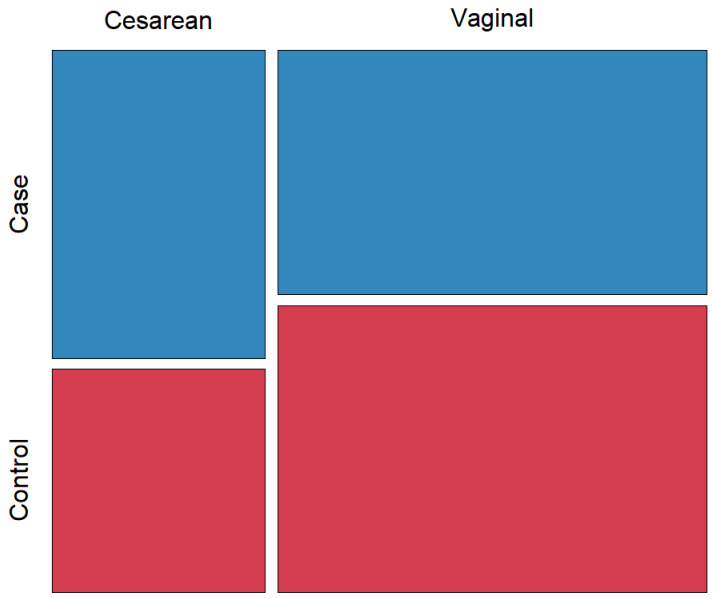


Figure A.5: Mosaic plot of 2×2 contingency table for joint distribution of case-control status and mode of delivery.

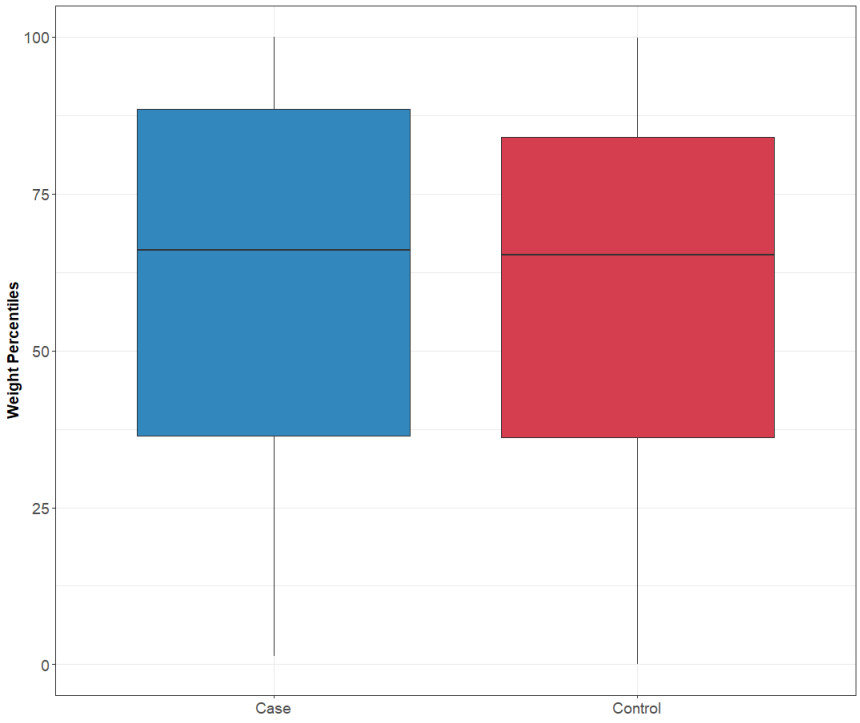


Figure A.6: Boxplots of weight percentiles conditional on gestational age, stratified by case/-ontrol status. Calculated using the INTERGROWTH-21st standards.

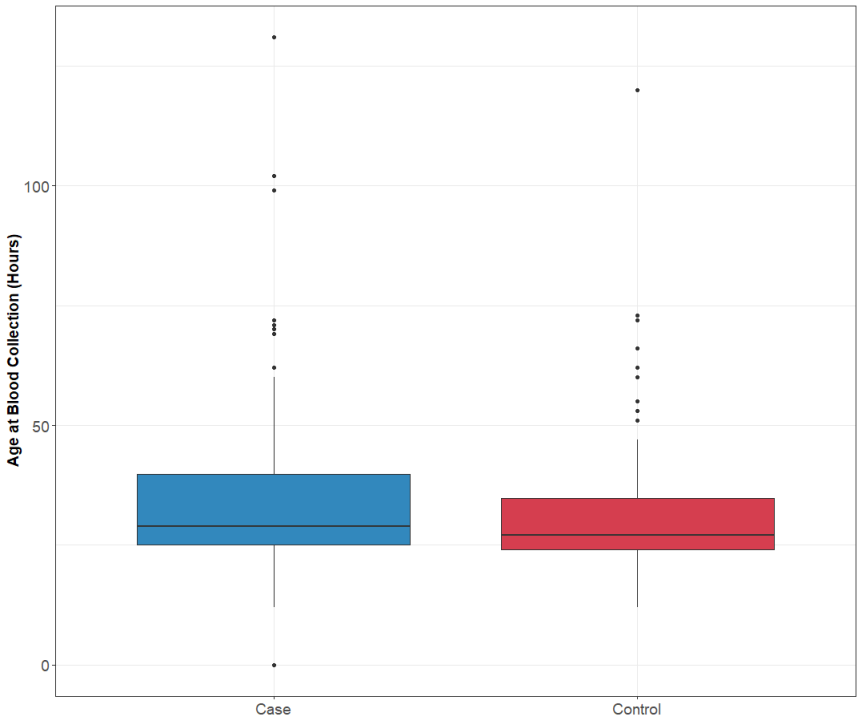


Figure A.7: Boxplots of age at blood collection stratified by case-control status.

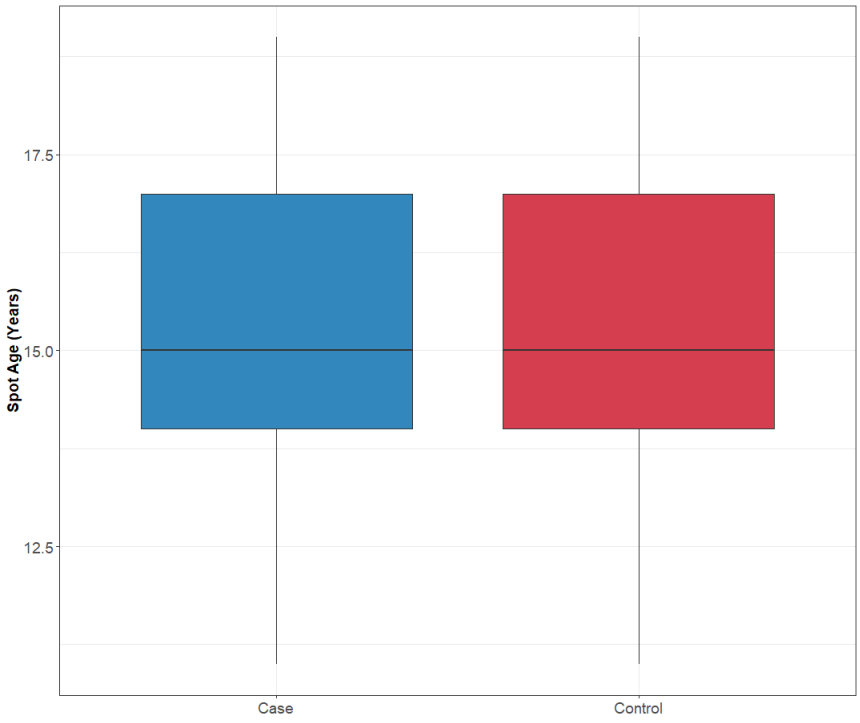


Figure A.8: Boxplots of number of years from sample collection to sample processing, stratified by case-control status.

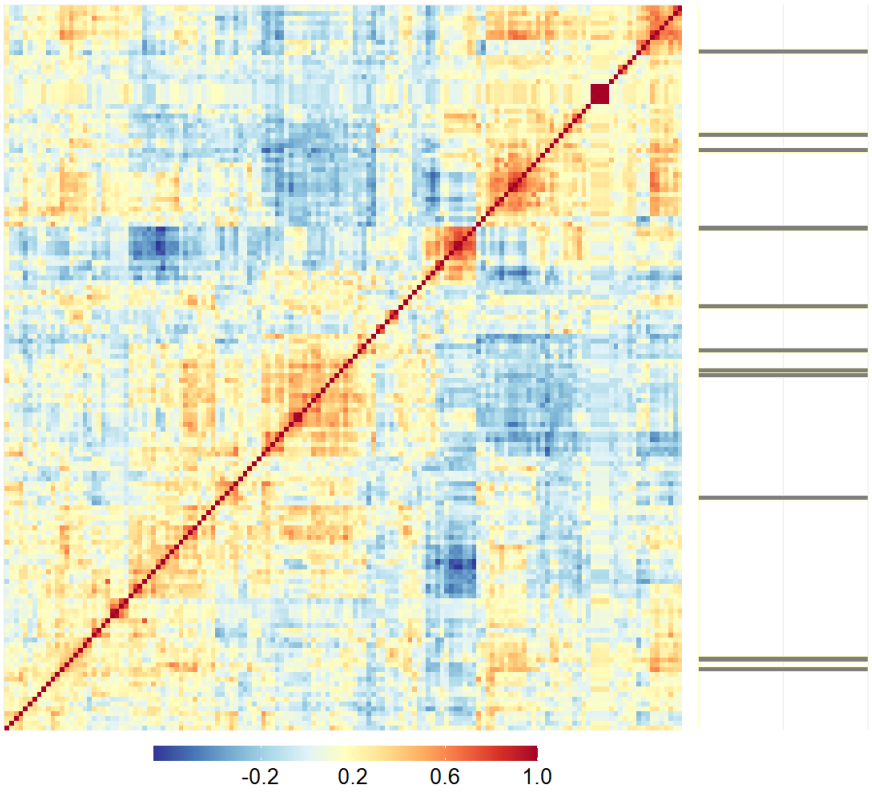


Figure A.9: Pseudo-color image of Spearman correlation matrix for all 148 metabolites, with rows and columns ordered by complete linkage hierarchical clustering. Darker red indicates a stronger positive correlation and a darker blue indicates a stronger negative correlation.

