# UC Berkeley

## UC Berkeley Electronic Theses and Dissertations

### Title

Untargeted Adductomics of Newborn Dried Blood Spots to Identify Exposures Associated with Childhood Leukemia

### Permalink

https://escholarship.org/uc/item/9q0290zp

### Author

Yano, Yukiko

### Publication Date

2018

Peer reviewed|Thesis/dissertation

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

# Untargeted Adductomics of Newborn Dried Blood Spots to Identify Exposures Associated with Childhood Leukemia

By

Yukiko Yano

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Health Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Stephen M. Rappaport, Chair
Professor Martyn T. Smith
Professor Evan R. Williams
Professor John R. Balmes

Fall 2018

Untargeted Adductomics of Newborn Dried Blood Spots
to Identify Exposures Associated with Childhood Leukemia

© 2018
by Yukiko Yano

Abstract

Untargeted Adductomics of Newborn Dried Blood Spots
to Identify Exposures Associated with Childhood Leukemia

by

Yukiko Yano

Doctor of Philosophy in Environmental Health Sciences

University of California, Berkeley

Professor Stephen M. Rappaport, Chair

The causes of cancer are likely to involve a combination of genetic factors, exposures, and random chance. However, current evidence suggests between a third and a half of cancers may be preventable, and that exposures may play an important role in the etiology of cancer. In fact, it is possible that the most common type of cancer among children, childhood leukemia, may be a preventable disease. Current scientific knowledge suggests early life exposures, including exposures occurring *in utero*, play critical roles in the development of childhood leukemia. However, few risk factors have been established, and the underlying disease mechanisms remain elusive.

The overarching goal of this dissertation research was to address this gap in knowledge by performing an untargeted adductomics analysis of archived newborn dried blood spots (DBS) to identify risk factors for childhood leukemia resulting from *in utero* exposures. Metabolism of chemicals derived from the diet, exposures to xenobiotics, the microbiome, and lifestyle factors (*e.g.*, smoking, alcohol consumption) produce electrophiles that react with nucleophilic sites in circulating proteins, notably Cys34 of human serum albumin (HSA). Since HSA has a residence time of 28 days, HSA-Cys34 adducts measured in archived newborn DBS reflect systemic exposures during the last month of gestation and allow us to explore potential risk factors resulting from *in utero* exposures.

Here, we have developed an untargeted adductomics method to detect HSA-Cys34 adducts in 4.7-mm punches from DBS (equivalent to 5 – 8 $\mu$L of whole blood). The workflow includes extraction of proteins from DBS, measurement of hemoglobin to normalize for blood volume, addition of methanol to enrich HSA by precipitation of hemoglobin and other interfering proteins, digestion with trypsin, and detection of HSA-Cys34 adducts via nanoflow liquid chromatography-high resolution mass spectrometry. As proof-of-concept, we tested the DBS-adductomics method with 49 archived DBS collected from newborns whose mothers either actively smoked or were nonsmokers during pregnancy. A novel normalization method ('scone') was used to remove unwanted technical variation arising from: HSA digestion, blood volume, DBS age, mass spectrometry analysis, and batch effects. A total of 26 HSA-Cys34 adducts were detected, including Cys34 oxidation products, mixed disulfides with low-molecular-weight thiols

(*e.g.*, cysteine, homocysteine, glutathione, cysteinylglycine, etc.), and other modifications. Using an ensemble of variable selection methods, including both linear and nonlinear models, we found that the Cys34 adduct of cyanide consistently discriminated between newborns of smoking and nonsmoking mothers with a fold change (smoking/nonsmoking) of 1.31 and a cross-validated area under the estimated receiver operating characteristic curve (cvAUC) of 0.79. Indeed, hydrogen cyanide is a component of cigarette smoke, and these results indicated that DBS-based adductomics is suitable for investigating *in utero* exposures to reactive electrophiles that may influence disease risks later in life.

We then applied the DBS-adductomics method to analyze 783 archived newborn DBS collected from childhood leukemia cases and matched controls participating in the California Childhood Leukemia Study. Childhood leukemia cases included the two main subtypes, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), with additional molecular subtypes for ALL. After data preprocessing and normalization, a combination of linear and nonlinear models were used to identify adducts that discriminated between leukemia cases and controls. Of the 28 Cys34 adducts that were detected, none was predictive of ALL overall. However, several adducts showed increases in adduct abundances in subgroups of cases with T-cell ALL and B-cell ALL with t(12;21) translocations. For both subgroups, elevated levels of adducts of reactive carbonyl species in cases may suggest that oxidative stress and lipid peroxidation had influenced adduct production. Regarding AML, the Cys34 homocysteine adduct (with loss of $H_2O$) was found to consistently discriminate between AML cases and controls with a fold change (case/control) of 0.66. Since homocysteine is an important intermediate in the folate-mediated one-carbon metabolism, this indicates alterations in epigenetic regulations and folate status may be involved in the etiology of AML. Moreover, because lower homocysteine levels were detected in newborn DBS collected years before AML cases were diagnosed, biological changes involved in the initiation of AML may be present at birth and that there may be avenues for preventing the disease. Since this was a hypothesis-generating study, these findings warrant replication in follow-up studies with larger sample sizes of the various subtypes. Future integrated analyses with other omics (*e.g.*, genomics, metabolomics, epigenomics) will be key to obtaining a full picture of the disease mechanisms regarding childhood leukemia.

*To Eric and Aika*

# Table of Contents

# List of tables

## Chapter 2.



## Chapter 3.



## Chapter 3 Supplemental Tables.



## Chapter 4.



## Chapter 4 Supplemental Tables.

# List of figures

## Chapter 2.



## Chapter 2 Supplemental Figures.



## Chapter 3.



## Chapter 3 Supplemental Figures.

## Chapter 4.



## Chapter 4 Supplemental Figures.

# List of abbreviations

| | |
|---|---|
| ALL | Acute lymphoblastic leukemia |
| AML | Acute myeloid leukemia |
| ANBS | Archived newborn dried blood spots |
| ANOVA | Analysis of variance |
| AUC | Area under the curve |
| CBS | Cystathionine β-synthase |
| CCLS | California Childhood Leukemia Study |
| CI | Confidence interval |
| CID | Collision-induced dissociation |
| cvAUC | Cross-validated area under the estimated receiver operating characteristic curve |
| Cys34 | Cysteine amino acid at position 34 in human serum albumin |
| DBS | Dried blood spots |
| E:P | Enzyme-to-protein ratio |
| FC | Fold change |
| FDR | False discovery rate |
| GWAS | Genome-wide association studies |
| Hb | Hemoglobin |
| Hct | Hematocrit |
| HSA | Human serum albumin |
| IAA-iT3 | Carbamidomethylated isotopically labeled T3 |
| $K^+$ | Potassium |
| lasso | Least absolute shrinkage and selection operator |
| MIM | Monoisotopic mass |
| MS | Mass spectrometry |
| MS2 | Tandem mass spectrometry |
| *m/z* | Mass-to-charge ratio |
| nLC-HRMS | nanoflow liquid chromatography-high resolution mass spectrometry |
| OR | Odds ratio |
| RCS | Reactive carbonyl species |
| RLA | Relative log abundance |
| ROC | Receiver operating characteristic |
| ROS | Reactive oxygen species |
| RT | Retention time |
| RUV | Removal of unwanted variation |
| SNP | Single nucleotide polymorphisms |
| T3 | The third largest tryptic peptide of human serum albumin containing Cys34 |

# Acknowledgements

I would like to thank the many people who have contributed to the work presented in this dissertation.

First, I would like to express my sincere gratitude to my advisor, Dr. Stephen Rappaport, for his valuable guidance throughout my Ph.D. His work inspired me to pursue this degree, and I thank him for giving me this incredible opportunity to be a part of an exciting and important research field. I would also like to thank my committee members: Dr. Martyn Smith, Dr. Evan Williams, and Dr. John Balmes for their scientific guidance and external insight. Dr. Martyn Smith has been a vital part of my Ph.D., and I am particularly grateful for the tremendous amount of support he has provided me throughout the years.

Thank you to the Center for Integrative Research on Childhood Leukemia and the Environment (CIRCLE) for making this study possible. Special mention goes to Dr. Catherine Metayer, Dr. Todd Whitehead, Alice Kang, and Lucie McCoy for their collaboration and for providing me with necessary resources for this work.

I am appreciative of other faculty and staff in the Department of Environmental Health Sciences who have supported me in many ways. I would like to give a special thank you to Dr. Kathie Hammond, who has been an incredible mentor to me. She has been a great source of wisdom, strength, and motivation. I owe many thanks to Norma Firestone, who has not only helped me navigate through the Ph.D. program but has also continuously advocated for my success. I am also deeply appreciative of Dr. Ellen Eisen, who provided me with the support I needed to complete this degree.

I was very fortunate to have worked with incredibly talented and intelligent fellow lab members. Former members, Dr. Lauren Petrick and Dr. Kelsi Perttula, were amazing mentors and their friendship has been invaluable. Courtney Schiffman's statistical insight was vital to this work, and she has been a truly inspiring teammate and friend. I'm very grateful for the many insightful discussions with Dr. Josie Hayes and Dr. Henrik Carrlson who have given me continuous encouragement and motivation. I thank Dr. Hasmik Grigoryan for her expertise in mass spectrometry and for guiding me through the challenges of method development. Dr. William Edmands made profound contributions developing the bioinformatics for this adductomics work. His work has been key to much of the progress in adductomics, and I will always remember his humor and laughter.

I am deeply indebted to my friends and family for their continuous support. My fellow graduate students, Erica Fuhrmeister, Emily Marron, Daniel Drew, Jorge Marchand, Dr. Joey Greenspun, Riley Scheid, and Hannah Evans have cared for me and always kept me smiling and laughing. I will always cherish their friendship. None of this would have been possible without my loving sister, Aika. I would not have left Japan to pursue my goals without her encouragement. And finally, to Eric. Thank you for your endless love and support. You've always believed in me.

# Chapter 1.

# Introduction

## 1.1 Causes of cancer – is cancer preventable?

The global disease burden continues to shift from communicable diseases to cancer and other non-communicable diseases (*e.g.*, cardiovascular diseases, chronic respiratory diseases, and diabetes).[1,2] Cancer is becoming the predominant challenge in global health as the leading driver of morbidity, mortality, and health-care costs.[1,3,4] Cancer risk involves a composite of inherited genes, exposures, and random chance.[5–8] Here, exposures broadly encompass non-genetic factors including chemical exposures from the diet, xenobiotics (*e.g.*, occupational or environmental pollutants), the microbiome, and lifestyle factors such as smoking and alcohol consumption. While the relative contributions of genes, exposures, and chance have been debated for some time,[5,6] it is clear that cancer is a multifactorial disease that probably manifests as a result of gene-environment interactions.[7,9] In the last decade, genome-wide association studies (GWAS) have identified many genetic variants associated with cancer, but most common variants individually or in combination showed modest effects on cancer risk (odds ratios 1.1 – 1.5).[10–12] In fact, familial and twin studies investigating the heritability of cancers [13,14] suggest exposures may play a predominant role in cancer etiology. This is in agreement with epidemiological evidence showing the large geographical and temporal variations in incidence rates of specific cancers in different parts of the world,[4,15] and from the observed increase in cancer incidence when populations migrated from a low-risk to a high-risk country.[16]

Evidence suggests that while both genetics and exposures are causal components of cancer,[7] the vast majority of cancers may be preventable.[6,9,17–19] Most cancers require more than one genetic mutation in order to progress into the clonal expansion of cancer cells which result in overt cancer.[20,21] Premalignant clones occur more frequently than clinical cancer rates,[21] meaning that cancer is commonly initiated but rarely promoted to full malignancy.[22] This observation along with the complexity of cancer mutations and the long time period often required for cancer symptoms to emerge (anywhere between 1 and 50 years), suggest cancer is promoted through sequential events in which additional mutational and epigenetic changes occur.[21,23] This multi-stage model of carcinogenesis presents opportunities for preventive interventions to disrupt cancer development. Ideally, cancer can be prevented by avoiding or reducing exposures that can directly or indirectly lead to cancer-initiating or promoting mutations.[7] Identification of promotional exposures that drive the emergence of overt cancer among susceptible individuals with premalignant clones is particularly crucial.[21,23]

At the population level, between a third and a half of cancers are thought to be preventable based on current scientific knowledge.[17,18,24,25] Known modifiable risk factors include tobacco smoking and secondhand smoke exposure, excessive alcohol consumption, being overweight or obese, an imbalanced diet (*i.e.*, diets low in fruits and vegetables and high in processed food), physical inactivity, chronic infections (*e.g.*, human papillomavirus), and ultraviolet radiation exposure.[18,25] However, with the exception of a handful of specific cancer types such as lung cancer, in which >80% of the cases can be attributed to smoking, much of the phenotypic variability in most cancers remains unexplained.[26] In addition to this large unknown

group of risk factors, what is still lacking is an understanding of the cumulative effects of multiple exposures and how they contribute to cancer risk.[27] It is also necessary to address which exposures are most important and to identify critical periods (*e.g.*, *in utero*, puberty, pregnancy) during which individuals are most susceptible to exposures that contribute to cancer risk.[28]

## 1.2 Discovery of exposures impacting cancer risk using omics

In order to address these gaps in knowledge within cancer etiology, a discovery-based approach using recent advances in omics is better suited than reductionist approaches that focus on one-to-one relations between a specified exposure and a disease outcome.[26] Following the example of GWAS, there has been an emergence of various omics platforms (*e.g.*, metabolomics, proteomics, transcriptomics, epigenomics), which allow for an untargeted interrogation of exposures that impact human health.[29] Humans are continuously exposed to numerous chemicals of highly variable dynamic range[26,30] that originate from both endogenous (*e.g.*, human and microbial metabolism) and exogenous (*e.g.*, diet, xenobiotics in the environment and occupational settings, smoking, alcohol, etc.) sources. While the genome remains static through one's lifetime, these exposures can result in important changes in small-molecule metabolites, proteins, DNA and RNA over time.[31] Omics generally refers to the rigorous measurement of the global collection of such biological molecules in biospecimens using high-throughput techniques.[28] For example, advances in analytical instruments such as liquid chromatography-high resolution mass spectrometry have driven the recent progress in metabolomics studies which entail simultaneous measurements of thousands of small molecules in biological samples such as urine and blood.[32]