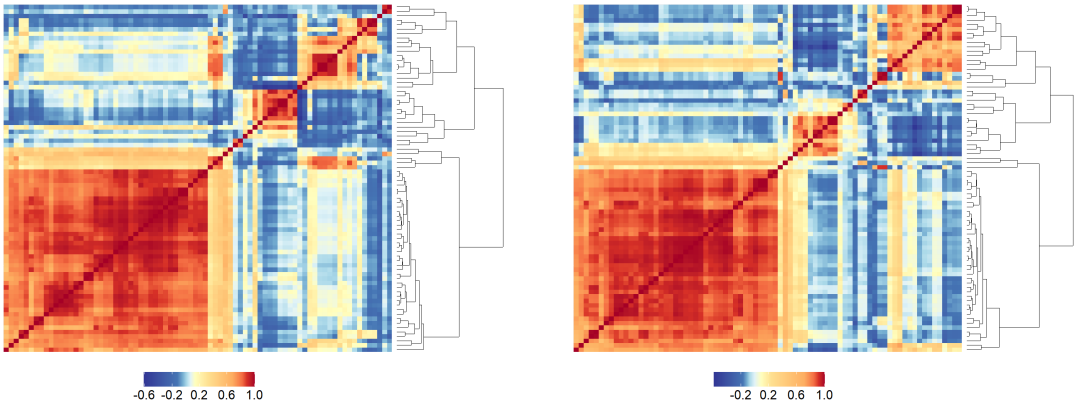
Appendix B

B.1 Supplementary Methods

LC-HRMS data acquisition

Samples were maintained at 5°C in the autosampler module. For nonpolar metabolites separation, 2 μL of sample sandwiched between 10 μL of water was injected onto a Zorbax Eclipse Plus C18, RRHD column (50 mm × 2.1 mm, 1.8 μm particle size, Agilent Technologies, Santa Clara, USA) coupled to a guard column (5 mm × 2 mm, 1.8 μm Agilent Technologies, Santa Clara, USA) maintained at 50°C. The analysis runtime was 19 minute. Mobile phases for gradient elution included mobile phase A of 100% water with 0.1% formic acid and mobile phase B of 90%IPA/10%ACN with 0.1% formic acid. The gradient was set at 5% B for 2 min, increasing to 30% B in 4 min, increasing to 98% B in 13.5 min followed by a 1.5 min hold and a 4 min equilibration.

B.2 Supplementary Figures and Tables



(a) Cases (b) Controls

Figure B.1: Heatmaps of hierarchical clustering results using pairwise spearman correlation values between neonatal qPMDs for a) cases and b) controls.

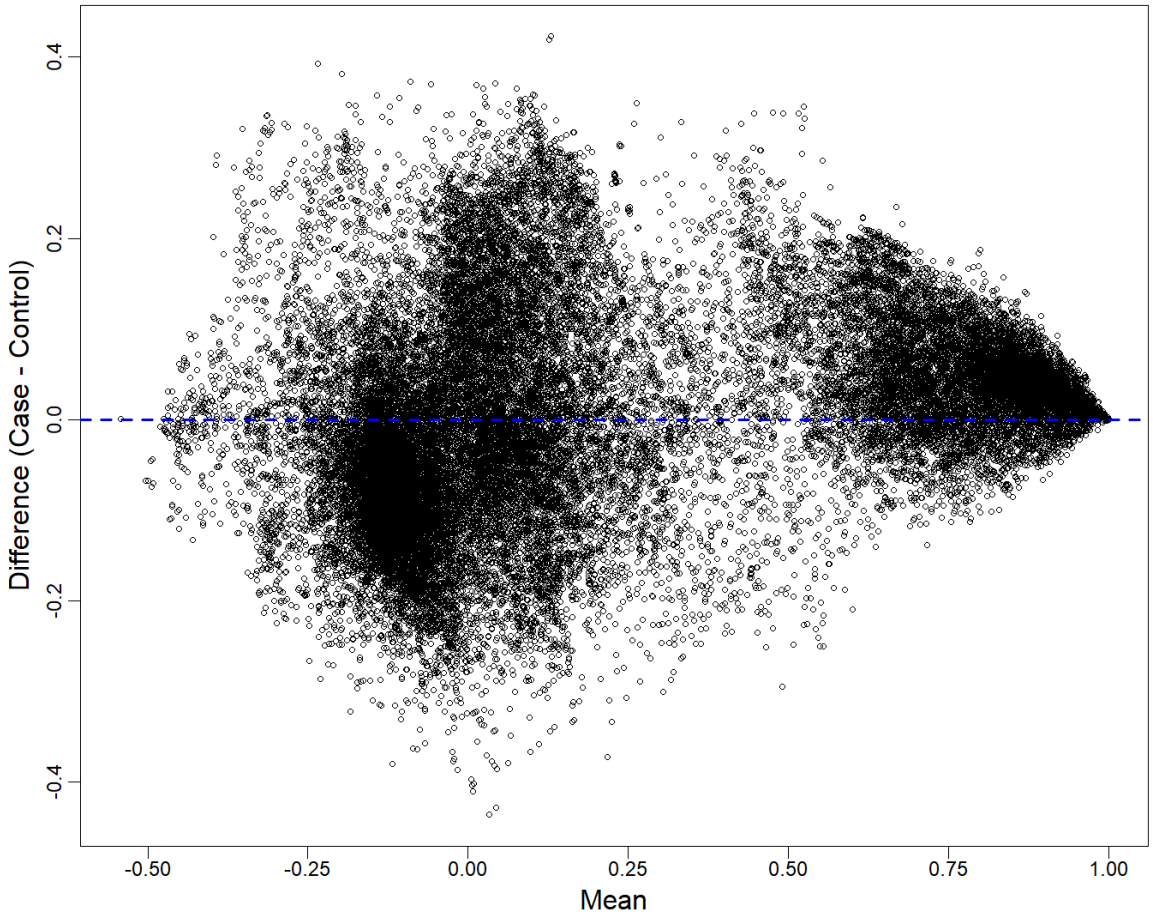
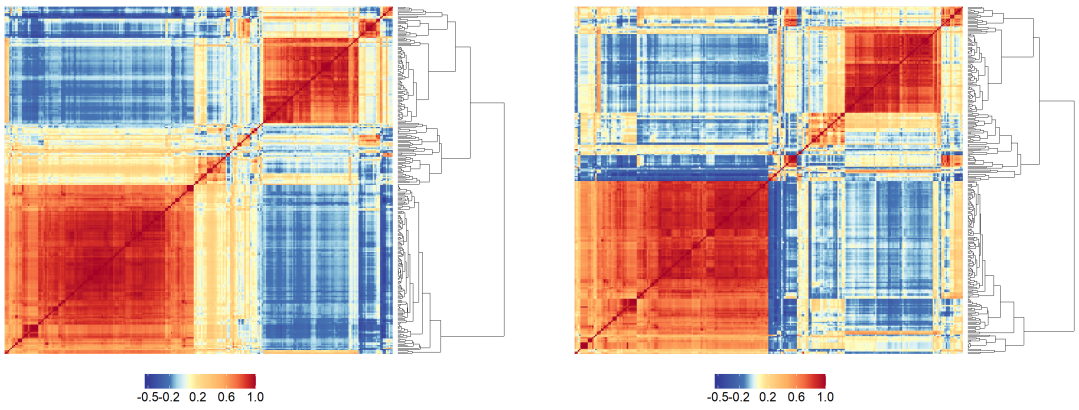


Figure B.2: Mean-difference plot of spearman correlation values for full set of maternal qPMDs with 36,046 correlation points (pairwise for 269 qPMDs). X-axis shows the mean and y-axis shows the difference (cases – controls). Blue line shows a difference of zero.



(a) Cases (b) Controls

Figure B.3: Heatmaps of hierarchical clustering results using pairwise spearman correlation values between maternal qPMDs for a) cases and b) controls.