Omics can be readily applied to samples from epidemiologic case-control or cohort studies to explore the interplay between multiple exposures, identify unrecognized risk factors, and to discover intermediate biomarkers along the exposure-disease continuum that may help to better characterize the pathways from cancer initiation to development.[33] Specifically, if the biological samples are collected long before the onset of cancer, these intermediate biomarkers are likely to reflect the impact of the exposures. If the samples are collected closer to diagnosis, then the intermediate markers are likely to be associated with the progression of cancer.[34] One of the advantages of using omics to study complex diseases like cancer is that the findings are not limited by a single hypothesis known a priori. Given its agnostic approach, omics allows for the exploration of both known and unknown exposures, which may lead to the discovery of novel biomarkers resulting in the identification of unrecognized risk factors. Thus, omics provides opportunities to generate new hypotheses that may improve our understanding of the mechanisms of carcinogenesis.[35]

## 1.3 Adductomics to characterize exposures to reactive electrophiles

Although thousands of chemicals can be measured with current omics technologies, it is still not feasible to obtain accurate measurements of all chemicals in complex sample matrices like blood. Blood is a reservoir of chemical exposures, with concentrations of endogenous chemicals, food chemicals, drugs, and exogenous pollutants spanning 11 orders of magnitude.[30] With many diverse chemicals covering such a wide concentration range, it is impractical to measure everything. Therefore, classes of chemicals that are likely to have toxic effects should be prioritized for investigation. Reactive electrophiles (*i.e.*, electron-seeking species) are

inherently toxic because they can directly react with nucleophilic (*i.e.*, electron-rich) sites in DNA and proteins, thereby damaging these vital macromolecules.[36–38] These reactive species are constantly generated *in vivo* through metabolism of endogenous and exogenous chemicals. Although reactive electrophiles cannot be measured directly because of their short lifetimes in blood, their exposures can be characterized by measuring their stable reaction products (*i.e.*, adducts) with DNA or proteins.[39,40] Protein adducts have been studied for decades to monitor human exposures to reactive electrophiles from both endogenous (*e.g.*, lipid peroxidation products)[41] and exogenous (*e.g.*, occupational benzene exposure) sources.[42] However, these studies have been limited to targeted approaches focusing on electrophiles selected *a priori*. While these targeted analyses can unequivocally identify and quantify known exposures and generate dose-response relationships, they are not optimal for discovering the many unknown toxicants to which humans are exposed.

To characterize the totality of exposures to reactive electrophiles, our laboratory has developed an untargeted adductomics method, focusing on adducts bound to the highly nucleophilic Cys34 of human serum albumin (HSA).[43] HSA is the most abundant protein in plasma (~40 mg/mL, 60% of total plasma protein).[44] Of the many nucleophilic loci on proteins, cysteine thiol (-SH) groups are particularly important for their strong reactivity towards oxidants and electrophiles.[38] Cys34 is the only free thiol group within HSA, and it is a powerful scavenger of reactive species, accounting for more than 80% of the antioxidant activity in plasma.[45] The reactivity of HSA-Cys34 is attributed to its unusually low p$K_a$ (<6.7 compared to about 8.0 – 8.5 for thiols in other peptides and proteins), which causes Cys34 to exist primarily in the highly nucleophilic thiolate ($-S^-$) form at physiological pH of 7.4.[46,47] HSA also has a residence time of 28 days,[48] meaning that HSA-Cys34 adducts represent exposures integrated over the course of about one month. These qualities make HSA-Cys34 adducts ideal for exploring exposures to a wide range of reactive electrophiles.

Moreover, as the major extracellular antioxidant, the redox state of HSA may serve as a global biomarker for the redox state of the body in human diseases.[49] Oxidative stress, a condition with an increased amount of reactive oxygen species (ROS), has been related to the development of many chronic diseases including cancer.[50] Upon exposure to ROS, the HSA nucleophilic hotspot Cys34 (Cys34-SH: mercaptalbumin) is oxidized to sulfenic (Cys34-SOH), sulfinic (Cys34-$SO_2H$) and sulfonic (Cys34-$SO_3H$) acids. Sulfenic acid is an intermediate in the formation of disulfides (Cys34-S-SR) between Cys34 and cysteine, homocysteine, cysteinylglycine, glutathione, and other low-molecular-weight thiols.[45] In plasma of healthy young adults, 70 – 80% of HSA is in its reduced free thiol form (mercaptalbumin), and 20 – 30% of Cys34 forms reversible mixed disulfides with low-molecular weight thiols, with cysteinylated HSA being the predominant modification.[51] A small fraction of Cys34 is more highly oxidized to the sulfinic or sulfonic forms (2 – 5%),[49] which are likely to be irreversible.[52] Oxidation of HSA has been shown to increase with age.[53] Increases in oxidized HSA have also been observed in various chronic diseases and pathological conditions such as intrauterine growth restriction,[54] liver and kidney diseases,[55] diabetes mellitus,[55] and mild cognitive impairment and Alzheimer's disease.[56] In these studies, increases in oxidized forms of HSA have been correlated with decreases in mercaptalbumin, which may indicate the depletion and/or impairment of the antioxidant capacity of plasma.

In our HSA-Cys34 adductomics pipeline, HSA is first isolated from plasma or serum, digested with trypsin, and analyzed via nanoflow liquid chromatography-high resolution mass spectrometry (nLC-HRMS) to find and quantitate modifications at the third largest tryptic

peptide (T3) ([21]ALVLIAFAQYLQQC[34]PFEDHVK[41]), which contains the Cys34 site of modification.[43] In four previous studies, we applied this adductomics method to plasma/serum from healthy smokers and nonsmokers in the U.S.,[43] nonsmoking women in China exposed to indoor combustion products and local controls,[57] nonsmoking British patients with lung & heart disease and local controls,[58] and nonsmoking Chinese workers exposed to benzene and local controls.[59] Various HSA-Cys34 adducts have been detected in these studies, including Cys34 adducts with reactive oxygen and carbonyl species, mixed disulfides with low-molecular-weight thiols, and other modifications. Some adducts were associated with certain exposures such as Cys34 adducts of cigarette smoke constituents (*e.g.*, ethylene oxide and acrylonitrile)[43] and Cys34 adducts of benzene oxide and benzoquinones derived from occupational benzene exposures.[59] Cys34 sulfoxidation products and Cys34 mixed disulfides with low-molecular-weight thiols were commonly seen in all studies, and these adducts showed some indication of an altered HSA redox state among smokers,[43] subjects exposed to indoor combustion products,[57] and among subjects with chronic obstructive pulmonary disease and ischemic heart disease.[58]

## 1.4 Extending the adductomics assay to analyze archived newborn dried blood spots

While the adductomics assay was originally developed for the analysis of plasma and serum samples, we now wanted to extend this assay for the analysis of archived newborn dried blood spots (DBS). Newborn DBS are routinely collected by heel prick within 24 – 48 h of birth to screen for congenital disorders and inborn errors of metabolism in the U.S. and worldwide.[60] Almost every child in the U.S. participates in newborn screening, and some states, like California,[61] archive residual DBS from newborns for more than 21 years.[62] Thus, these archived newborn DBS provide an avenue for investigating the etiologies of diseases initiated *in utero*. For example, retrospective investigations of chromosome translocations in DNA from archived newborn DBS have provided direct evidence of the prenatal origin of childhood leukemia, the most common childhood cancer.[63–65] There is now increasing evidence suggesting that many non-communicable diseases in adult life, such as type 2 diabetes mellitus, cardiovascular disease, and the metabolic syndrome, can also have fetal origins.[66,67] Importantly, since HSA has a residence time of 28 days,[48] measuring Cys34 adducts in archived newborn DBS allows us to backtrack exposures to reactive electrophiles occurring during the month prior to birth. Therefore, Cys34 adducts may be able to directly capture fetal exposures impacting disease risks later in life.

One of the main issues preventing the widespread use of DBS in quantitative assays is the hematocrit (Hct) effect, which relates to variability in blood volume that results from the heterogeneity in hematocrit values across individuals.[68–70] Hct is a determinant of blood viscosity which affects the spreading of a blood drop through the filter paper and consequently leads to non-uniform blood volumes in fixed-size DBS punches.[68,71,72] Variation in blood volume can impact the accuracy and precision of the assay, and can decrease the statistical power to detect small differences between diseased and healthy populations.[71,72] Therefore, the first aim of this dissertation research was to develop a normalization technique to overcome the Hct issue by adjusting for differences in blood volume across DBS (described in Chapter 2). Here, we describe a simple and rapid normalization method employing UV-Visible spectroscopy to measure hemoglobin (Hb) concentrations in microliter volumes of extracts from DBS punches as a surrogate for Hct. The method was developed with DBS prepared with blood from adult volunteers and validated with archived newborn DBS. This normalization technique permitted

efficient adjustment for variation in blood volume across newborn DBS and preserved the DBS punches for downstream adductomics measurements.

Another major challenge of using newborn DBS for adductomics analysis involved the presence of interfering species from the sample matrix, specifically lysed red blood cells and Hb that are not abundant in serum or plasma.[73,74] Indeed, Hb has a 7-fold higher concentration than HSA in whole blood,[75] and the presence of Hb interfered with tryptic digestion of HSA that releases the T3 peptide containing Cys34 for analysis.[76] Thus, while Hb was crucial for blood volume normalization, Hb needed to be subsequently removed from the DBS extracts in order to measure HSA-Cys34 adducts. For this reason, the second aim of the dissertation research was to develop a sample processing method to remove Hb from DBS extracts prior to digestion and detection of HSA-Cys34 adducts via nLC-HRMS (described in Chapter 3). Extraction and digestion of HSA was optimized for the analysis of 4.7-mm punches from DBS (equivalent to 5 to 8-$\mu$L of whole blood). As proof-of-concept, we analyzed archived DBS from 49 newborns with mothers who either actively smoked during pregnancy or were nonsmokers. Data normalization was performed using a novel method called 'scone,'[77,78] and we used an ensemble of statistical methods to identify Cys34 adducts that discriminated between newborns of smoking and nonsmoking mothers.

The final aim of this dissertation research was to apply the developed DBS-adductomics method to analyze 783 archived newborn DBS collected from participants of a population-based childhood leukemia case-control study conducted in California (described in Chapter 4). As described in detail below, the purpose of this study was to use adductomics to discover potential risk factors for childhood leukemia resulting from *in utero* exposures.

## 1.5 Etiology of childhood leukemia

Leukemia is the most common cancer among children under the age of 15.[79] The two main types of childhood leukemia are acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). ALL and AML are leukemias characterized by the uncontrolled proliferation of early cells of the lymphoid (both B- and T-cells) or myeloid lineage, respectively.[80] ALL accounts for approximately 80% of leukemia cases in children while AML comprises 15% of cases.[81] Survival rates are as high as 90% for ALL but progress has been slower for AML, and many childhood leukemia survivors suffer long-term consequences of the treatments.[82] In the U.S., approximately 3,800 children are diagnosed with ALL or AML each year, and the overall incidence has increased significantly from 1975 to 2012 by ~1% per year.[83] Childhood leukemia is also more common in affluent countries such as the U.S. and Europe.[84] In these high income countries, ALL (particularly B-cell ALL) shows a marked incidence peak between ages 2 – 5 years, but a similar peak is not seen in lower income countries such as Africa and India.[19,85] Such temporal and geographic variations in incidence suggest exposures, rather than genetic factors, may be contributing to this upwards trend, and that causal factors for childhood leukemia are likely to have become more prevalent in the past few decades.[79,85] However, the only established risk factors are ionizing radiation and congenital genetic syndromes such as Down's, and these together explain less than 10% of the cases.[85]

Like many cancers, inherited genes are likely to contribute to the risk of developing childhood leukemia.[86] In the last decade, numerous studies have looked at the contribution of inherited genetic variants by examining associations between single nucleotide polymorphisms (SNPs) in candidate genes involved in xenobiotic metabolism, folate metabolism, DNA repair,

and immune function, but oftentimes with conflicting results.[87] GWAS utilizing high-throughput genotyping to test associations of SNPs across the entire human genome simultaneously, have identified risk loci for ALL in genes involved in lymphoid differentiation.[88] However, the molecular mechanism by which these variants are associated with childhood leukemia are unknown, and GWAS have shown that the risk associated with individual SNPs tend to be modest, with expected odds ratios between 1.1 and 1.6.[12,89] While additional genetic factors are likely to be discovered in the future, such modest effect sizes associated with genetic variants further implicate exposures may also be involved in the etiology of childhood leukemia.[79,89,90]

Epidemiologic studies of childhood leukemia have investigated perinatal/reproductive factors, as well as prenatal and early life exposures such as parental smoking and alcohol consumption, diet and supplement intake, home use of pesticides, parental occupational chemical exposures (*e.g.*, solvents, paint), and outdoor air pollution, but the findings have been mostly inconsistent.[90–97] On the other hand, there is increasing evidence showing that immune function and response to infections in early life, as well as dietary factors and commensal microbes may play critical roles in the development of childhood leukemia.[19] Improper immune development may result in an abnormal response to common infections in early life, which could trigger the development of childhood leukemia.[19] This hypothesis is corroborated by observations of a protective effect for day-care attendance,[90] breastfeeding,[98] and an increased risk in children born by caesarean delivery, where newborns do not experience initial microbial colonization by the mother's vaginal and fecal microbiota during birth.[99] However, a better understanding of the underlying disease mechanisms and identification of etiologic factors are still necessary in order to implement primary prevention of childhood leukemia.[79]

## 1.6 Heterogeneity of childhood leukemia

Childhood leukemia is a biologically heterogeneous disease, and both ALL and AML consist of a number of subtypes defined by cell lineage as well as molecular and cytogenetic characteristics, such as the presence of chromosome translocations and changes in chromosome number.[100] About 80% of ALL originate from precursor-B cells (B-cell ALL) and 20% are of T-cell origin (T-cell ALL).[101] The most common chromosomal abnormalities observed in B-cell ALL are hyperdiploidy (>50 chromosomes) and the *ETV6-RUNX1* (*TEL-AML1*) fusion gene generated by the t(12;21)(p13;q22) translocation.[102] Cytogenetic analyses have been rather uninformative for T-cell ALL, but T-cell ALL can be classified into distinct genetic subgroups corresponding to the T-cell development stage using gene expression profiling.[103–106] Classification of AML has primarily been based on morphology and immunophenotype. Seventy percent of AML cases belong to the following cytogenetic categories: normal karyotype; core binding factor AML (t(8;21) and inv(16)); acute promyelocytic leukemia (*PML-RARA*); and rearrangements involving the *MLL* gene.[107,108] The diversity of mutated genes in AML have become increasingly clear with recent technological improvements in DNA sequencing.[109–111] Currently, the most widely accepted classification and prognostic markers include *CEBPA*, *NPM1*, and *FLT3*-ITD, but there are likely to be many more genetic markers for AML classification.[109–111] Infant leukemias occurring in the first year of life is a biologically distinct subtype of childhood leukemia marked by their almost singular association with one particular genetic mutation - the fusion of the *MLL* gene with a variety of partner genes.[85,112] Classification of childhood leukemia cases into these various subtypes has important implications for both

diagnosis and prognosis, and it is hypothesized that each of these subtypes may have distinct etiologies.[89,91]

## 1.7 *In utero* origins of childhood leukemia

There is compelling evidence showing that the majority of childhood leukemia cases originate *in utero*.[19,90] Direct evidence for this was obtained by backtracking preleukemic translocations in archived newborn DBS or stored cord blood of children with leukemia up to 14 years before these individuals subsequently developed leukemia.[64,113] This has been confirmed for *ETV6-RUNX1* and *AML1-ETO*, the most common translocations for ALL and AML, respectively.[64] Furthermore, the protracted postnatal latency of the disease and the modest concordance rates for ALL and AML in monozygotic twin children (5 – 25%)[114] suggest that chromosome translocations occurring *in utero* may initiate leukemogenesis, but are insufficient by themselves to transition to overt leukemia.[115] In fact, preleukemic clones are present in 1% or more of newborns, but only 1% of them progress to overt leukemia.[116] Thus, the etiology of childhood leukemia is described by a two-hit model, where leukemia is initiated by common chromosomal rearrangements, but additional, rarer postnatal genetic or epigenetic mutations are required to transition to overt leukemia.[115]

Exposures occurring during fetal development may contribute to the initial genetic aberrations involved in the first hit, which can increase the risk of developing childhood leukemia. Fetuses are exposed to chemicals derived from endogenous process such as maternal, fetal, and microbial metabolism, and to exogenous chemicals from maternal exposures. The mother's diet, exposure to xenobiotics, and lifestyle factors, including smoking and alcohol consumption, can all influence fetal exposures. In addition, a number of xenobiotic chemicals (*e.g.*, tobacco smoke constituents, illicit drugs, prescription pharmaceuticals, alcohol, pesticides, heavy metals, etc.) have been shown to pass through the placental barrier, thereby exposing the fetus to potentially toxic chemicals.[117] Because fetuses have immature organ systems and their metabolic systems are not fully developed, they are particularly sensitive to *in utero* exposures that may result in genetic alterations impacting future disease risks.[67,118] Therefore, *in utero* exposures may play a critical role in the etiology of childhood leukemia, and identification of risk factors resulting from *in utero* exposures may have important implications for preventing the disease.

## 1.8 Application of adductomics to investigate *in utero* exposures associated with childhood leukemia

As mentioned earlier, the final aim of this dissertation research (described in Chapter 4) was to apply the DBS-based adductomics method (developed in work from Chapters 2 and 3) to discover potential risk factors for childhood leukemia resulting from *in utero* exposures. Untargeted measurements of HSA-Cys34 adducts were performed using archived newborn DBS collected from 783 participants in the California Childhood Leukemia Study, a previously described population-based case-control study.[119] Cys34 adducts were compared between 387 cases (including both ALL and AML) and 396 matched controls to identify adducts that discriminated the two groups. Hypothesizing that different leukemia subtypes have distinct etiologies, we performed separate analyses for ALL and AML, with additional stratified analyses for the major subtypes of ALL (*i.e.*, B-cell ALL, B-cell ALL with high-hyperdiploidy, B-cell

ALL with t(12;21) chromosome translocation, and T-cell ALL). We hypothesized that Cys34 adducts discriminating childhood leukemia cases and controls would reflect differences in *in utero* exposures to reactive electrophiles. Discriminating Cys34 adducts may represent adducts associated with the first hit, susceptibility to the second hit, or intermediate markers involved in the etiology of childhood leukemia. With a residence time of 28 days,[48] we anticipated that detected HSA-Cys34 adducts would capture a wide range of exposures to reactive electrophiles occurring during the last month of gestation. Given this was the first adductomics investigation of childhood leukemia, our goal was to select *in utero* exposures that may lead to new hypotheses about the etiology of childhood leukemia and to provide insight into the underlying mechanisms of the disease.

## 1.9 References


(1) Naghavi, M.; Abajobir, A. A.; Abbafati, C.; Abbas, K. M.; Abd-Allah, F.; Abera, S. F.; Aboyans, V.; Adetokunboh, O.; Ärnlöv, J.; Afshin, A.; et al. Global, Regional, and National Age-Sex Specifc Mortality for 264 Causes of Death, 1980-2016: A Systematic Analysis for the Global Burden of Disease Study 2016. *Lancet* **2017**, *390* (10100), 1151–1210.

(2) Franceschi, S.; Wild, C. P. Meeting the Global Demands of Epidemiologic Transition - The Indispensable Role of Cancer Prevention. *Mol. Oncol.* **2013**, *7* (1), 1–13.

(3) Bauer, U. E.; Briss, P. A.; Goodman, R. A.; Bowman, B. A. Prevention of Chronic Disease in the 21st Century: Elimination of the Leading Preventable Causes of Premature Death and Disability in the USA. *Lancet* **2014**, *384* (9937), 45–52.

(4) Bray, F.; Jemal, A.; Grey, N.; Ferlay, J.; Forman, D. Global Cancer Transitions According to the Human Development Index (2008–2030): A Population-Based Study. *Lancet Oncol.* **2012**, *13* (8), 790–801.

(5) Tomasetti, C.; Li, L.; Vogelstein, B. Stem Cell Divisions, Somatic Mutations, Cancer Etiology, and Cancer Prevention. *Science (80-. ).* **2017**, *355* (6331), 1330–1334.

(6) Wu, S.; Powers, S.; Zhu, W.; Hannun, Y. A. Substantial Contribution of Extrinsic Risk Factors to Cancer Development. *Nature* **2016**, *529* (7584), 43–47.

(7) Greaves, M. Cancer Causation: The Darwinian Downside of Past Success? *Lancet Oncol.* **2002**, *3* (4), 244–251.

(8) Davey Smith, G.; Relton, C. L.; Brennan, P. Chance, Choice and Cause in Cancer Aetiology: Individual and Population Perspectives. *Int. J. Epidemiol.* **2016**, *45* (3), 605–613.

(9) Thomas, D. Gene–environment-Wide Association Studies: Emerging Approaches. *Nat. Rev. Genet.* **2010**, *11* (4), 259–272.

(10) Manolio, T. A.; Collins, F. S.; Cox, N. J.; Goldstein, D. B.; Hindorff, L. a; Hunter, D. J.; McCarthy, M. I.; Ramos, E. M.; Cardon, L. R.; Chakravarti, A.; et al. Finding the Missing Heritability of Complex Diseases. *Nature* **2009**, *461* (7265), 747–753.

(11) Galvan, A.; Ioannidis, J. P. A.; Dragani, T. A. Beyond Genome-Wide Association Studies: Genetic Heterogeneity and Individual Predisposition to Cancer. *Trends Genet.* **2010**, *26* (3), 132–141.

(12) Hindorff, L. a; Sethupathy, P.; Junkins, H. a; Ramos, E. M.; Mehta, J. P.; Collins, F. S.; Manolio, T. a. Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits. *Proc. Natl. Acad. Sci.* **2009**, *106* (23), 9362–9367.

(13) Hemminki, K.; Lorenzo Bermejo, J.; Försti, A. The Balance between Heritable and Environmental Aetiology of Human Disease. *Nat. Rev. Genet.* **2006**, *7* (12), 958–965.

(14) Lichtenstein, P.; Holm, N. V.; Verkasalo, P. K.; Iliadou, A.; Kaprio, J.; Koskenvuo, M.; Pukkala, E.; Skytthe, A.; Hemminki, K. Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **2000**, *343* (2), 78–85.

(15) Fitzmaurice, C.; Allen, C.; Barber, R. M.; Barregard, L.; Bhutta, Z. A.; Brenner, H.; Dicker, D. J.; Chimed-Orchir, O.; Dandona, R.; Dandona, L.; et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-Years for 32 Cancer Groups, 1990 to 2015: A Systematic

Analysis for the Global Burden of Disease Study Global Burden . *JAMA Oncol.* **2017**, *3* (4), 524–548.
(16) Ziegler, R. G.; Hoover, R. N.; Pike, M. C.; Hildesheim, a; Nomura, a M.; West, D. W.; Wu-Williams, a H.; Kolonel, L. N.; Horn-Ross, P. L.; Rosenthal, J. F.; et al. Migration Patterns and Breast Cancer Risk in Asian-American Women. *J. Natl. Cancer Inst.* **1993**, *85* (22), 1819–1827.
(17) Vineis, P.; Wild, C. P. Global Cancer Patterns: Causes and Prevention. *Lancet* **2014**, *383* (9916), 549–557.
(18) Golemis, E. A.; Scheet, P.; Beck, T. N.; Scolnick, E. M.; Hunter, D. J.; Hawk, E.; Hopkins, N. Molecular Mechanisms of the Preventable Causes of Cancer in the United States. *Genes Dev.* **2018**, *32* (13–14), 868–902.
(19) Greaves, M. A Causal Mechanism for Childhood Acute Lymphoblastic Leukaemia. *Nat. Rev. Cancer* **2018**, *18* (8), 471–484.
(20) Knudson, A. G. Two Genetic Hits (More or Less) to Cancer. *Nat. Rev. Cancer* **2001**, *1* (2), 157–162.
(21) Greaves, M.; Maley, C. C. Clonal Evolution in Cancer. *Nature* **2012**, *481* (7381), 306–313.
(22) Greaves, M. Darwinian Medicine: A Case for Cancer. *Nat. Rev. Cancer* **2007**, *7* (3), 213–221.
(23) Umar, A.; Dunn, B. K.; Greenwald, P. Future Directions in Cancer Prevention. *Nat. Rev. Cancer* **2012**, *12* (12), 835–848.
(24) Anand, P.; Kunnumakara, A. B.; Sundaram, C.; Harikumar, K. B.; Tharakan, S. T.; Lai, O. S.; Sung, B.; Aggarwal, B. B. Cancer Is a Preventable Disease That Requires Major Lifestyle Changes. *Pharm. Res.* **2008**, *25* (9), 2097–2116.
(25) Colditz, G. A.; Wolin, K. Y.; Gehlert, S. Applying What We Know to Accelerate Cancer Prevention. *Sci. Transl. Med.* **2012**, *4* (127), 127rv4-127rv4.
(26) Manrai, A. K.; Cui, Y.; Bushel, P. R.; Hall, M.; Karakitsios, S.; Mattingly, C. J.; Ritchie, M.; Schmitt, C.; Sarigiannis, D. A.; Thomas, D. C.; et al. Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health. *Annu. Rev. Public Health* **2017**, *38* (1), 279–294.
(27) Vineis, P.; Khan, A. E.; Vlaanderen, J.; Vermeulen, R. The Impact of New Research Technologies on Our Understanding of Environmental Causes of Disease: The Concept of Clinical Vulnerability. *Environ. Health* **2009**, *8* (3), 54.
(28) Vineis, P.; van Veldhoven, K.; Chadeau-hyam, M.; Athersuch, T. J. Advancing the Application of Omics-Based Biomarkers in Environmental Epidemiology. *Environ. Mol. Mutagen.* **2013**, *00* (0), 00.
(29) Coughlin, S. S. Toward a Road Map for Global -Omics: A Primer on -Omic Technologies. *Am. J. Epidemiol.* **2014**, *180* (12), 1188–1195.
(30) Rappaport, S. M.; Barupal, D. K.; Wishart, D.; Vineis, P.; Scalbert, A. The Blood Exposome and Its Role in Discovering Causes of Disease. *Environ. Health Perspect.* **2014**, *122* (8), 769–774.
(31) Sun, Y. V; Hu, Y. Integrative Analysis of Multi-Omics Data for Discovery and Functional Studies of Complex Human Diseases. In *Advances in Genetics*; 2016; pp 147–190.
(32) Jones, D. P.; Park, Y.; Ziegler, T. R. Nutritional Metabolomics: Progress in Addressing Complexity in Diet and Health. *Annu. Rev. Nutr.* **2012**, *32*, 183–202.
(33) Vineis, P.; Perera, F. Molecular Epidemiology and Biomarkers in Etiologic Cancer