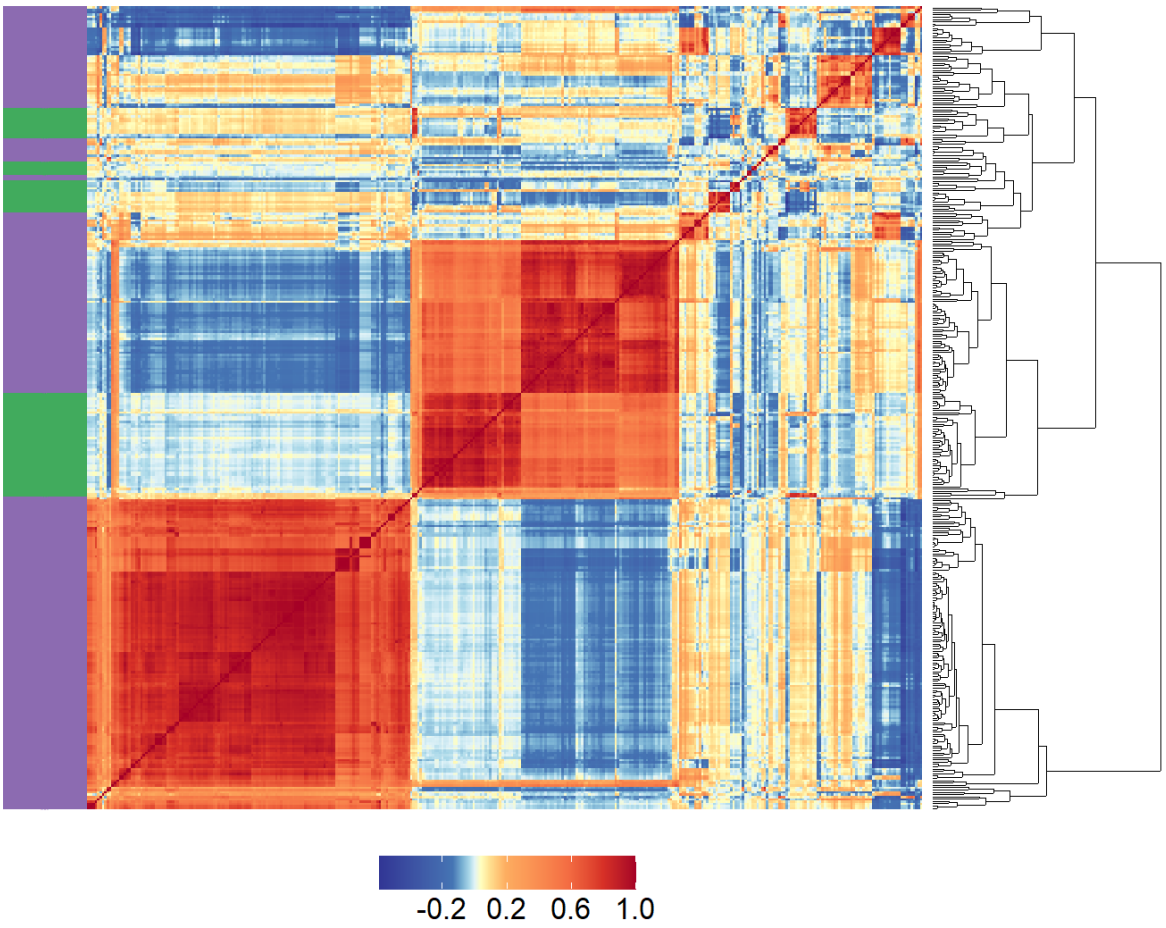
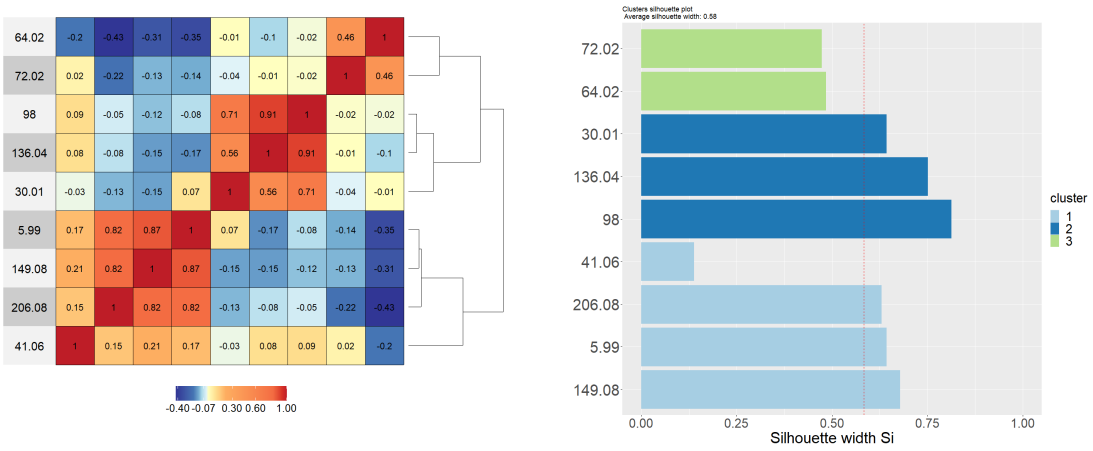
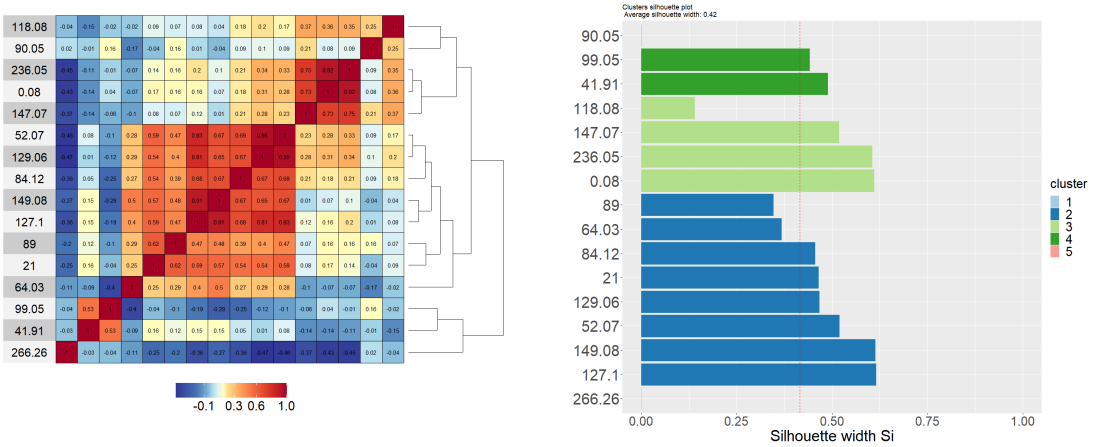


Figure B.4: Heatmap showing hierarchical clustering results for neonatal and maternal PMDs jointly based on pairwise spearman correlation values. The color bar along the left side indicates whether the qPMD is from the neonatal data (green) or from the maternal data (purple).



(a) (b)

Figure B.5: Clustering using Spearman correlation values of selected DBS qPMDs associated with childhood leukemia. a) Heatmap of hierarchical clustering results in three distinct correlation clusters. b) Silhouette widths for PAM clustering results on the same qPMDs. Red line indicates average silhouette width of 0.58.



(a) (b)

Figure B.6: Clustering using Spearman correlation values of selected maternal serum qPMDs associated with leukemia. a) Heatmap of hierarchical clustering results in three distinct correlation clusters. b) Silhouette widths for PAM clustering results on the same qPMDs. Red line indicates average silhouette width of 0.58.

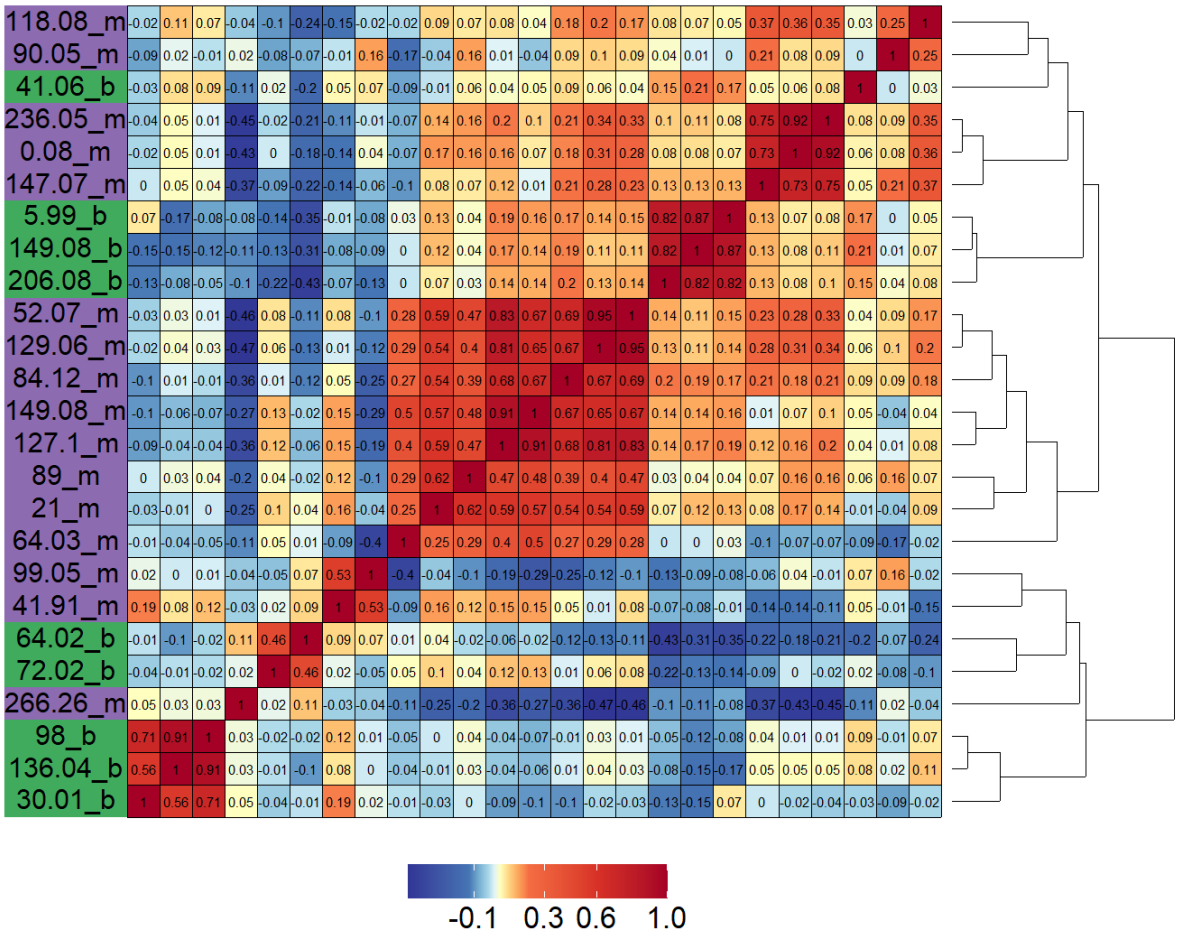


Figure B.7: Heatmaps of hierarchical clustering results using pairwise Spearman correlation values between selected neonatal and maternal PMDs.