Research: The New in Light of the Old. *Cancer Epidemiol. Biomarkers & Prev.* **2007**, *16* (10), 1954–1965.
(34) Vineis, P.; Chadeau-Hyam, M. Integrating Biomarkers into Molecular Epidemiological Studies. *Curr. Opin. Oncol.* **2011**, *23* (1), 100–105.
(35) Rappaport, S. M. Biomarkers Intersect with the Exposome. *Biomarkers* **2012**, *17* (6), 483–489.
(36) Rappaport, S. M.; Smith, M. T. Environment and Disease Risks. *Science (80-. ).* **2010**, *330* (6003), 460–461.
(37) Liebler, D. C. Protein Damage by Reactive Electrophiles: Targets and Consequences. *Chem. Res. Toxicol.* **2008**, *21* (1), 117–128.
(38) Dennehy, M. K.; Richards, K. A. M.; Wernke, G. R.; Shyr, Y.; Liebler, D. C. Cytosolic and Nuclear Protein Targets of Thiol-Reactive Electrophiles. *Chem. Res. Toxicol.* **2006**, *19* (1), 20–29.
(39) Rubino, F. M.; Pitton, M.; Di Fabio, D.; Colombi, A. Toward an "Omic " Physiopathology of Reactive Chemicals: Thirty Years of Mass Spectrometric Study of the Protein Adducts with Endogenous and Xenobiotic Compounds. *Mass Spectrom. Rev.* **2009**, *28* (5), 725–784.
(40) Törnqvist, M.; Fred, C.; Haglund, J.; Helleberg, H.; Paulsson, B.; Rydberg, P. Protein Adducts: Quantitative and Qualitative Aspects of Their Formation, Analysis and Applications. *J. Chromatogr. B* **2002**, *778* (1–2), 279–308.
(41) Aldini, G.; Gamberoni, L.; Orioli, M.; Beretta, G.; Regazzoni, L.; Maffei Facino, R.; Carini, M. Mass Spectrometric Characterization of Covalent Modification of Human Serum Albumin by 4-Hydroxy-Trans-2-Nonenal. *J. Mass Spectrom.* **2006**, *41* (9), 1149–1161.
(42) Lin, Y. S.; Vermeulen, R.; Tsai, C. H.; Waidyanatha, S.; Lan, Q.; Rothman, N.; Smith, M. T.; Zhang, L.; Shen, M.; Li, G.; et al. Albumin Adducts of Electrophilic Benzene Metabolites in Benzene-Exposed and Control Workers. *Environ. Health Perspect.* **2007**, *115* (1), 28–34.
(43) Grigoryan, H.; Edmands, W.; Lu, S. S.; Yano, Y.; Regazzoni, L.; Iavarone, A. T.; Williams, E. R.; Rappaport, S. M. Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin. *Anal. Chem.* **2016**, *88* (21), 10504–10512.
(44) Beck, J. L.; Ambahera, S.; Yong, S. R.; Sheil, M. M.; Jersey, J. De; Ralph, S. F. Direct Observation of Covalent Adducts with Cys34 of Human Serum Albumin Using Mass Spectrometry. *Anal. Biochem.* **2004**, *325* (2), 326–336.
(45) Carballal, S.; Alvarez, B.; Turell, L.; Botti, H.; Freeman, B. A.; Radi, R. Sulfenic Acid in Human Serum Albumin. *Amino Acids* **2007**, *32* (4), 543–551.
(46) Aldini, G.; Regazzoni, L.; Orioli, M.; Rimoldi, I.; Facino, R. M.; Carini, M. A Tandem MS Precursor-Ion Scan Approach to Identify Variable Covalent Modification of Albumin Cys34: A New Tool for Studying Vascular Carbonylation. *J. Mass Spectrom.* **2008**, *43* (11), 1470–1481.
(47) Stewart, A. J.; Blindauer, C. A.; Berezenko, S.; Sleep, D.; Tooth, D.; Sadler, P. J. Role of Tyr84 in Controlling the Reactivity of Cys34 of Human Albumin. *FEBS J.* **2005**, *272* (2), 353–362.
(48) Rappaport, S. M.; Li, H.; Grigoryan, H.; Funk, W. E.; Williams, E. R. Adductomics: Characterizing Exposures to Reactive Electrophiles. *Toxicol. Lett.* **2012**, *213* (1), 83–90.

(49) Colombo, G.; Clerici, M.; Giustarini, D.; Rossi, R.; Milzani, A.; Dalle-Donne, I. Redox Albuminomics: Oxidized Albumin in Human Diseases. *Antioxid. Redox Signal.* **2012**, *17* (11), 1515–1527.
(50) Giustarini, D.; Dalle-Donne, I.; Tsikas, D.; Rossi, R. Oxidative Stress and Human Diseases: Origin, Link, Measurement, Mechanisms, and Biomarkers. *Crit. Rev. Clin. Lab. Sci.* **2009**, *46* (5–6), 241–281.
(51) Carballal, S.; Radi, R.; Kirk, M. C.; Barnes, S.; Freeman, B. A.; Alvarez, B. Sulfenic Acid Formation in Human Serum Albumin by Hydrogen Peroxide and Peroxynitrite †. *Biochemistry* **2003**, *42* (33), 9906–9914.
(52) Turell, L.; Botti, H.; Carballal, S.; Radi, R.; Alvarez, B. Sulfenic Acid-A Key Intermediate in Albumin Thiol Oxidation. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2009**, *877* (28), 3384–3392.
(53) Giustarini, D.; Dalle-Donne, I.; Lorenzini, S.; Milzani, A.; Rossi, R. Age-Related Influence on Thiol, Disulfide, and Protein-Mixed Disulfide Levels in Human Plasma. *J. Gerontol. A. Biol. Sci. Med. Sci.* **2006**, *61* (10), 1030–1038.
(54) Bar-Or, D.; Heyborne, K. D.; Bar-Or, R.; Rael, L. T.; Winkler, J. V.; Navot, D. Cysteinylation of Maternal Plasma Albumin and Its Association with Intrauterine Growth Restriction. *Prenat. Diagn.* **2005**, *25* (3), 245–249.
(55) Nagumo, K.; Tanaka, M.; Chuang, V. T. G.; Setoyama, H.; Watanabe, H.; Yamada, N.; Kubota, K.; Tanaka, M.; Matsushita, K.; Yoshida, A.; et al. Cys34-Cysteinylated Human Serum Albumin Is a Sensitive Plasma Marker in Oxidative Stress-Related Chronic Diseases. *PLoS One* **2014**, *9* (1).
(56) Greilberger, J.; Koidl, C.; Greilberger, M.; Lamprecht, M.; Schroecksnadel, K.; Leblhuber, F.; Fuchs, D.; Oettl, K. Malondialdehyde, Carbonyl Proteins and Albumin-Disulphide as Useful Oxidative Markers in Mild Cognitive Impairment and Alzheimer's Disease. *Free Radic. Res.* **2008**, *42* (7), 633–638.
(57) Lu, S. S.; Grigoryan, H.; Edmands, W. M. B.; Hu, W.; Iavarone, A. T.; Hubbard, A.; Rothman, N.; Vermeulen, R.; Lan, Q.; Rappaport, S. M. Profiling the Serum Albumin Cys34 Adductome of Solid Fuel Users in Xuanwei and Fuyuan, China. *Environ. Sci. Technol.* **2017**, *51* (1), 46–57.
(58) Liu, S.; Grigoryan, H.; Edmands, W. M. B.; Dagnino, S.; Sinharay, R.; Cullinan, P.; Collins, P.; Chung, K. F.; Barratt, B.; Kelly, F. J.; et al. Cys34 Adductomes Differ between Patients with Chronic Lung or Heart Disease and Healthy Controls in Central London. *Environ. Sci. Technol.* **2018**, *52* (4), 2307–2313.
(59) Grigoryan, H.; Edmands, W. M. B.; Lan, Q.; Carlsson, H.; Vermeulen, R.; Zhang, L.; Yin, S.-N.; Li, G.-L.; Smith, M. T.; Rothman, N.; et al. Adductomic Signatures of Benzene Exposure Provide Insights into Cancer Induction. *Carcinogenesis* **2018**, *39* (5), 661–668.
(60) Therrell, B. L.; Padilla, C. D.; Loeber, J. G.; Kneisser, I.; Saadallah, A.; Borrajo, G. J. C.; Adams, J. Current Status of Newborn Screening Worldwide: 2015. *Semin. Perinatol.* **2015**, *39* (3), 171–187.
(61) California Department of Public Health (CDPH). Background and History of the California Biobank Program (CBP) https://www.cdph.ca.gov/programs/GDSP/Pages/MoreAboutTheCBP.aspx (accessed Feb 22, 2017).
(62) Therrell, B. L.; Hannon, W. H. Newborn Dried Blood Spot Screening: Residual Specimen Storage Issues. *Pediatrics* **2012**, *129* (2), 365–366.

(63) Gale, K. B.; Ford, A. M.; Repp, R.; Borkhardt, A.; Keller, C.; Eden, O. B.; Greaves, M. F. Backtracking Leukemia to Birth: Identification of Clonotypic Gene Fusion Sequences in Neonatal Blood Spots. *Proc. Natl. Acad. Sci.* **1997**, *94* (25), 13950–13954.
(64) Greaves, M. In Utero Origins of Childhood Leukaemia. *Early Hum. Dev.* **2005**, *81* (1), 123–129.
(65) Wiemels, J. L.; Walsh, K. M.; de Smith, A. J.; Metayer, C.; Gonseth, S.; Hansen, H. M.; Francis, S. S.; Ojha, J.; Smirnov, I.; Barcellos, L.; et al. GWAS in Childhood Acute Lymphoblastic Leukemia Reveals Novel Genetic Associations at Chromosomes 17q12 and 8q24.21. *Nat. Commun.* **2018**, *9* (1), 286.
(66) Gluckman, P. D.; Hanson, M. A.; Cooper, C.; Thornburg, K. L. Effect of In Utero and Early-Life Conditions on Adult Health and Disease. *N. Engl. J. Med.* **2008**, *359* (1), 61–73.
(67) Barouki, R.; Gluckman, P. D.; Grandjean, P.; Hanson, M.; Heindel, J. J. Developmental Origins of Non-Communicable Disease: Implications for Research and Public Health. *Environ. Heal.* **2012**, *11* (1), 42.
(68) Mei, J. V.; Alexander, J. R.; Adam, B. W.; Hannon, W. H. Use of Filter Paper for the Collection and Analysis of Human Whole Blood Specimens. *J. Nutr.* **2001**, *131*, 1631–1636.
(69) Wagner, M.; Tonoli, D.; Varesio, E.; Hopfgartner, G. The Use of Mass Spectrometry to Analyze Dried Blood Spots. *Mass Spectrom. Rev.* **2016**, *35* (3), 361–438.
(70) Hall, E.; Flores, S.; De Jesús, V. Influence of Hematocrit and Total-Spot Volume on Performance Characteristics of Dried Blood Spots for Newborn Screening. *Int. J. Neonatal Screen.* **2015**, *1* (2), 69–78.
(71) Petrick, L.; Edmands, W.; Schiffman, C.; Grigoryan, H.; Perttula, K.; Yano, Y.; Dudoit, S.; Whitehead, T.; Metayer, C.; Rappaport, S. An Untargeted Metabolomics Method for Archived Newborn Dried Blood Spots in Epidemiologic Studies. *Metabolomics* **2017**, *13* (3), 27.
(72) O'Broin, S. D.; Kelleher, B. P.; Gunter, E. Evaluation of Factors Influencing Precision in the Analysis of Samples Taken from Blood Spots on Filter Paper. *Clin. Lab. Haematol.* **1995**, *17* (2), 185–188.
(73) Chambers, A. G.; Percy, A. J.; Hardie, D. B.; Borchers, C. H. Comparison of Proteins in Whole Blood and Dried Blood Spot Samples by LC/MS/MS. *J. Am. Soc. Mass Spectrom.* **2013**, *24* (9), 1338–1345.
(74) Martin, N. J.; Cooper, H. J. Challenges and Opportunities in Mass Spectrometric Analysis of Proteins from Dried Blood Spots. *Expert Rev. Proteomics* **2014**, *11* (6), 685–695.
(75) Chambers, A. G.; Percy, A. J.; Yang, J.; Borchers, C. H. Multiple Reaction Monitoring Enables Precise Quantification of 97 Proteins in Dried Blood Spots. *Mol. Cell. Proteomics* **2015**, *14* (11), 3094–3104.
(76) Hustoft, H. K.; Reubsaet, L.; Greibrokk, T.; Lundanes, E.; Malerod, H. Critical Assessment of Accelerating Trypsination Methods. *J. Pharm. Biomed. Anal.* **2011**, *56* (5), 1069–1078.
(77) Cole, M.; Risso, D. Scone: Single Cell Overview of Normalized Expression Data. *R Packag. version 1.2.0* **2018**.
(78) Schiffman, C.; Petrick, L.; Perttula, K.; Yano, Y.; Carlsson, H.; Whitehead, T.; Metayer, C.; Hayes, J.; Edmands, W. M. B.; Rappaport, S.; et al. Data-Adaptive Pipeline for Filtering and Normalizing Metabolomics Data. **2018**, 1–12.

(79) Whitehead, T. P.; Metayer, C.; Wiemels, J. L.; Singer, A. W.; Miller, M. D. Childhood Leukemia and Primary Prevention. *Curr. Probl. Pediatr. Adolesc. Health Care* **2016**, *46* (10), 317–352.
(80) Bomken, S. N.; Josef Vormoor, H. Childhood Leukaemia. *Paediatr. Child Health (Oxford).* **2009**, *19* (8), 345–350.
(81) Ward, E.; DeSantis, C.; Robbins, A.; Kohler, B.; Jemal, A. Childhood and Adolescent Cancer Statistics, 2014. *CA. Cancer J. Clin.* **2014**, *64* (2), 83–103.
(82) Madhusoodhan, P. P.; Carroll, W. L.; Bhatla, T. Progress and Prospects in Pediatric Leukemia. *Curr. Probl. Pediatr. Adolesc. Health Care* **2016**, *46* (7), 229–241.
(83) Metayer, C.; Dahl, G.; Wiemels, J.; Miller, M. Childhood Leukemia: A Preventable Disease. *Pediatrics* **2016**, *138* (Supplement), S45–S55.
(84) Erdmann, F.; Kielkowski, D.; Schonfeld, S. J.; Kellett, P.; Stanulla, M.; Dickens, C.; Kaatsch, P.; Singh, E.; Schüz, J. Childhood Cancer Incidence Patterns by Race, Sex and Age for 2000-2006: A Report from the South African National Cancer Registry. *Int. J. Cancer* **2015**, *136* (11), 2628–2639.
(85) Wiemels, J. Perspectives on the Causes of Childhood Leukemia. *Chem. Biol. Interact.* **2012**, *196* (3), 59–67.
(86) Zhang, J.; Walsh, M. F.; Wu, G.; Edmonson, M. N.; Gruber, T. A.; Easton, J.; Hedges, D.; Ma, X.; Zhou, X.; Yergeau, D. A.; et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N. Engl. J. Med.* **2015**, *373* (24), 2336–2346.
(87) Vijayakrishnan, J.; Houlston, R. S. Candidate Gene Association Studies and Risk of Childhood Acute Lymphoblastic Leukemia: A Systematic Review and Meta-Analysis. *Haematologica* **2010**, *95* (8), 1405–1414.
(88) Moriyama, T.; Relling, M. V; Yang, J. J. Inherited Genetic Variation in Childhood Acute Lymphoblastic Leukemia. *Blood* **2015**, *125* (26), 3988–3995.
(89) Chokkalingam, A. P.; Buffler, P. A. Genetic Susceptibility to Childhood Leukaemia. *Radiat Prot Dosim.* **2008**, *132* (2), 119–129.
(90) Greaves, M. Infection, Immune Responses and the Aetiology of Childhood Leukaemia. *Nat. Rev. Cancer* **2006**, *6* (3), 193–203.
(91) Buffler, P.; Kwan, M.; Reynolds, P.; Urayama, K. Environmental and Genetic Risk Factors for Childhood Leukemia: Appraising the Evidence. *Cancer Invest.* **2005**, *23* (1), 60–75.
(92) Eden, T. Aetiology of Childhood Leukaemia. *Cancer Treat. Rev.* **2010**, *36* (4), 286–297.
(93) Lightfoot, T. J.; Roman, E. Causes of Childhood Leukaemia and Lymphoma. *Toxicol. Appl. Pharmacol.* **2004**, *199*, 104–117.
(94) Belson, M.; Kingsley, B.; Holmes, A. Risk Factors for Acute Leukemia in Children: A Review. *Environ. Health Perspect.* **2007**, *115* (1), 138–145.
(95) Singer, A. W.; Carmichael, S. L.; Selvin, S.; Fu, C.; Block, G.; Metayer, C. Maternal Diet Quality before Pregnancy and Risk of Childhood Leukaemia. *Br. J. Nutr.* **2016**, No. 21, 1–10.
(96) Metayer, C.; Milne, E.; Dockerty, J. D.; Clavel, J.; Pombo-De-Oliveira, M. S.; Wesseling, C.; Spector, L. G.; Schüz, J.; Petridon, E.; Ezzat, S.; et al. Maternal Supplementation with Folic Acid and Other Vitamins and Risk of Leukemia in Offspring: A Childhood Leukemia Internotionol Consortium Study. *Epidemiology* **2014**, *25* (6), 811–822.
(97) Puumala, S. E.; Ross, J. A.; Aplenc, R.; Spector, L. G. Epidemiology of Childhood Acute Myeloid Leukemia. *Pediatr. Blood Cancer* **2013**, *60* (5), 728–733.

(98) Amitay, E. L.; Keinan-Boker, L. Breastfeeding and Childhood Leukemia Incidence. *JAMA Pediatr.* **2015**, *169* (6), e151025.
(99) Marcotte, E. L.; Thomopoulos, T. P.; Infante-Rivard, C.; Clavel, J.; Petridou, E. T.; Schüz, J.; Ezzat, S.; Dockerty, J. D.; Metayer, C.; Magnani, C.; et al. Caesarean Delivery and Risk of Childhood Leukaemia: A Pooled Analysis from the Childhood Leukemia International Consortium (CLIC). *Lancet Haematol.* **2016**, *3* (4), e176–e185.
(100) Greaves, M. Childhood Leukemia. *Bmj* **2002**, *324*, 283–287.
(101) Teitell, M. A.; Pandolfi, P. P. Molecular Genetics of Acute Lymphoblastic Leukemia. *Annu. Rev. Pathol. Mech. Dis.* **2009**, *4* (1), 175–198.
(102) Pui, C.-H.; Relling, M. V.; Downing, J. R. Acute Lymphoblastic Leukemia. *N. Engl. J. Med.* **2004**, *350*, 1535–1548.
(103) Ferrando, A. A.; Neuberg, D. S.; Staunton, J.; Loh, M. L.; Huard, C.; Raimondi, S. C.; Behm, F. G.; Pui, C. H.; Downing, J. R.; Gilliland, D. G.; et al. Gene Expression Signatures Define Novel Oncogenic Pathways in T Cell Acute Lymphoblastic Leukemia. *Cancer Cell* **2002**, *1* (1), 75–87.
(104) Chiaretti, S.; Foà, R. T-Cell Acute Lymphoblastic Leukemia. *Haematologica* **2009**, *94* (2), 160–162.
(105) Pui, C.-H.; Carroll, W. L.; Meshinchi, S.; Arceci, R. J. Biology, Risk Stratification, and Therapy of Pediatric Acute Leukemias: An Update. *J. Clin. Oncol.* **2011**, *29* (5), 551–565.
(106) Szczepański, T.; Harrison, C. J.; van Dongen, J. J. Genetic Aberrations in Paediatric Acute Leukaemias and Implications for Management of Patients. *Lancet Oncol.* **2010**, *11* (9), 880–889.
(107) Arceci, R. J.; Meshinchi, S. Childhood Leukemia. In *Pediatrics in review / American Academy of Pediatrics*; Reaman, G. H., Smith, F. O., Eds.; Pediatric Oncology; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; Vol. 31.
(108) Arber, D. A.; Orazi, A.; Hasserjian, R.; Thiele, J.; Borowitz, M. J.; Le Beau, M. M.; Bloomfield, C. D.; Cazzola, M.; Vardiman, J. W. The 2016 Revision to the World Health Organization Classification of Myeloid Neoplasms and Acute Leukemia. *Blood* **2016**, *127* (20), 2391–2405.
(109) Rose, D.; Haferlach, T.; Schnittger, S.; Perglerová, K.; Kern, W.; Haferlach, C. Subtype-Specific Patterns of Molecular Mutations in Acute Myeloid Leukemia. *Leukemia* **2017**, *31* (1), 11–17.
(110) Papaemmanuil, E.; Gerstung, M.; Bullinger, L.; Gaidzik, V. I.; Paschka, P.; Roberts, N. D.; Potter, N. E.; Heuser, M.; Thol, F.; Bolli, N.; et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **2016**, *374* (23), 2209–2221.
(111) Bolouri, H.; Farrar, J. E.; Triche, T.; Ries, R. E.; Lim, E. L.; Alonzo, T. A.; Ma, Y.; Moore, R.; Mungall, A. J.; Marra, M. A.; et al. The Molecular Landscape of Pediatric Acute Myeloid Leukemia Reveals Recurrent Structural Alterations and Age-Specific Mutational Interactions. *Nat. Med.* **2018**, *24* (1), 103–112.
(112) Biondi, A.; Cimino, G.; Pieters, R.; Pui, C.-H. Biological and Therapeutic Aspects of Infant Leukemia. *Blood* **2000**, *96* (1), 24–33.
(113) Wiemels, J. L.; Ford, A. M.; Van Wering, E. R.; Postma, A.; Greaves, M. Protracted and Variable Latency of Acute Lymphoblastic Leukemia after TEL-AML1 Gene Fusion in Utero. *Blood* **1999**, *94* (3), 1057–1062.
(114) Greaves, M. F.; Maia, A. T.; Wiemels, J. L.; Ford, A. M. Leukemia in Twins: Lessons in Natural History. *Blood* **2003**, *102* (7), 2321–2333.

(115) Greaves, M. F.; Wiemels, J. Origins of Chromosome Translocations in Childhood Leukaemia. *Nat. Rev. Cancer* **2003**, *3* (9), 639–649.
(116) Mori, H.; Colman, S. M.; Xiao, Z.; Ford, A. M.; Healy, L. E.; Donaldson, C.; Hows, J. M.; Navarrete, C.; Greaves, M. Chromosome Translocations and Covert Leukemic Clones Are Generated during Normal Fetal Development. *Proc. Natl. Acad. Sci.* **2002**, *99* (12), 8242–8247.
(117) Barr, D. B.; Bishop, A.; Needham, L. L. Concentrations of Xenobiotic Chemicals in the Maternal-Fetal Unit. *Reprod. Toxicol.* **2007**, *23* (3), 260–266.
(118) Perera, F.; Herbstman, J. Prenatal Environmental Exposures, Epigenetics, and Disease. *Reprod. Toxicol.* **2011**, *31* (3), 363–373.
(119) Metayer, C.; Zhang, L.; Wiemels, J. L.; Bartley, K.; Schiffman, J.; Ma, X.; Aldrich, M. C.; Chang, J. S.; Selvin, S.; Fu, C. H.; et al. Tobacco Smoke Exposure and the Risk of Childhood Acute Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer Epidemiol. Biomarkers Prev.* **2013**, *22* (9), 1600–1611.

# Chapter 2.

# Analysis of Hemoglobin to Adjust for Blood Volumes in Newborn Dried Blood Spots

Yukiko Yano,[1] Lauren Petrick,[1] Todd Whitehead,[2,3] Catherine Metayer,[2,3] and Stephen Rappaport[1,3]

1. Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA 94720 USA
2. Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720 USA
3. Center for Integrative Research on Childhood Leukemia and the Environment, University of California, Berkeley, CA 94720 USA

## 2.1 Abstract

Archived newborn dried blood spots (ANBS) provide valuable opportunities to investigate the impact of prenatal and early life exposures on disease risks. Recently, there has been an increasing interest in performing metabolomics and proteomics with ANBS to discover novel biomarkers associated with causes and effects of pediatric diseases. One challenge is the variation in blood volumes across ANBS punches that is introduced by differences in blood hematocrit (Hct) in neonates. This unknown variability in Hct reduces the statistical power to detect differences between groups (such as diseased vs. non-diseased) in epidemiological studies. Previously, potassium ($K^+$) concentrations of ANBS extracts were measured as a surrogate for Hct using a micro $K^+$ ion-selective electrode to normalize for differences in blood volume. However, this method was time-consuming and could not easily be integrated into high-throughput omics analyses. Here, we describe an alternative normalization method using UV-Vis spectrophotometry to rapidly quantify hemoglobin (Hb) in a few microliters of ANBS extracts. We found that Hb and $K^+$ concentrations were highly correlated in ANBS from a case-control study of childhood leukemia (n ~ 860). We demonstrate that Hb concentrations of ANBS extracts can be used as proxies for $K^+$ (and Hct) to effectively remove blood volume as a variable in omics analyses.

## 2.2 Introduction

The analysis of archived newborn blood spots (ANBS) offers new avenues for epidemiological investigations. ANBS have been routinely collected by heel prick within 24 – 48 h of birth as part of newborn screening programs in the U.S. and worldwide.[1] There are currently 14 programs in the U.S., serving approximately half of the newborn population, that archive residual blood spots from neonates for more than 21 years.[2] The state of California has retained these specimen in freezer storage since 1982.[3] In the past, retrospective analyses of ANBS were performed to provide direct evidence for the prenatal origin of childhood leukemia,[4,5] thus

demonstrating the utility of ANBS in making important discoveries about disease etiology. With recent advances in high-throughput omics (i.e., genomics, metabolomics, proteomics), there is growing interest in using ANBS for population-based research to discover novel markers of disease processes.[6–8]

Of the many analytical challenges that limit use of ANBS for studies of disease etiology, one critical factor is the variability in sample volume that results from the heterogeneity in hematocrit (Hct) values across neonates.[9–11] Since Hct is a determinant of blood viscosity, it affects the spreading of a blood drop on the filter paper (Guthrie card) and thus leads to non-uniform blood volumes within and between ANBS that affect the accuracy and precision of assays.[7,9,12] While many laboratories rely on the assumption that equally-sized punches removed from ANBS contain identical volumes of blood,[11] Hct values vary by sex, gestational age, and birth weight.[10,13–15] Therefore, accounting for the variation in blood volumes in ANBS can increase the statistical power to detect small differences between diseased and healthy populations.[7,12]

Capiau et al. showed that potassium ($K^+$) concentrations were highly correlated with Hct in extracts from dried blood spots (DBS),[16] and that normalization for $K^+$ effectively removed biases in the quantitation of targeted analytes.[17] However, analysis of $K^+$ with a clinical chemistry analyzer resulted in loss of the sample and thus precluded analyses of other analytes in the same DBS extract. To circumvent this shortcoming, we used a micro ion-selective electrode to measure $K^+$ concentrations in extracts from ANBS punches and then performed metabolomics analysis with the same extracts.[7] Normalization for $K^+$ removed nuisance variation and allowed us to detect biologically relevant differences in small-molecule features across ANBS in our samples. However, using the $K^+$ micro-electrode was time-consuming and measurement drift was observed over time due to protein build-up.

While the hemoglobin (Hb) content of DBS has also been suggested as a surrogate for Hct,[12,18–21] these Hb-based normalization techniques require expensive equipment and/or do not meet the high-throughput needs of omics studies. Here, we describe an alternative technique, which employs UV-Vis spectrophotometry to rapidly quantify Hb concentrations in microliter volumes of extracts from ANBS punches. The method was developed with blood from adult volunteers and validated with ANBS. We found that $K^+$ and Hb concentrations were highly correlated in ANBS extracts, and that this method permits efficient adjustment for variation in blood volume across ANBS.

## 2.3 Materials and methods

### *Chemicals and Reagents*

Water was prepared with a PureLab purification system (18.2 mΩ cm resistivity at 25 °C; Elga LabWater, Woodridge, IL). Potassium chloride (>99%) and sodium chloride (>99%) were purchased from Sigma Aldrich (St. Louis, MO) and Fisher Scientific (Pittsburgh, PA), respectively. Purified human Hb was purchased from MP Biomedicals, LLC (Santa Ana, CA).