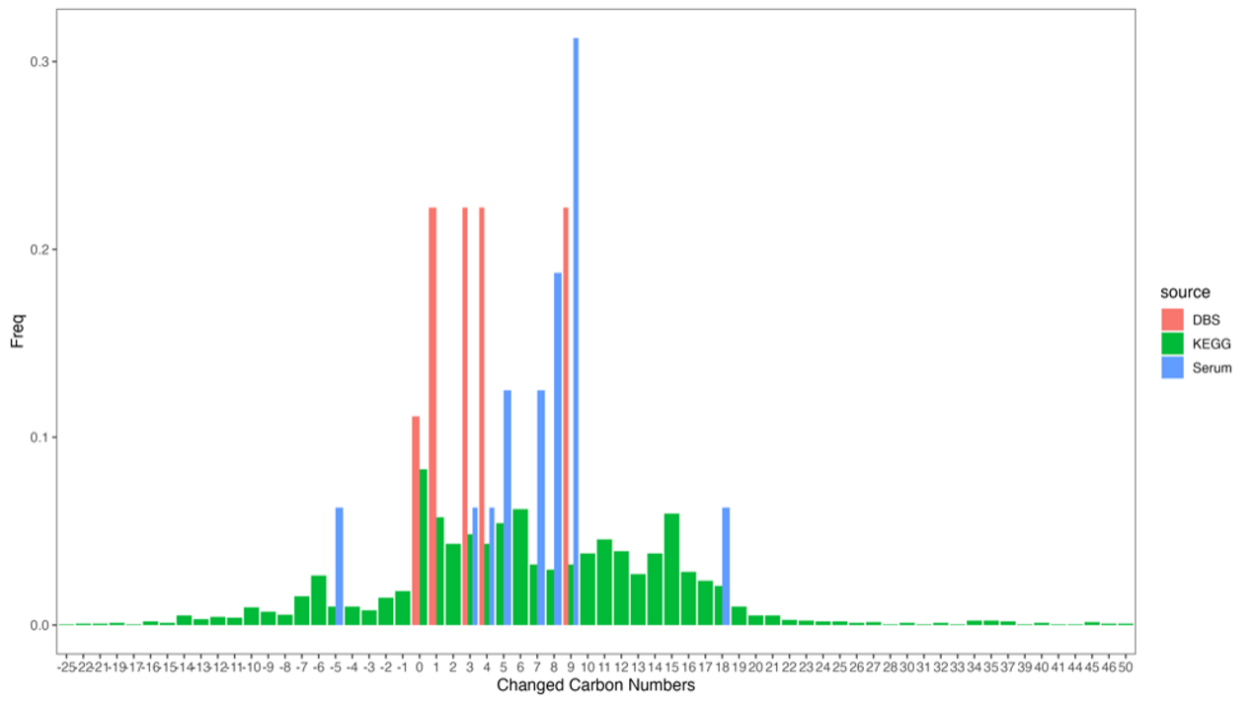


Figure B.8: Distribution of elemental carbon changes of qPMDs from the KEGG database compared to those selected qPMDs in DBS and maternal serum associated with child leukemia status.

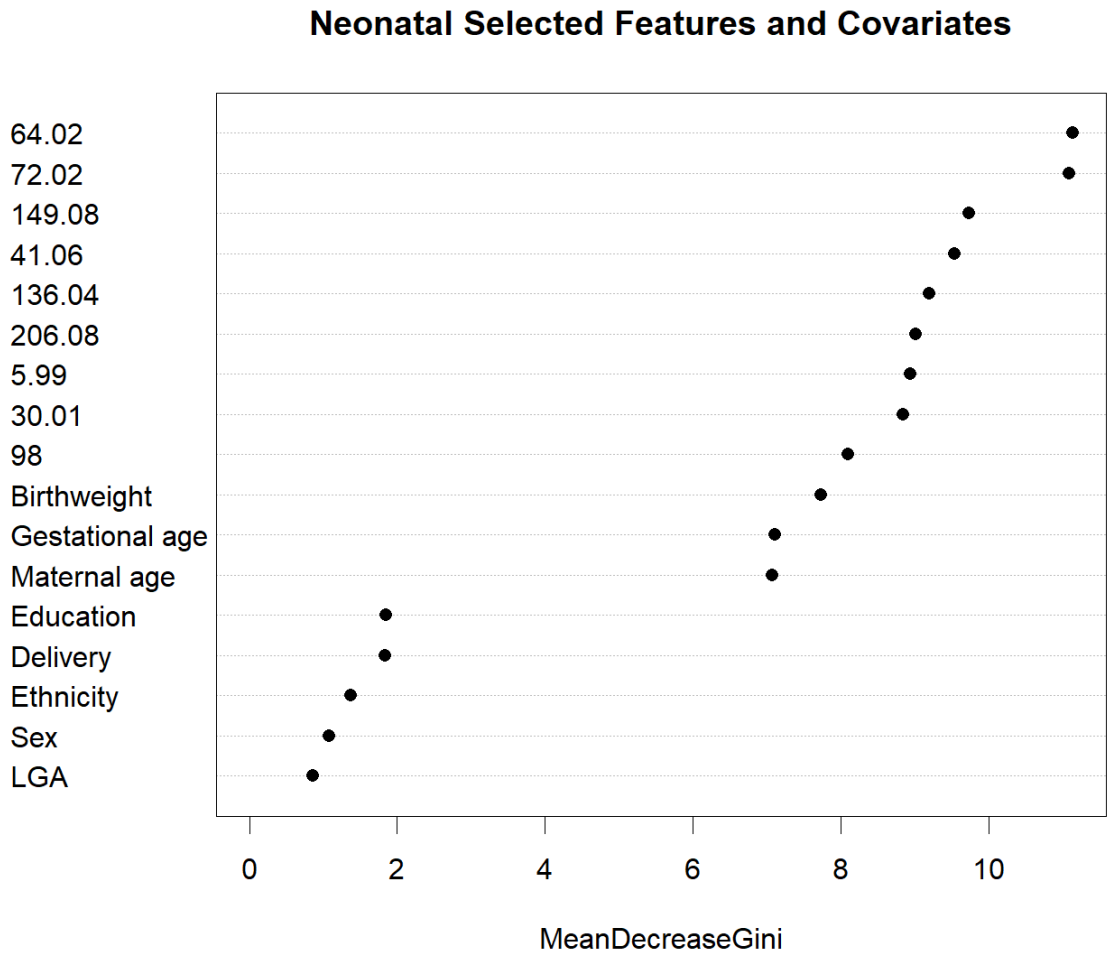
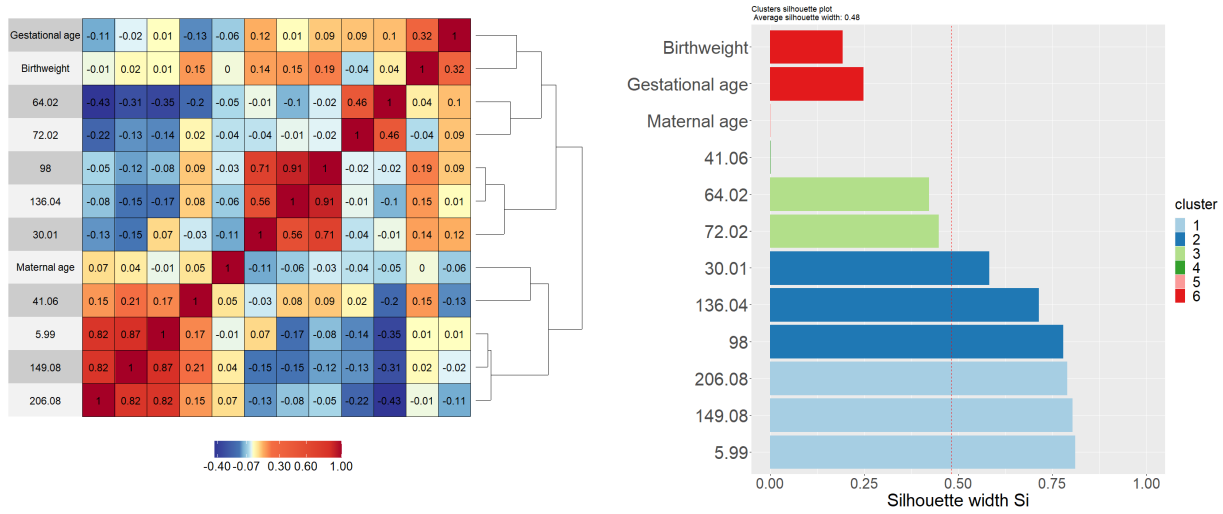


Figure B.9: Random Forests variable importance plot using the mean decrease in Gini index for prediction of case-control status using the selected neonatal qPMDs and covariates.



(a) (b)

Figure B.10: Clustering using Spearman correlation values of selected neonatal DBS qPMDs associated with childhood leukemia and continuous covariates. Left: Heatmap of hierarchical clustering results in four distinct correlation clusters. Right: Silhouette widths for PAM clustering clusters the same qPMDs. Red line indicates average silhouette width of 0.48.