### *Preparation of Experimental DBS for Method Development*

Venous blood was collected in sodium heparin tubes (BD Vacutainer, BD) with informed consent from two adult female volunteers. Experimental DBS were prepared the same day of blood collection by aliquoting 50 $\mu$L of whole blood on Whatman 903 Guthrie cards (GE Healthcare), which were air dried for a minimum of 4 days and stored at -20 ℃ in glassine

envelopes (GE Healthcare) prior to use (2 months – 1 year). To produce experimental DBS with various Hct levels for the same subject, venous blood was centrifuged at 1200 g and diluted or concentrated by adding or removing plasma, as described by Capiau et al.[16] Blood with very low, low, medium (unaltered), and high Hct were prepared and aliquoted onto DBS cards, dried, and stored prior to use.

***ANBS from the CCLS***

We received single 4.7-mm ANBS punches (~8 $\mu$L whole blood) for participants in the California Childhood Leukemia Study (CCLS, described elsewhere).[22] Briefly, the CCLS is a population-based case-control study of incident childhood leukemia conducted in 35 counties in California from 1995 to 2008. Cases were ascertained within 72 hours of diagnosis at pediatric hospitals, and controls were selected using birth certificates. ANBS of the CCLS participants in the present analysis were obtained from the California Department of Public Health, where the ANBS had been archived at -20 ℃. In addition, information of the child's sex and birth characteristics (i.e., gestational age and birth weight) were collected from interviews with the biological parents (mainly the mother).

***Method Development***

Four types of experiments were performed for method development as shown in Figure 2.1. For the first experiment (MD1), experimental DBS of various punch sizes were extracted with water to test the relationship between $K^+$ and Hb concentrations over a wide dynamic range. The experiment was performed in three batches, each with a different punch size range: 2.0 – 6.0-mm (batch 1), 6.0 – 7.8-mm (batch 2), and 2.0 – 7.8-mm (batch 3). In the second experiment (MD2), experimental DBS punches were extracted with different volumes of water to test the effect of the extraction volume on the $K^+$ and Hb concentrations. For the third experiment (MD3), the effect of Hct on the relationship between $K^+$ and Hb concentrations was investigated by extracting experimental DBS containing different Hct levels (described above). To account for differences between adult blood in experimental DBS and newborn blood, ANBS punches from 10 randomly selected control children from the CCLS were analyzed in the fourth experiment (MD4).

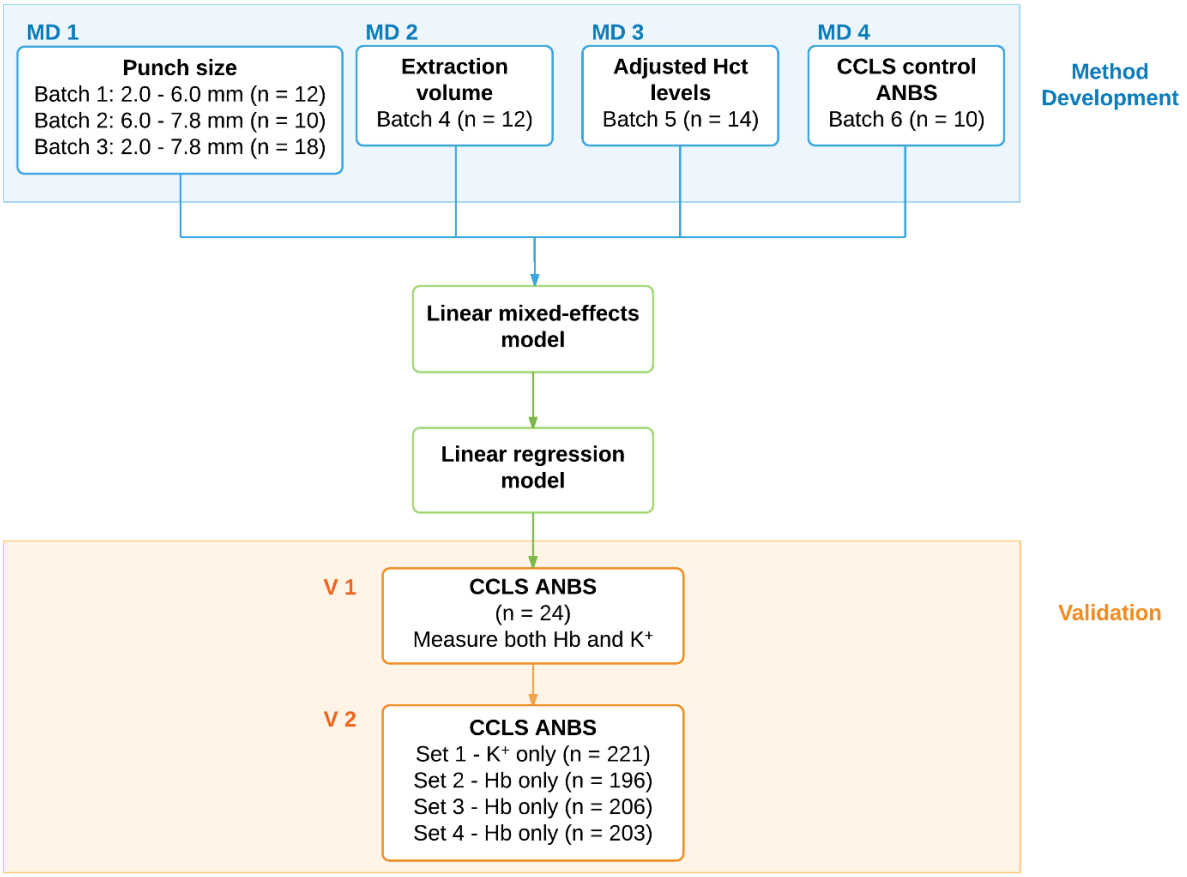



**Figure 2. 1** Flow chart of the experimental method: Four experiments were performed for method development (MD1 – 4). Linear mixed-effects models were applied to perform a combined analysis of all method development experiments. The resulting adjusted Hb and K+ values obtained from the mixed models were used to build a linear regression model to predict $\mathbf{K^+}$ concentrations from Hb concentrations in DBS extracts. This linear model was validated using ANBS from the CCLS (V1 and V2). Diagram made in Lucidchart (Lucid Software, Inc, South Jordan, Utah, USA).

### *Extraction of $K^+$ and Hb from blood spots*

In MD1, extracts were prepared in duplicate from DBS punches of 2.0, 3.0, 4.0, 5.0 and 6.0-mm diameters from a single donor. Extracts representing larger punches, i.e., 6.3, 6.7, 7.2, and 7.8-mm, were prepared in duplicate by combining a 6-mm punch with a smaller punch (e.g., a 6.3-mm punch represents the combined areas of a 6-mm and a 2-mm punch). The punches were placed in microcentrifuge tubes and extracted with 100 $\mu$L of water at room temperature for 15 min with constant agitation at 1400 rpm (Eppendorf Thermomixer). Samples were then centrifuged for 10 s and a 5 $\mu$L aliquot was transferred to a new tube and diluted with 20 $\mu$L of water for Hb measurements while $K^+$ concentrations were measured in the remaining extracts. For MD2, the effect of the extraction volume was investigated with duplicate 3-mm and 5-mm punches from a single donor. The 3-mm punches were extracted in 40, 50, 80, and 100 $\mu$L of water, and the 5-mm punches were extracted in 80 and 100 $\mu$L of water. The punches were analyzed for Hb and $K^+$ as described above for MD1.

In MD3, the relationship between $K^+$ and Hb was examined in relation to Hct by using duplicate 5-mm punches from the experimental DBS prepared with blood from the two donors. For the first donor, 5-mm punches were prepared from four Hct levels: very low, low, medium, and high, and for the second donor, 5-mm punches were prepared from three Hct levels: low, medium, and high. Each punch was extracted with 100 $\mu$L of water for $K^+$ and Hb measurements as described above.

The 10 ANBS from the CCLS in MD4 were extracted in 100 $\mu$L of water and prepared for $K^+$ and Hb measurements as described above. In addition to accounting for differences between newborn blood and adult blood, these ANBS were included to assess the effect of long-term freezer storage (~20 years) on $K^+$ and Hb concentrations in the extracts. Two 5- $\mu$L spots (each equivalent to a 3.8-mm punch[23]) from the experimental DBS, which were prepared by volumetrically applying blood from the adult volunteers onto the filter paper, were analyzed along with the ANBS in this experiment as reference points to ensure comparability with the method-development experiments that used experimental DBS.

***Validation Experiments with ANBS***

The relationship between $K^+$ and Hb concentrations was validated in two experiments (Figure 2.1). In the first validation experiment (V1 in Figure 2.1), both $K^+$ and Hb concentrations were measured in extracts from 4.7-mm punches of 24 randomly selected ANBS from the CCLS cases (n = 11) and controls (n = 13). This experiment was conducted to confirm that the same $K^+$ and Hb relationship seen in method-development experiments would be observed with ANBS extracts. In addition, we also assessed whether Hb concentrations could be used to predict $K^+$ concentrations by comparing the predicted values with the observed $K^+$ concentrations measured with the micro-electrode.

To further evaluate the utility of measuring Hb as a proxy for $K^+$ concentrations in ANBS for large scale studies, the distribution of $K^+$ concentrations obtained from an ion micro-electrode was compared to that predicted from Hb measurements in ANBS in the second validation experiment (V2). Four sets of independent samples of ANBS from the CCLS (cases and controls) were used. In the first set, $K^+$ measurements had been performed with a micro-electrode for 221 ANBS as described by Petrick et al.[7] In the second, third, and fourth sets, $K^+$ concentrations were predicted from Hb measurements in 196, 206, and 203 subjects respectively. We tested to see if the distributions of the estimated $K^+$ concentrations in sets 2, 3, and 4 were comparable with the observed $K^+$ concentrations from the first set.

Samples from V2 were also used to evaluate whether the Hb normalization method could be easily integrated into a large-scale omics workflow. Since omics studies typically entail hundreds of samples, it is important that the method be simple and that instrumental analysis be rapid. It is also necessary for the instrument to be stable over time without any loss of accuracy or precision because batch sizes can be large and the analyses can span days to weeks for omics studies. Finally, the sample volume required for the normalization method should be minimal to ensure there is sufficient sample left to perform downstream omics analyses.

***Extraction of $K^+$ and Hb from ANBS in Validation Experiments***

Similar to the method-development samples, ANBS punches in V1 were extracted with 100 $\mu$L of water to measure Hb concentrations. Aliquots of 5 $\mu$L were then diluted with 20 $\mu$L of water for Hb measurements. The $K^+$ concentrations were later obtained using the remainder of the diluted extracts from Hb measurements, after storage at -20 ℃ for approximately 1 year. This

required further diluting the remaining 14 $\mu$L of the extracts with an additional 12 $\mu$L of water to obtain sufficient volume for measurement of $K^+$ concentrations. In V2, all ANBS punches in sets 2, 3, and 4 were extracted with 100 $\mu$L of water and diluted for Hb measurements as described above.

***Analysis of $K^+$***

Potassium concentrations of DBS extracts were measured using a micro $K^+$ ion-selective electrode (MI-442 and MI-401 1-mm tip, Microelectrodes Inc., Accumet AB250 meter, Fisher Scientific), as previously described.[7] To convert voltage readings to $K^+$ concentrations, a five-point semi-logarithmic calibration curve was generated with standard solutions of 0.001 to 0.01 N KCl containing 0.1 N NaCl as a potentially interfering species (as per manufacturer's instructions). The 0.05 N KCl standard solution was measured after every 12 measurements in order to monitor measurement drift over time. $K^+$ measurements were performed in duplicate for each DBS extract. Calibration curves were obtained for each batch in the method-development experiments and for V1.

***Analysis of Hb***

Hb concentrations were determined by UV-Vis absorption spectrophotometry at room temperature using a Cytation 5 microplate spectrophotometer with a Take3 micro-volume 16-well plate (BioTek Instruments, Winooski, VT). The absorbance of duplicate 2.5-$\mu$L sample aliquots was measured at 407 nm, which was the experimentally-determined absorbance maximum for Hb. Absorbance readings were converted into Hb concentrations with five-point linear calibration curves using Hb standard solutions ranging from 0.1 to 3.0 mg/mL for each batch in the method-development experiments and for V1. In V2, duplicate Hb measurements were made for each extract in batches of 48 samples, and a seven-point calibration curve was prepared for each batch (Hb concentration range 0.5 to 3.5 mg/mL).

***Statistical Analyses***

All statistical analyses were performed using the R statistical programming environment.[24] Pearson correlation between $K^+$ and Hb concentrations was initially determined separately in each of the method-development experiments, i.e., punch sizes (MD1), extraction volumes (MD2), varying Hct levels (MD3), and ANBS from the CCLS (MD4).

Data from all method-development experiments were then combined in a series of statistical models to evaluate the overall relationship between $K^+$ and Hb (Figure 2.1). The first step of this combined analysis was to adjust measured concentrations of $K^+$ and Hb for technical variation (i.e., batch effects) and biological variation (i.e., differences between adult DBS and ANBS) using a linear mixed-effects model applied with the lme4 R package.[25] The following equation was applied to both $K^+$ and Hb measurements separately after log-transformation:

$$Y_{\mathrm{ijkl}} = \ln\left(X_{\mathrm{ijk}}\right) = \beta_0 + \beta_1 X_{\mathrm{i}} + \beta_2 X_{\mathrm{S}} + \beta_3 X_{\mathrm{N}} + a_{\mathrm{j}} + b_{\mathrm{k(j)}} + e_{\mathrm{l(jk)}} \quad \text{(Eq. 1)}$$

where $Y_{\mathrm{ijkl}}$ is the natural logarithm of $X_{\mathrm{ijkl}}$, which represents the concentration of $K^+$ (mM) or Hb (mg/mL) for the *l*th duplicate measurement ($l$ = 1,2), of the *k*th biological replicate ($k$ = 1,2), from the *j*th sample ($j$ = 1,…,43), in the *i*th batch ($i$ = 1,…,6). Method-development experiments were performed in the following batches: MD1 (batches 1 – 3), MD2 (batch 4), MD3 (batch 5), and MD4 (batch 6). The intercept $\beta_0$ is the overall mean, and $\beta_1$, $\beta_2$, and $\beta_3$ represent the coefficients for the fixed effects of $X_{\mathrm{i}}$ (batch number), $X_{\mathrm{S}}$ (punch size [mm]), and $X_{\mathrm{N}}$ (categorical variable for newborn or adult, with newborn as the reference), respectively. The batch variable

$\beta_1$ accounts for nuisance variation in measurements associated with time. In Eq. 1, $a_{\mathrm{j}}$ and $b_{\mathrm{k(j)}}$ are the random effects for the *k*th biological replicate and the *j*th sample, respectively, and $e_{\mathrm{l(jk)}}$ is the random error for the *l*th duplicate measurement. Adjusted, log-transformed $K^+$ and Hb concentrations were predicted as $Y_{\mathrm{ijkl}} = \beta_0 + \beta_2 X_{\mathrm{S}} + a_{\mathrm{j}}$, where $\beta_2$ is the best linear unbiased estimator for the fixed effect for punch size and $a_{\mathrm{j}}$ is the best linear unbiased predictor of the *j*th sample random effect [26]. Variance components were estimated using the restricted maximum likelihood method. The precision of $K^+$ and Hb measurements were evaluated by their respective coefficients of variation (CV) estimated as $\sqrt{\mathrm{e}^{\sigma_{\mathrm{e}}^2} - 1}$, where $\sigma_{\mathrm{e}}^2$ is the error variance obtained from the log-scale mixed model (Eq. 1).

In the second step of combined analysis of the method-development experiments, the adjusted, log-transformed $K^+$ and Hb concentrations estimated by the mixed models (Eq. 1) were used as inputs for a linear regression model with $K^+$ concentrations as the dependent variable and Hb concentrations as the independent variable. This linear model was then used in the validation experiments (V1 and V2) to estimate $K^+$ concentrations in ANBS extracts based on Hb concentration measurements.

In V1, where both $K^+$ and Hb concentrations were measured in the same ANBS extracts from 24 subjects, Pearson correlation between the two concentrations was determined, and a linear model was fit with $K^+$ concentrations as the dependent variable and Hb concentrations as the independent variable. In linear regression analyses, model diagnostics including a plot of residuals against fitted values and a normal Q-Q plot of the residuals were used to identify deviations from the assumptions of linearity and normality. Influential points were identified with the Cook's distance statistic.

Analysis of variance (ANOVA) was used for V2 to test for differences between $K^+$ concentrations in the four sets of ANBS from the CCLS. The $K^+$ concentrations measured directly in ANBS extracts with the micro-electrode in the first set and $K^+$ concentrations predicted from Hb with the developed linear model in the remaining three sets were compared with a significance level of $\alpha = 0.05$. ANOVA was performed with and without adjusting for case status and covariates known to affect Hct levels (i.e., sex, gestational age and birth weight). Missing values in the covariate data were excluded from the analysis.

## 2.4 Results and discussion

***Method development***

All experiments using the experimental DBS showed a high correlation between $K^+$ and Hb concentrations ($r \geq 0.99$) (Figure S2.1) regardless of punch size (MD1), extraction volume (MD2), and Hct level (MD3). In addition, $K^+$ and Hb were also highly correlated in the ten ANBS in MD4 ($r = 0.85$).

To perform a combined analysis of the $K^+$ and Hb relationship using all data points from the method-development experiments, the linear mixed-effects model in Eq. 1 was used to adjust for technical variation and biological variation associated with age (adult vs. newborn) in the measurements of either $K^+$ or Hb. Although $K^+$ calibration curves showed good linearity in each batch, there were large inter-batch differences in the curves (Figure S2.2A), particularly for batches where the $K^+$ micro-electrode was not in optimal working condition. The resulting batch effect was clearly seen when data points were presented across all batches in a scatter plot (Figure S2.3). Therefore, the mixed model was used to remove this unwanted systematic variation in the $K^+$ measurements. On the other hand, Hb calibration curves showed excellent

linearity and inter-batch stability (Figure S2.2B). The estimated CVs of the $K^+$ and Hb measurements were 2.21 % and 1.02 %, respectively. Thus, while both methods showed good reproducibility between duplicate measurements, Hb measurements showed greater inter-batch stability over time.

Simple linear regression was used to estimate the association between adjusted $K^+$ and Hb concentrations. When $K^+$ concentrations were regressed on Hb concentrations in the natural log scale (Figure 2.2), the results showed a strong correlation ($r^2 = 0.98$). The regression coefficient was 1.04 (95 % confidence interval [CI]: 0.99, 1.09) and the intercept was 1.19 (95 % CI: 1.16, 1.22). Similar results were obtained with untransformed data with $r^2 = 0.97$ showing strong agreement between the two measurements (Figure S2.4). These results show that Hb concentrations of DBS extracts can be used as proxies for $K^+$ concentrations to adjust for Hct.

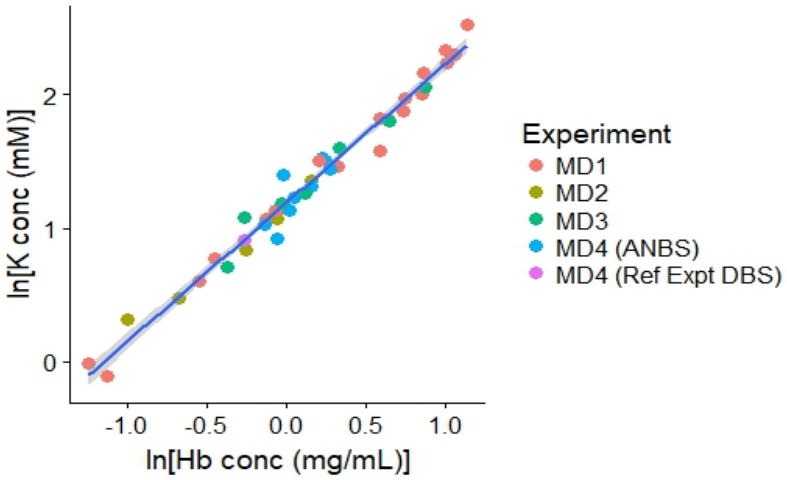



**Figure 2. 2** Linear relationship between $\mathbf{K^+}$ (mM) and Hb (mg/mL) concentrations in DBS extracts measured in the method-development experiments (in log scale). Colors represent the experiments that were performed for method development: size (MD1), extraction volume (MD2), adjusted Hct (MD3), ANBS (MD4), and experimental DBS analyzed with ANBS as reference points (Ref Expt DBS, MD4). The regression line (blue) and the pointwise 95% confidence band (shaded) are shown.

Using the adjusted Hb and $K^+$ values obtained from the mixed models in the linear regression model, the following equation was developed to predict $K^+$ concentrations of ANBS extracts from their observed Hb concentrations:

$$\ln(\mathrm{K^+[mM]}) = 1.04 \times \left(\ln(\mathrm{Hb[mg/mL]}) - \beta_{3[\mathrm{Hb}]}\right) + 1.19 + \beta_{3[\mathrm{K^+}]} \quad \text{(Eq. 2)}$$

where $\ln(\mathrm{K^+[mM]})$ and $\ln(\mathrm{Hb[mg/mL]})$ are the natural logarithm of the adjusted mM concentration of $K^+$and mg/mL concentration of Hb, respectively. The terms $\beta_{3[\mathrm{Hb}]}$ and $\beta_{3[\mathrm{K^+}]}$ in Eq. 2 are the fixed effect coefficients from Eq. 1 that adjusted for differences between newborns and adults, estimated at 0.46 and 0.35, respectively. Therefore, the equation for predicting newborn $K^+$ concentrations in ANBS from Hb concentrations is simplified to:

$$\ln(\mathrm{K^+[mM]}) = 1.04 \times \ln(\mathrm{Hb[mg/mL]}) + 1.06 \quad \text{(Eq. 3)}$$

Eq. 3 was used to estimate $K^+$concentrations to normalize for Hct in the analysis of ANBS in the validation experiments.

***Validation***

As described in the Methods section, the relationship between $K^+$ and Hb was validated with 24 ANBS from the CCLS in V1. While Hb concentrations of the ANBS extracts were measured as described, $K^+$ concentrations were obtained from the remaining extracts a year later and required further dilution to obtain the necessary volume for measurement of $K^+$. Despite differences in dilution and aging of approximately 1 year, the measured $K^+$ and Hb concentrations from these ANBS demonstrated good correlation ($r = 0.82$). After removal of one outlier identified from model diagnostics (Figure S2.5), the $K^+$ and Hb concentrations showed a linear relationship ($r^2 = 0.81$) with a slope of 2.04 (95 % CI: 1.60, 2.48) and intercept of 0.28 (95 % CI: -0.64, 1.20) (Figure 2.3). Using Eq. 3, $K^+$ concentrations were predicted from the measured Hb concentrations, and predicted $K^+$ concentrations were compared with $K^+$ concentrations determined with the micro-electrode. After removing the outlier, the linear relationship between the predicted and observed $K^+$ concentrations had a slope of 1.23 (95 % CI: 0.97, 1.50), intercept of 0.58 (95 % CI: -0.63, 1.80), and $r^2 = 0.81$ (Figure S2.6). Since the 95 % CI for the slope included one, and the 95 % CI for the intercept included zero, there was no evidence of fixed or proportional bias, respectively, between the predicted and observed $K^+$ concentrations.[27] However, predicted $K^+$ concentrations were slightly higher than observed $K^+$ concentrations (Figure S2.6). This difference is likely to have been caused by the different dilutions of the extracts in which Hb and $K^+$ measurements were made. Furthermore, this systematic error between predicted and observed $K^+$ concentrations does not alter the strong correlation between Hb and $K^+$ concentrations in ANBS extracts (Figure 2.3), and thus, does not diminish the utility of Hb measurements in normalizing for blood volume in the analysis of ANBS.

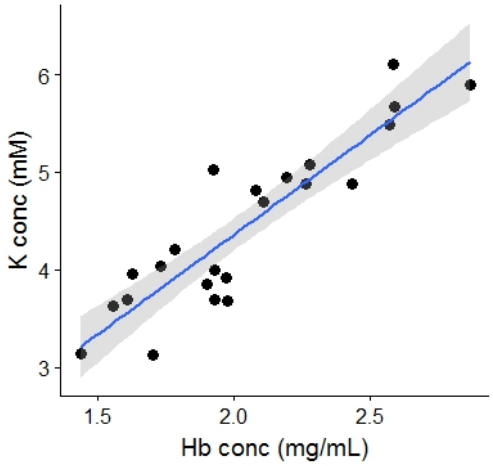



**Figure 2. 3** Plot of Hb (mg/mL) and **$K^+$** concentrations (mM) in extracts from 24 ANBS from CCLS analyzed in the validation experiment, V1. The linear regression line (blue) and the pointwise 95% confidence band (shaded) are shown.

We further validated the use of Hb measurements to estimate $K^+$ concentrations for 605 ANBS from the CCLS in V2. In our previous metabolomics analysis, $K^+$ concentrations were measured with the micro-electrode for 221 ANBS punches (Dataset 1).[7] For the present validation experiment, Hb concentrations were measured from 196, 206, and 203 ANBS punches

in three independent analyses (Datasets 2, 3, and 4, respectively) and $K^+$concentrations were estimated using Eq. 3. The mean Hb concentrations of Dataset 2, 3, and 4 were 1.93 $\pm$ 0.39, 1.95 $\pm$ 0.36, and 2.00 $\pm$ 0.39 mg/mL, respectively. We compared the observed $K^+$concentrations from Dataset 1 with the estimated $K^+$concentrations from Datasets 2, 3, and 4 to see if the four independent, random sample sets from the CCLS had comparable $K^+$distributions. As can be seen from the box plot (Figure 2.4) and summary statistics (Table 2.1), $K^+$concentrations from the four datasets were similar with no statistically significant differences (ANOVA *p*-value = 0.20; unadjusted for case status and covariates). Furthermore, predicted $K^+$concentrations in Datasets 2, 3, and 4 were not higher than the observed $K^+$concentrations in Dataset 1, suggesting that the difference between predicted and observed $K^+$concentrations in V1 were due to the different dilutions. Covariate characteristics of the four CCLS datasets in V2 are presented in Table 2.2. The four sets were comparable in terms of the number of cases and controls, sex, gestational age, and birth weight. There was no statistically significant difference in the $K^+$concentrations across the four sets after adjusting for case status, sex, gestational age, and birth weight (ANOVA *p*-value = 0.24). These validation results show that Hb concentrations can be used as proxies for $K^+$concentrations to normalize for Hct in ANBS analyses.

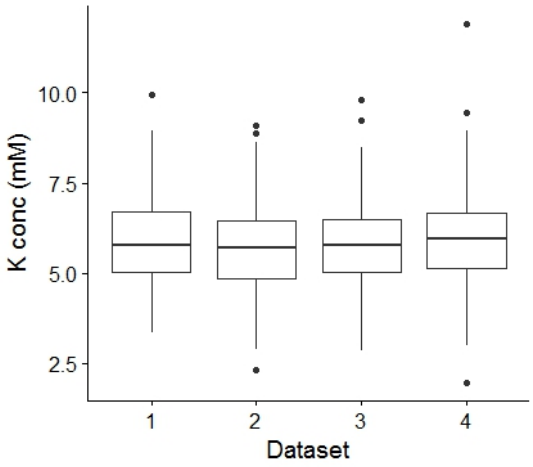



**Figure 2. 4** Box plot of $\mathbf{K^+}$ concentrations (mM) from the four CCLS datasets in the validation experiment, V2. In Dataset 1, $\mathbf{K^+}$ concentrations were directly measured in ANBS extracts with a micro-electrode. In Datasets 2, 3, and 4, $\mathbf{K^+}$ concentrations were predicted from measured Hb concentrations.