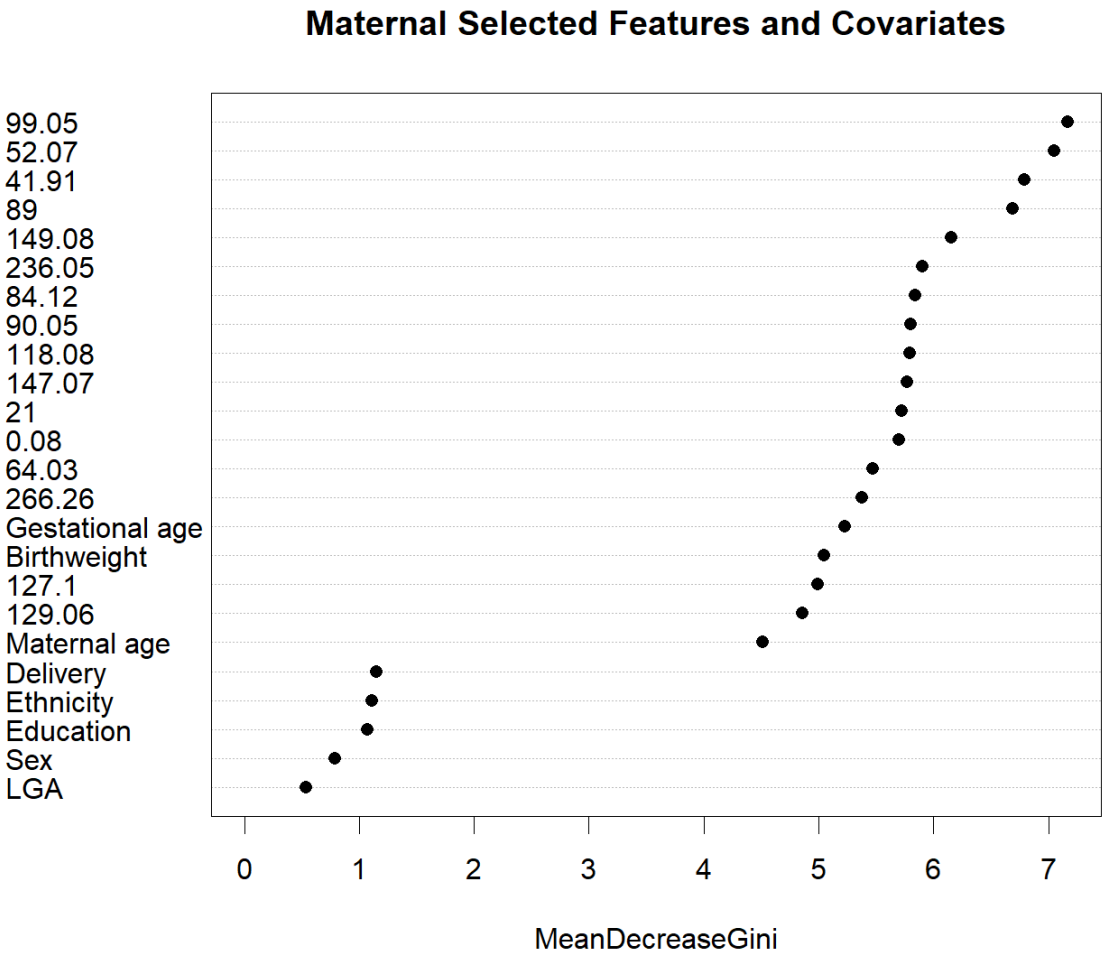
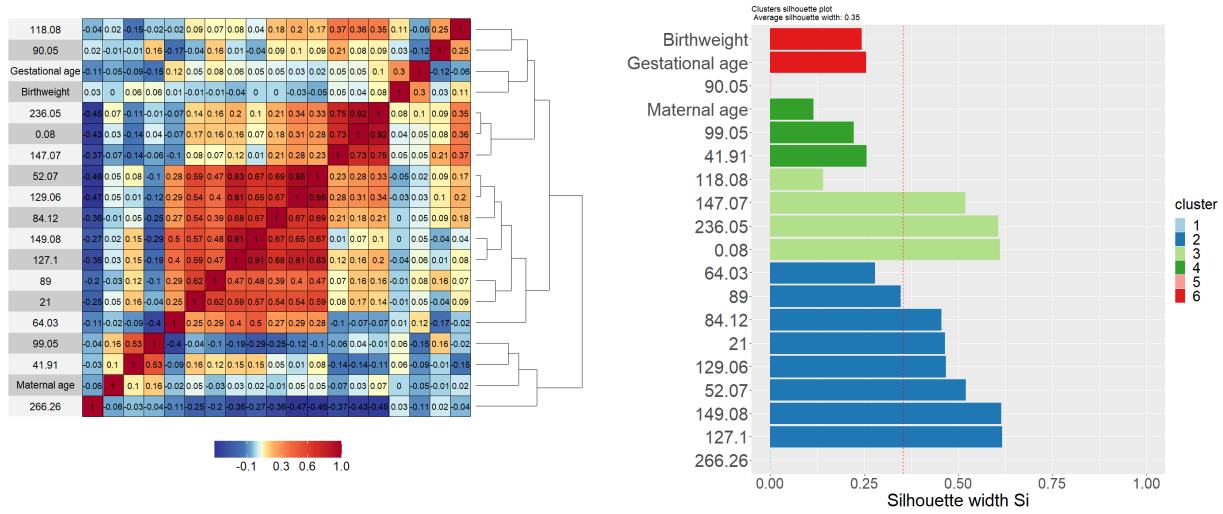


Figure B.11: Random Forests variable importance plot using the Mean decrease in Gini index for prediction of case-control status using the selected maternal serum qPMDs and covariates.



(a) (b)

Figure B.12: Clustering using Spearman correlation values of selected maternal serum qPMDs associated with childhood leukemia and continuous covariates. a) Heatmap of hierarchical clustering results in four distinct correlation clusters. b) Silhouette widths for PAM clustering clusters the same qPMDs. Red line indicates average silhouette width of 0.35.

Appendix C

C.1 Supplementary Methods

Measurement of Cys34/Lys525 adducts using nLC-HRMSMS as summarized from [34, 36]

Nano-liquid chromatography-high resolution mass spectrometry/mass spectrometry (nLC-HRMSMS) was performed in data-dependent mode with duplicate injections as described in (1,2). Purified HSA (≥ 75 %) from 5 μL of plasma was digested with trypsin without prior reduction of disulfide bonds at 37°C using high-pressure cycling for 30 min (NEP2320, Pressure Biosciences Inc., South Easton, MA). Prior to nLC-HRMSMS, 1 μL (20 pmol) of an internal standard, consisting of the isotopically labeled T3 peptide modified at Cys34 with iodoacetamide (IAA-iT3), was added to normalize data for instrument performance. One microliter of each digest was then injected into the nLC-HRMSMS, consisting of a Dionex Ultimate 3000 nanoflow LC system equipped with a Dionex PepSwift monolithic column (100-μm i.d. x 25 cm) and connected via a Flex Ion nano-electrospray-ionization source with an Orbitrap Elite Hybrid HRMS (Thermo Scientific, Sunnyvale, CA) that was operated in positive-ion mode with a spray voltage of 2.0 kV, a capillary temperature of 200°C and voltage of 35 V. Peptides were separated at a flow rate of 750 nL/min. Mobile phase A was water/0.1% formic acid (v/v) and mobile phase B was acetonitrile/0.1% formic acid (v/v). The gradient involved isocratic flow at 2% B for 5 min, a linear gradient from 2% to 45% B over 30 min, rapid increase to 98% B to wash the column (3 min), and reset to initial conditions over 5 min. The retention time range for Lys525 containing miss-cleaved peptides was set from 7 to 17 min compared to that of the Cys34 peptides (25 to 35 min). The precursor mass range for the search of triply-charged precursors of Cys34 adducts and doubly-charged precursors of Lys525 adducts was from 500 *m/z* to 1000 *m/z* which represents added masses from -129 to 370 Da and -185 to 565 Da for Lys525 and Cys34, respectively (negative added masses reflect losses due to elimination reactions, deletions, rearrangements and truncations). After sample injections, blanks and column washes were injected to maintain instrument performance as described in [34].

The mass spectra were acquired over the range *m/z* = 300-1200 using the Orbitrap mass analyzer in positive profile mode, with a resolution of 6×10^4 at *m/z* = 400. In data-dependent mode, up to 12 triply-charged (Cys34) and doubly-charged (Lys525) precursor ions exceeding 15,000 counts were selected from each MS1 scan and fragmented by collision-induced dissociation (CID) and analyzed in the Velos Pro linear ion trap in centroid mode. To avoid the occurrence of redundant MS2 measurements, real-time dynamic exclusion was enabled using the following parameters: repeat count 2, repeat duration 20 s, exclusion list size 500, exclusion duration 75 s, and exclusion precursor ion width ±10 ppm. Charge-state screening and monoisotopic-ion selection were enabled. The lock-mass option was enabled to provide a real-time internal mass calibration using a reference list of 8 identified background ions [50]. External mass calibration of the mass spectrometry and the LC diagnostics were performed weekly. MS1 and MS2 spectra were acquired and analyzed using Xcalibur software (version 3.0.63) and Chromeleon Xpress (v. 6.80) (Thermo Fisher Scientific, Sunnyvale, CA). Adducts were located via MS2 spectra on the triply charged T3 peptide, that contains Cys34, (^{21}ALVLIAFAQYLQQC^{34}PFEDHVK^{41}, *m/z* 811.7593) and the doubly charged miscleaved Lys525 peptide (^{525K}QTALVELVK^{534}, m/z 564.8529) that displayed the signature b^+- and/or y^{2+}-series ions [34]. The corresponding precursor ions were then extracted from the total ion chromatogram (TIC) to obtain a monoisotopic mass (MIM) for each adduct feature. To normalize peak areas for the amount of HSA in each sample, the MIM was also extracted for the HK peptide 42LVNEVTEFAK51 (*m/z*=575.3111), which is adjacent to T3. The peak area ratio (PAR), representing the ratio of the abundance of the adduct peak to that of HK is a robust measure of the adduct concentration. Peak picking and integration were performed using the Xcalibur Processing Method (version 3.0, Thermo Fisher Scientific, Waltham, MA) based on the average MIMs and retention times. Peak integration employed the Genesis algorithm after normalizing for instrument performance via iT3-IAA. Added masses relative to the Cys34 thiolate ion were estimated as Madduct = (m/z_{adduct}-$m/z_{T3-peptide}$)*3 + 1.0078, where m/z_{adduct} and $m/z_{T3-peptide}$ are the observed m/z values for the triply-charged MIMs of a given precursor ion for an adduct and the unmodified T3 peptide, respectively, and 1.0078 is the mass of a hydrogen atom.

The mass added to the $\mathcal{E}$-amino group of Lys525 was calculated in two ways to account for different mechanisms of adduct formation. The first calculation of added mass is consistent with formation of Michael-addition products that can occur in HSA at Lys525. In this case, the MIM of the doubly charged precursor ion is subtracted from the corresponding MIM of the modified Lys525-containing miss-cleaved peptide and the difference is multiplied by two and the mass of one hydrogen atom is added. The second calculation applies to Schiff-base formation that can occur with the $\mathcal{E}$-amino group of Lys525 and involves loss of a water molecule. To contend with this mechanism, added masses for putative Lys525 adducts were calculated corresponding to addition of two hydrogens to Lys525. Notably, the added mass representing the reactive electrophile per se would show loss of the oxygen atom. All data processing utilized in-house software written in R [1].

LC-MS/MS analysis of urinary metabolites of exogenous and amino-PAHs.

Several urinary metabolites of exogenous PAH and nitro-PAH were measured by LC-MSMS as described by Yang et al. [127] and Xu et al. [123] with modifications. Two composite sets of urinary PAH metabolites were generated from these measurements, namely hydroxylated-PAH metabolites ('Hydroxy-PAHs': sum of 2-, 3-, and 4-hydroxyphenanthrene, 1,9-dihydroxyphenanthrene, and 10-hydroxypyrene) that are products of metabolism of unmodified PAHs, and aminated PAHs ('Amino-PAHs': sum of 9-aminophenanthrene, 3-aminobenzanthracene, 1-, and 2-aminonaphthalene, 2-aminofluorene, and 1-aminopyrene) that are metabolites of nitro-PAHs. For analysis of these PAH features, two mL of urine were incubated with 20 μL of β-glucuronidase from Helix pomatia Type H-2 (Sigma-Aldrich, St. Louis, MO) in 2 mL of 0.1 M Na Acetate buffer (pH 5.0) at 37°C overnight. A sample enrichment and purification cartridge, packed with C18 was used for sample extraction and was eluted with 6 mL of methanol. The eluent was concentrated by evaporation and 20 μL was injected into the LC-MS/MS system consisting of a Nova-Pak C18 column (Waters 3.9 x 150 mm) and a Thermo TSQ Quantum Access Max triple quadrupole mass spectrometer with an electrospray ionization interface operated in negative ion mode with a spray voltage of 4 kV. Analytes were separated with a methanol/water mobile phase at a flow rate of 0.2 mL/min using stepwise elution starting with 55% methanol for 2 min., followed by 80% methanol for 25 min., then returning to 55% methanol for 10 min. The capillary temperature was optimized at 300°C, the sheath-gas pressure was 10 arbitrary units, and the auxiliary gas pressure was 5 arbitrary units. The selected reaction monitoring (SRM) transitions *m/z* 193.12/165.08 for hydroxyphenanthrenes and 199.20/171.20 for 13C6 3-hydroxyphenanthrene (internal standard) were used for detection.

C.2 Supplementary Figures and Tables

Feature	RT, min	MIM observed (*m/z*, +3)	MIM theoret. (*m/z*, +3)	ΔMass (ppm)	Added mass (+H, Da)	Elemental composition	Annotation	Mean peak area (x10^6)
796.43[a,b,c, f,g,h]	27.717	796.4298	796.4301	-0.3767	-45.983	$-CH_2S$	Cys34→Gly	6.16
805.76[a,b,c,f,h]	27.365	805.7614	805.7618	-0.4964	-17.989	$-SH_2$, +O	Cys34→oxoalanine or fGly	0.182
808.73[a,b,c,d,e,f,g,h]	27.868	808.7281			-9.088		Not Cys34 adduct	42.9
810.45[a,d,f]	28.263	810.4521			-3.916		Not Cys34 adduct	0.168
811.42[a,b,c,d,e,f,g,h]	30.803	811.4242	811.4234	0.9859	-1.000		T3 dimer*	3.45
811.76[a,b,c,d,e,f,g,h]	28.682	811.7576	811.7594	-2.2174	0.000		Unmodified T3*	9.25
816.42[a,b,c,d,e,f,g,h]	27.390	816.4188	816.4191	-0.3675	13.984	$-H_2$, +O	Cys34-Gln crosslink*	1.39
816.43[a,b,c,d,e,f,g,h]	29.102	816.4306	816.4312	-0.7349	15.027	$+CH_3$	Methylation (at Glu37)	3.94
819.09[b,e,g]	28.514	819.0863			22.994		Na adduct of T3	0.469
822.42[a,b,c,d,e,f,g,h]	27.207	822.4222	822.4226	-0.4864	33.002	$+HO_2$	Cys34 sulfinic acid*	16.4
827.088[b,c,f,g,h]	29.566	827.0883	827.0886	-0.3627	47.000	$+CH_3S$	S-Methanethiol*	40.9
827.094[c,d,f,g,h]	27.790	827.0942	827.0945	-0.3627	47.018	$+CH_3O_2$	S-(O)-O-CH_3	1.23
827.75[a,b,c,d,e,f,g,h]	27.361	827.7537	827.7543	-0.7249	48.996	$+HO_3$	Cys34 Sulfonic acid*	2.08
829.390[a,b,g,h]	28.575	829.3959			53.923		Not Cys34 adduct	0.321
835.11[a,c,e,,g,f,h]	28.809	835.1062	835.1066	-0.4790	71.054	$+C_4H_7O$	Crotonaldehyde*	1.58
841.75[a,b,c,d,e,f,g,h]	28.411	841.7513	841.7519	-0.7128	90.989	$+C_2H_3O_2S$	S-Mercaptoacetic acid	0.061
842.07	28.121	842.0734			91.955		Unknown	NA
845.42[a,b,c,d,e,f,g,h]	27.383	845.4238	845.4239	-0.1183	102.006	$+C_3H_4NOS$	S-Cys (−H_2O)	0.487
849.07[a,b,c,h]	28.440	849.0684	849.0689	-0.5889	112.940	$+HO_3S_2$	S-Sulfonic acid trisulfide	0.880
851.43[a,b,c,d,e,f,g,h]	26.980	851.4266	851.4274	-0.9396	120.015	$+C_3H_6NO_2S$	S-Cys*	67.2
851.76[a,b,c,d,e,f,g]	27.632	851.7554	851.7554	0.0000	121.000	$+C_3H_5O_3S$	S-Cys(NH2→OH)	1.23
853.78[a,b,f]	27.503	853.7823			127.082		Unknown	9.20
856.10[a,b,c,d,e,f,g,h]	26.746	856.0988	856.0993	-0.5840	134.032	$+C_4H_8NO_2S$	S-hCys*	200
858.75[a,b,c,d,e,g,h]	26.391	858.7546			141.999	$+C_3H_5NO_2SNa$	Na adduct of S-Cys	1.52
860.77[b,d,e,f,g]	26.866	860.7703	860.7712	-1.0456	148.046	$+C_5H_{10}NO_2S$	S-hCys (+CH3)	5.10
864.08[a,b,c,e,f,g]	26.305	864.0763			157.964		Not Cys34 adduct	0.760
870.43[a,b,c,d,e,f,g,h]	26.031	870.4342	870.4345	-0.3447	177.038	$+C_5H_9N_2O_3S$	S-CysGly*	0.751
894.44[a,b,c,d,e,f,g,h]	26.684	894.4414	894.4416	-0.2236	249.059	$+C_8H_{13}N_2O_5S$	S-γ-GluCys*	0.092
910.18[g,h]	32.954	910.1767			296.265	$+C_{18}H_{34}NO_2$	Unknown	0.009
913.45[a,b,c,d,e,f,g,h]	26.554	913.4487	913.4487	0.0000	306.081	$+C_{10}H_{16}N_3O_6S$	S-Glutathione*	0.250
914.83[h]	32.931	914.8359			310.243		Unknown	0.552
931.82[b,c,d,f,g]	24.921	931.8204			361.196		Unknown	0.685
965.49[b,c,d,e,f,h]	25.207	965.4910			462.208		Unknown	2.24
970.16[d,f,h]	25.473	970.1631			476.224		Unknown	NA

Legend: MIM, monoisotopic mass; NA, not available; RT, retention time.
* Identity confirmed with a synthetic standard.
[a] Grigoryan et al, *Analyt. Chem.* 2016, 88, 10504−10512.
[b] Lu et al, *Environ. Sci. Technol.,* 2017 Jan 3;51(1), 46-57.
[c] Grigoryan, et al., *Carcinogenesis*, 2018, 39(5), 661-668.
[d] Liu et al, *Environ. Sci. Technol.,* 2018, 52, 2307−2313.
[e] Dagnino et al., *Int. J. Cancer,* 2020, 146, 3294–3303.
[f] Grigoryan et al., *Cancer Res.* 2019;79, 6024–6031.
[g] Wong et al., *Environ. Toxicol. Pharmacol,* 2022;95,103966.
[h] Grigoryan et al, *Cancer Epidemiol Biomarkers Prev, 2023, 32, 1217-26.*

Figure C.1: T3 peptide (mostly Cys34) features with sufficient data for statistical analysis. (Mean peak areas are based on normalized data after imputation for missing values).

Feature	RT, min	MIM observed (*m/z*, +3)	MIM theoretical (*m/z*, +3)	ΔMass (ppm)	Added mass (+H, Da)	Elemental composition (+H)	Annotation	Mean peak area (x10^8)
500.81[a,b,c]	12.063	500.8045	500.8055	-1.9968	-128.095	$-C_6H_{12}N_2O$	Loss of Lysine	156
556.34[a,b,c]	10.701	556.3396	556.3397	-0.1797	-17.024	$-NH_3$	Loss of ammonia	0.037
556.84	10.995	556.8430	556.8436	-1.0775	-16.017	$-NH_2$	Deamination	0.309
564.85[a,b,c]	10.822	564.8518	564.8529	-1.8474	0.000		Lys525 containing peptide	7.96
566.77[a,c]	10.684	566.7741			4.853		Unknown	0.014
571.84[c]	11.074	571.8423	571.8426	-0.5246	14.989	$-H_2,+O$	Lys525 oxidation product	3.87
577.86[a]	12.454	577.8607	577.8608	-0.1731	27.026	$+C_2H_3$	Acetaldehyde	1.91
578.32	12.541	578.3164			27.937		Unknown	0.036
580.85[c]	11.110	580.8476	580.8479	-0.5165	32.999	$+O_2H$	Lys525 oxidation product	NA
586.36[a,b,c]	13.265	586.3557	586.3559	-0.3411	44.016	$+CH_2NO$	Carbamylation	9.59
587.31[c]	13.362	587.3109			45.926		Unknown	1.01
645.88[a,b,c]	10.904	645.8791	645.8794	-0.4645	163.063	$+C_6H_{11}O_5$	Fructosyl lysine (glycation)	18.4
647.34[c]	10.778	647.3378			165.980		Unknown	1.03

Legend: MIM, monoisotopic mass; NA, not available; RT, retention time.

[a] Grigoryan et al., *Chem. Res.Toxicol.*, 2021,34, 2549−2557.

[b] Wong et al, *Environ. Toxicol. Pharmacol*. 2022, 95, 103966.

[c] Grigoryan et al, *Cancer Epidemiol. Biomarkers Prev., 2023, 32, 1217-26.*

Figure C.2: Lys525 peptide features with sufficient data for statistical analysis. (Mean peak areas are based on normalized data after imputation for missing values).

Feature	ICC	CV	Feature	ICC	CV
500.81	0.497	0.285	827.10	0.425	0.370
556.34	0.904	0.291	827.75	0.579	0.312
556.84	0.742	0.300	829.39	0.430	0.387
564.85	0.614	0.254	835.11	0.929	0.322
566.77	0.387	0.556	841.75	0.465	0.403
571.84	0.655	1.067	842.07	0.384	0.432
577.86	0.348	0.483	845.42	0.611	0.348
578.32	0.923	0.267	849.07	0.445	0.447
580.85	0.492	1.429	851.43	0.771	0.185
586.36	0.352	0.638	851.75	0.398	0.437
587.31	0.875	0.270	853.78	0.529	0.264
645.88	0.756	0.234	856.10	0.679	0.255
647.34	0.577	0.444	858.75	0.555	0.397
796.43	0.793	0.197	860.77	0.548	0.334
805.76	0.675	0.238	864.08	0.501	0.481
808.73	0.889	0.308	870.43	0.704	0.280
810.45	0.501	0.354	894.44	0.260	0.519
811.76	0.715	0.254	910.18	0.408	0.795
811.42	0.915	0.316	913.45	0.427	0.490
816.42	0.498	0.372	914.83	0.335	0.732
816.43	0.518	0.388	931.82	0.629	0.388
819.09	0.492	0.444	965.49	0.721	0.446
822.42	0.619	0.242	970.16	0.370	0.421
827.09	0.448	0.512			
Median ICC:	**0.552**		**Median CV:**	**0.380**	

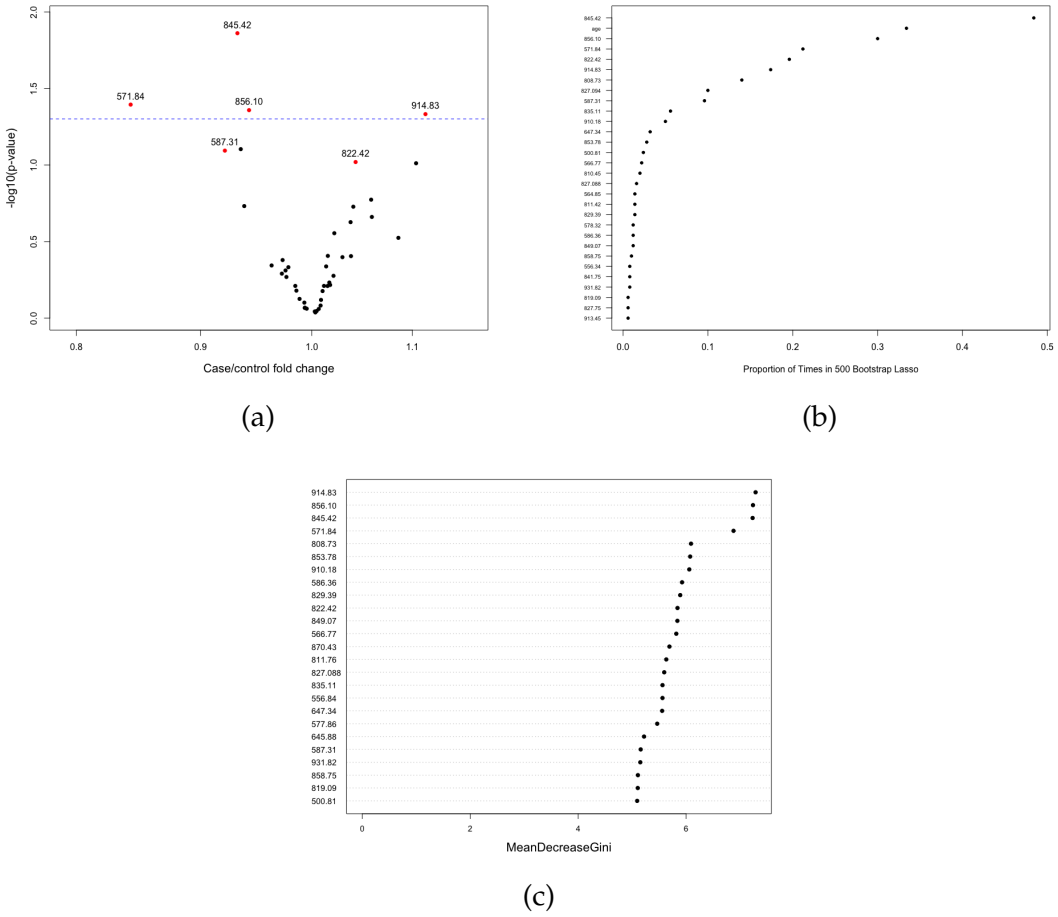
Table C.1: Intraclass correlation coefficients (ICC) and coefficients of variation (CV) from the random effects model for features abundances from duplicate nLC-HRMSMS injections.

Feature	**FC**	***p*-value**	**Feature**	**FC**	***p*-value**
500.81	1.015	0.392	827.09	0.936	0.453
556.34	1.012	0.616	827.10	1.040	0.187
556.84	1.010	0.666	827.75	1.017	0.607
564.85	1.014	0.459	829.39	0.973	0.417
566.77	0.972	0.512	835.11	1.037	0.236
571.84	0.842	0.040	841.75	0.993	0.857
577.86	1.003	0.918	845.42	0.932	0.014
578.32	1.085	0.299	849.07	0.935	0.079
586.36	1.059	0.218	851.43	1.003	0.905
587.31	0.921	0.081	851.75	1.005	0.899
645.88	1.009	0.760	853.78	1.022	0.279
647.34	1.058	0.168	856.10	0.942	0.044
796.43	0.985	0.661	858.75	1.029	0.400
805.76	0.995	0.868	860.77	1.017	0.586
808.73	0.938	0.185	864.08	1.008	0.828
810.45	0.978	0.465	870.43	0.984	0.616
811.76	0.993	0.792	894.44	0.988	0.749
811.42	1.038	0.394	910.18	1.104	0.097
816.42	1.015	0.617	913.45	1.007	0.871
816.43	0.993	0.855	914.83	1.114	0.047
819.09	0.975	0.488	931.82	0.976	0.538
822.42	1.042	0.096	965.49	1.021	0.529

Table C.2: Case-control fold changes (FC) and nominal *p*-values after normalization and imputation for HC lung cancer cases and controls.

Feature	**FC**	***p*-value**	**Feature**	**FC**	***p*-value**
500.81	1.014	0.471	827.09	0.999	0.992
556.34	1.006	0.804	827.10	1.035	0.291
556.84	1.004	0.864	827.75	1.024	0.522
564.85	1.013	0.508	829.39	1.003	0.939
566.77	0.956	0.344	835.11	1.027	0.467
571.84	0.823	0.040	841.75	1.024	0.583
577.86	0.985	0.671	845.42	0.937	0.047
578.32	1.060	0.520	849.07	0.912	0.028
586.36	1.033	0.539	851.43	1.022	0.394
587.31	0.896	0.040	851.75	1.030	0.444
645.88	1.005	0.865	853.78	1.023	0.306
647.34	1.050	0.297	856.10	0.946	0.087
796.43	0.989	0.764	858.75	1.070	0.087
805.76	1.014	0.671	860.77	0.988	0.685
808.73	0.950	0.352	864.08	1.019	0.650
810.45	0.978	0.503	870.43	1.014	0.691
811.76	0.989	0.715	894.44	0.995	0.900
811.42	1.044	0.374	910.18	1.106	0.127
816.42	1.012	0.718	913.45	1.011	0.797
816.43	0.979	0.603	914.83	1.113	0.073
819.09	0.981	0.629	931.82	1.004	0.921
822.42	1.018	0.504	965.49	1.009	0.808

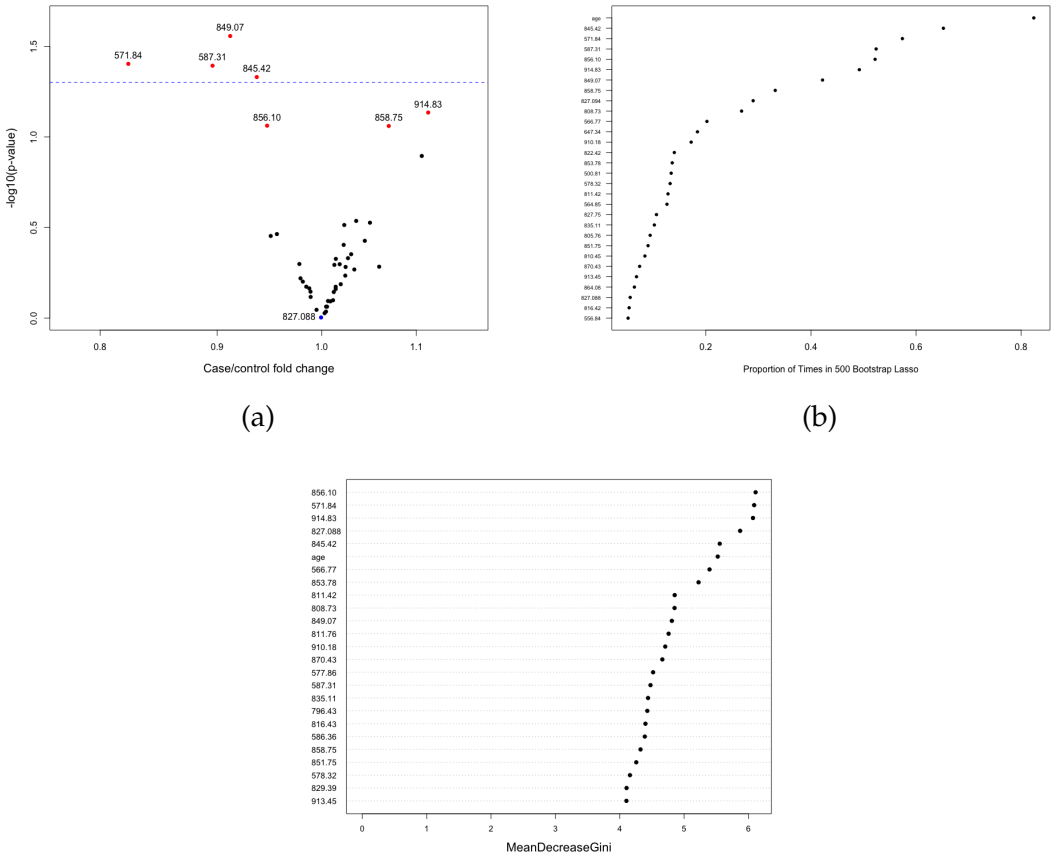
Table C.3: Case-control fold changes (FC) and nominal *p*-values after normalization and imputation for LUAD cases and controls.



(a) (b)

(c)

Figure C.3: Ensemble of regression and classification methods for selection of adduct features associated with case-control status for HC cases. A) Volcano plot of nominal p-values for the case-controls status regression coefficient in a multivariate linear regression of each adduct's log abundance on case-control status and other covariates, with selected adducts shown in red (dashed line indicates nominal p-value=0.05); B) proportion of times that a given adduct was selected by a regularized logistic regression (LASSO) of HC case-control status; C) ranked variable importance measures from a Random Forests classification of case-control status.



(a) (b)

(c)

Figure C.4: Ensemble of regression and classification methods for selection of adduct features associated with case-control status for LUAD cases. a) Volcano plot of nominal *p*-values for the case-control status regression coefficient in a multivariate linear regression of each adduct's log abundance on case-control status and other covariates, with selected adducts shown in red (dashed line indicates p=0.05); b) proportion of times that a given adduct was selected by a regularized logistic regression (LASSO) of LUAD case-control status; c) ranked variable importance measures from a Random Forests classification of case-control status.

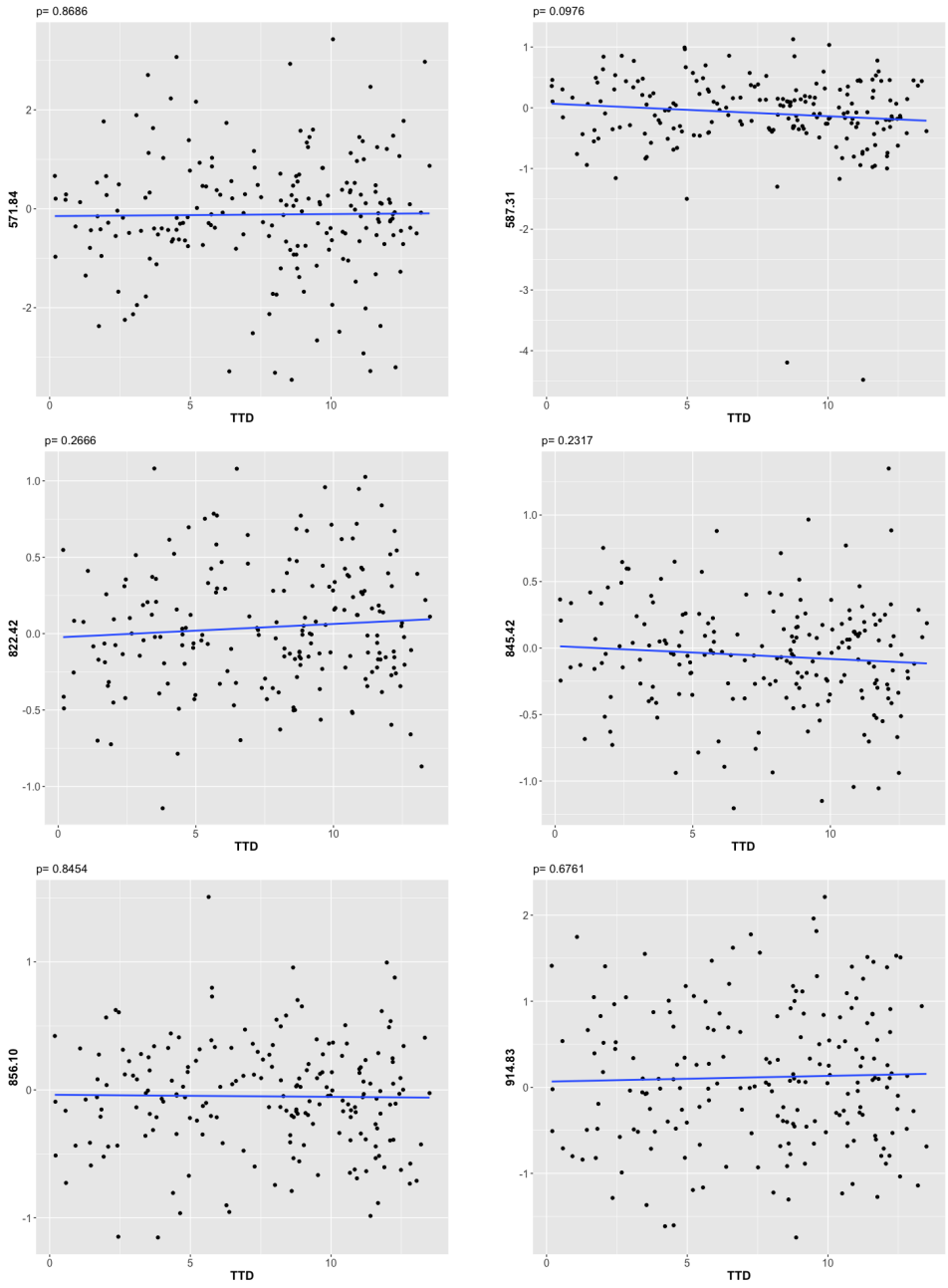


Figure C.5: Plots of ln(case-matched-control fold change) versus time to diagnosis (ttd) in years from recruitment for adduct features selected for HC cases-controls. *P*-values are shown for slopes of the relationship.

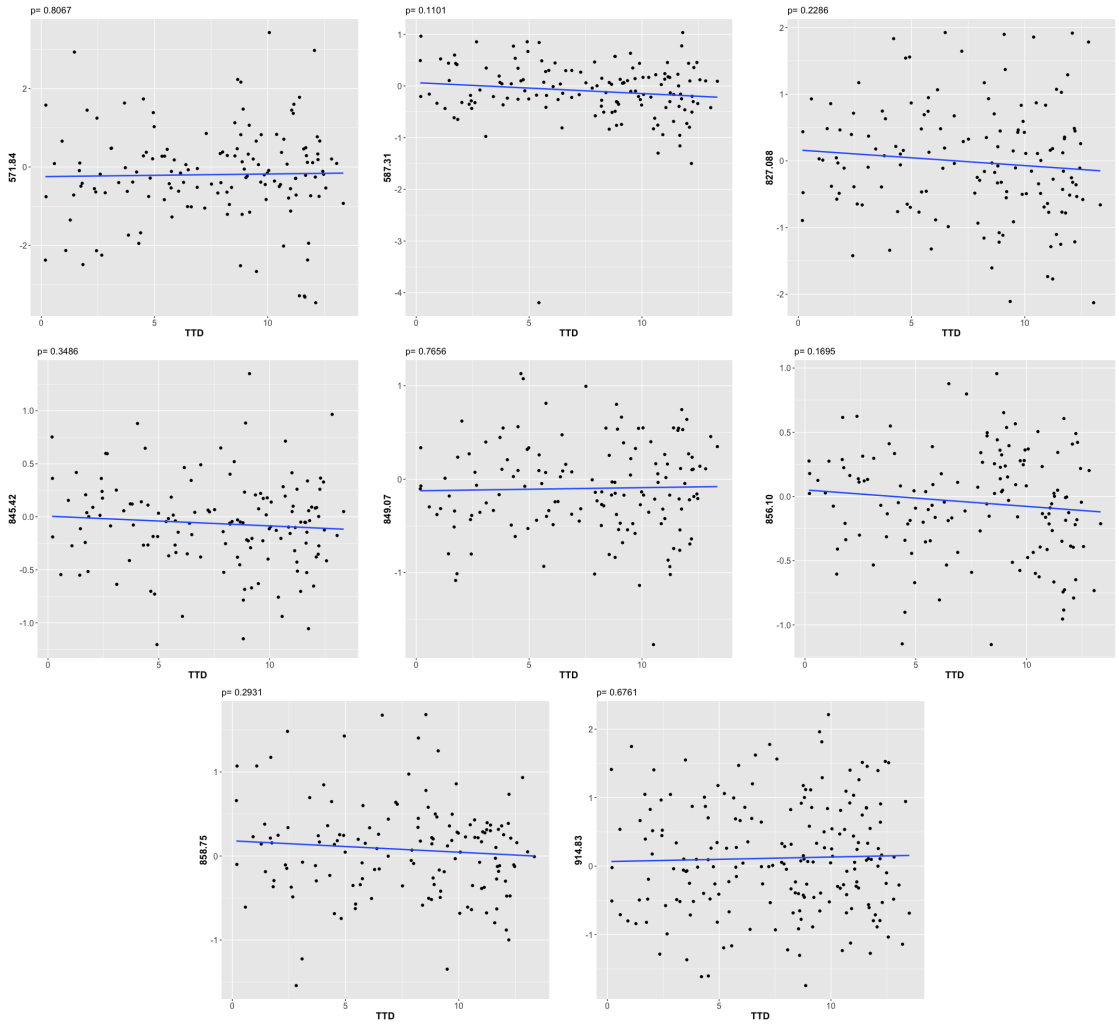


Figure C.6: Plots of ln(case-matched-control fold change) versus time to diagnosis (ttd) in years from recruitment for adduct features selected for LUAD cases-controls. *P*-values are shown for slopes of the relationship.