**Table 2. 1** Summary statistics of $\mathbf{K^+}$ concentrations (mM) from the four CCLS datasets in the validation experiment, V2.

| Dataset | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | n |
|---|---|---|---|---|---|---|---|
| 1 | 3.33 | 5.03 | 5.76 | 5.91 | 6.70 | 9.94 | 221 |
| 2 | 2.32 | 4.83 | 5.70 | 5.71 | 6.43 | 9.10 | 184 |
| 3 | 2.85 | 5.00 | 5.76 | 5.80 | 6.47 | 9.79 | 206 |
| 4 | 1.96 | 5.12 | 5.97 | 5.95 | 6.68 | 11.92 | 203 |

**Table 2. 2** Case status and covariate characteristics of the four CCLS datasets in the validation experiment, V2.

| | Dataset | | | |
|---|---|---|---|---|
| | **1** | **2** | **3** | **4** |
| Case status | | | | |
| Case | 106 | 88 | 104 | 103 |
| Control | 107 | 96 | 102 | 100 |
| Missing (%) | 8 (3.6) | 0 (0) | 0 (0) | 0 (0) |
| Sex | | | | |
| Female | 86 | 81 | 93 | 83 |
| Male | 127 | 103 | 113 | 120 |
| Missing (%) | 8 (3.6) | 0 (0) | 0 (0) | 0 (0) |
| Gestational age (weeks) | | | | |
| Mean (SD) | 39.2 (2.5) | 39.0 (2.4) | 39.4 (2.1) | 39.2 (2.4) |
| Missing (%) | 19 (8.6) | 26 (14.1) | 5 (2.4) | 10 (4.9) |
| Birth weight (g) | | | | |
| Mean (SD) | 3477 (567) | 3471 (635) | 3430 (537) | 3409 (604) |
| Missing (%) | 12 (5.4) | 12 (6.5) | 6 (2.9) | 5 (2.5) |

In Dataset 2 of V2, there were 12 ANBS punches for which Hb could not be extracted with water and thus Hb could not be quantified. While these samples were removed from the present analysis, we were able to measure $K^+$concentrations in these samples using the $K^+$ ion micro-electrode. It is likely that these ANBS samples were exposed to different storage or transport conditions since there is evidence that these factors influence quantitative analysis of Hb extracted from DBS.[12,19,28] We also observed decreased water solubility of Hb when DBS were stored at room temperature for several months compared to storage at -20 ℃ (data not shown). Hb was more susceptible to oxidation when DBS were kept at room temperature, which was demonstrated by color change from deep red to dark brown.[29] Thus, it appears that $K^+$ measurements are not as affected by these storage conditions. Since storage conditions of ANBS in the U.S. tend to vary by State,[30] it would be necessary to ensure that the ANBS used in future studies have been properly maintained at -20 ℃ (or colder) to enable precise measurement of Hb.

While the Hb normalization method is more sensitive to storage conditions, the Hb method is particularly attractive for omics studies for several reasons. First, measurement of Hb using the plate reader is very rapid (8 measurements/min) and fully automated; this is crucial for high-throughput analyses that require hundreds of samples. Hb measurements are highly reproducible, and minimal training is required to use the plate reader. In contrast, measuring $K^+$with the micro-electrode was performed manually and measurements tended to drift over time as sample deposits accumulated on the electrode. This limited the number of samples that could be analyzed in a batch and required the use of loess smoothing to adjust for the $K^+$ measurement drift.[7] The Hb method did not have any observed measurement drift even after ~200 samples were analyzed in V2, as confirmed by the reproducibility of the Hb calibration curve measured repeatedly throughout each set (data not shown). Moreover, the Hb method requires minimal sample volume per measurement (~2 $\mu$L), which is crucial for investigations of ANBS. Although absolute Hct values are not determined with this Hb method, the DBS Hb concentrations alone are sufficient for the purposes of adjusting for Hct in DBS, as was shown with the application of $K^+$-based normalization methods.[7,17]

## 2.5 Conclusion

Variation in Hct, which gives rise to variability in blood volume, has long hampered quantitative analyses of DBS[31] and epidemiologic applications of ANBS. We have developed and validated a simple and rapid method to measure Hb as a proxy for $K^+$, a recognized surrogate for Hct, in DBS extracts. Normalizing data for Hb should minimize measurement errors, thereby increasing statistical power to detect disease associations in studies employing ANBS. We are currently applying this Hb-based normalization method with data obtained from untargeted metabolomics to discover small molecules that discriminate between childhood leukemia cases and controls.[7]

## 2.6 Compliance with ethical standards

***Conflict of interest***
The authors declare they have no conflict of interest.

***Informed consent***
Written informed consent was obtained from all adult volunteer subjects providing blood samples. Written informed consent was obtained from the parents of all participating subjects in the CCLS.

***Ethics approval***
The CCLS and CIRCLE were approved by the University of California Committee for the Protection of Human Subjects, the California Health and Human Services Agency Committee for the Protection of Human Subjects, and the institutional review boards of all participating hospitals, as appropriate.

## 2.7 Acknowledgments

Research reported in this publication was conducted at the Center for Integrative Research on Childhood Leukemia and the Environment (CIRCLE), supported by the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH) Award Numbers P01/P50ES018172, and by the US Environmental Protection Agency (EPA) Award Numbers RD83451101 and RD8361590. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH and EPA. The biospecimens used in this study were obtained from the California Biobank Program, (SIS request number 26). The California Department of Public Health is not responsible for the results or conclusions drawn by the authors of this publication. We thank the families for their participation in the California Childhood Leukemia Study (CCLS) supported by the NIEHS Award Numbers R01ES009137 and P42ES004705.

## 2.8 References


(1) Therrell, B. L.; Padilla, C. D.; Loeber, J. G.; Kneisser, I.; Saadallah, A.; Borrajo, G. J. C.; Adams, J. Current Status of Newborn Screening Worldwide: 2015. *Semin. Perinatol.* **2015**, *39* (3), 171–187.

(2) Therrell, B. L.; Hannon, W. H. Newborn Dried Blood Spot Screening: Residual Specimen Storage Issues. *Pediatrics* **2012**, *129* (2), 365–366.

(3) California Department of Public Health (CDPH). Background and History of the California Biobank Program (CBP) https://www.cdph.ca.gov/programs/GDSP/Pages/MoreAboutTheCBP.aspx (accessed Feb 22, 2017).

(4) Gale, K. B.; Ford, A. M.; Repp, R.; Borkhardt, A.; Keller, C.; Eden, O. B.; Greaves, M. F. Backtracking Leukemia to Birth: Identification of Clonotypic Gene Fusion Sequences in Neonatal Blood Spots. *Proc. Natl. Acad. Sci.* **1997**, *94* (25), 13950–13954.

(5) Greaves, M. In Utero Origins of Childhood Leukaemia. *Early Hum. Dev.* **2005**, *81* (1), 123–129.

(6) Khoo, S. K.; Dykema, K.; Vadlapatla, N. M.; LaHaie, D.; Valle, S.; Satterthwaite, D.; Ramirez, S. A.; Carruthers, J. A.; Haak, P. T.; Resau, J. H. Acquiring Genome-Wide Gene Expression Profiles in Guthrie Card Blood Spots Using Microarrays. *Pathol. Int.* **2011**, *61* (1), 1–6.

(7) Petrick, L.; Edmands, W.; Schiffman, C.; Grigoryan, H.; Perttula, K.; Yano, Y.; Dudoit, S.; Whitehead, T.; Metayer, C.; Rappaport, S. An Untargeted Metabolomics Method for Archived Newborn Dried Blood Spots in Epidemiologic Studies. *Metabolomics* **2017**, *13* (3), 27.

(8) Ignjatovic, V.; Pitt, J.; Monagle, P.; Craig, J. M. The Utility of Dried Blood Spots for Proteomic Studies: Looking Forward to Looking Back. *Proteomics - Clin. Appl.* **2014**, *8* (11–12), 896–900.

(9) Mei, J. V.; Alexander, J. R.; Adam, B. W.; Hannon, W. H. Use of Filter Paper for the Collection and Analysis of Human Whole Blood Specimens. *J. Nutr.* **2001**, *131*, 1631–1636.

(10) Wagner, M.; Tonoli, D.; Varesio, E.; Hopfgartner, G. The Use of Mass Spectrometry to Analyze Dried Blood Spots. *Mass Spectrom. Rev.* **2016**, *35* (3), 361–438.

(11) Hall, E.; Flores, S.; De Jesús, V. Influence of Hematocrit and Total-Spot Volume on Performance Characteristics of Dried Blood Spots for Newborn Screening. *Int. J. Neonatal Screen.* **2015**, *1* (2), 69–78.

(12) O'Broin, S. D.; Kelleher, B. P.; Gunter, E. Evaluation of Factors Influencing Precision in the Analysis of Samples Taken from Blood Spots on Filter Paper. *Clin. Lab. Haematol.* **1995**, *17* (2), 185–188.

(13) Orsini, J. J.; Yeman, J.; Bodamer, O. A.; Mühl, A.; Caggana, M. Semi-Quantitative Method for Determination of Hematocrit in Dried Blood Spots, Using Data Collected in HPLC Hemoglobin Variant Testing. *Clin. Chim. Acta* **2010**, *411* (11–12), 894–895.

(14) Li, W.; Tse, F. L. S. Dried Blood Spot Sampling in Combination with LC-MS/MS for Quantitative Analysis of Small Molecules. *Biomed. Chromatogr.* **2010**, *24* (October 2009), 49–65.

(15) Jopling, J.; Henry, E.; Wiedmeier, S. E.; Christensen, R. D. Reference Ranges for Hematocrit and Blood Hemoglobin Concentration during the Neonatal Period: Data from

a Multihospital Health Care System. *Pediatrics* **2009**, *123* (2), e333–e337.
(16) Capiau, S.; Stove, V. V.; Lambert, W. E.; Stove, C. P. Prediction of the Hematocrit of Dried Blood Spots via Potassium Measurement on a Routine Clinical Chemistry Analyzer. *Anal. Chem.* **2013**, *85* (1), 404–410.
(17) De Kesel, P. M. M.; Capiau, S.; Stove, V. V.; Lambert, W. E.; Stove, C. P. Potassium-Based Algorithm Allows Correction for the Hematocrit Bias in Quantitative Analysis of Caffeine and Its Major Metabolite in Dried Blood Spots. *Anal. Bioanal. Chem.* **2014**, *406*, 6749–6755.
(18) Miller IV, J. H.; Poston, P. A.; Rutan, S. C.; Karnes H, T. An On-Card Approach for Assessment of Hematocrit on Dried Blood Spots Which Allows for Correction of Sample Volume. *J. Anal. Bioanal. Tech.* **2013**, *04* (01), 1–8.
(19) Capiau, S.; Wilk, L. S.; Aalders, M. C. G.; Stove, C. P. A Novel, Nondestructive, Dried Blood Spot-Based Hematocrit Prediction Method Using Noncontact Diffuse Reflectance Spectroscopy. *Anal. Chem.* **2016**, *88* (12), 6538–6546.
(20) Oostendorp, M.; El Amrani, M.; Diemel, E. C.; Hekman, D.; Van Maarseveen, E. M. Measurement of Hematocrit in Dried Blood Spots Using Near-Infrared Spectroscopy: Robust, Fast, and Nondestructive. *Clin. Chem.* **2016**, *62* (11), 1534–1536.
(21) O'Connor, G.; Molloy, A. M.; Daly, L.; Scott, J. M. Deriving a Useful Packed Cell Volume Estimate from Haemoglobin Analysis. *J.Clin.Pathol.* **1994**, *47* (0021–9746 (Print)), 78–79.
(22) Metayer, C.; Zhang, L.; Wiemels, J. L.; Bartley, K.; Schiffman, J.; Ma, X.; Aldrich, M. C.; Chang, J. S.; Selvin, S.; Fu, C. H.; et al. Tobacco Smoke Exposure and the Risk of Childhood Acute Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer Epidemiol. Biomarkers Prev.* **2013**, *22* (9), 1600–1611.
(23) ter Heine, R.; Rosing, H.; van Gorp, E. C. M.; Mulder, J. W.; van der Steeg, W. A.; Beijnen, J. H.; Huitema, A. D. R. Quantification of Protease Inhibitors and Non-Nucleoside Reverse Transcriptase Inhibitors in Dried Blood Spots by Liquid Chromatography-Triple Quadrupole Mass Spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2008**, *867* (2), 205–212.
(24) R Development Core Team, (2016). R: A Language and Environment for Statistical Computing. https://www.r-project.org/ (accessed Mar 8, 2017).
(25) Bates, D.; Mächler, M.; Bolker, B. M.; Walker, S. C. Fitting Linear Mixed-Effects Models Using Lme4. *ArXIV e-print; Press. J. Stat. Softw.* **2015**.
(26) Searle, S. R.; Casella, G.; McCulloch, C. E. *Variance Components*; John Wiley & Sons, Inc.: New York, 1992.
(27) Ludbrook, J. Linear Regression Analysis for Comparing Two Measurers or Methods of Measurement: But Which Regression? *Clin. Exp. Pharmacol. Physiol.* **2010**, *37* (7), 692–699.
(28) Makowski, G. S.; Davis, E. L.; Hopfer, S. M. The Effect of Storage on Guthrie Cards: Implications for Deoxyribonucleic Acid Amplification. *Ann. Clin. Lab. Sci.* **1996**, *26* (5), 458–469.
(29) Bremmer, R. H.; Nadort, A.; van Leeuwen, T. G.; van Gemert, M. J. C.; Aalders, M. C. G. Age Estimation of Blood Stains by Hemoglobin Derivative Determination Using Reflectance Spectroscopy. *Forensic Sci. Int.* **2011**, *206* (1–3), 166–171.
(30) Olney, R. S.; Moore, C. A.; Ojodu, J. A.; Lindegren, M. Lou; Hannon, W. H. Storage and Use of Residual Dried Blood Spots from State Newborn Screening Programs. *J. Pediatr.*

**2006**, *148* (5), 618–622.
(31) De Kesel, P. M. M.; Capiau, S.; Lambert, W. E. Current Strategies for Coping with the Hematocrit Problem in Dried Blood Spot Analysis. *Bioanalysis* **2014**, *6*, 1871–1874.

## 2.9 Supporting information

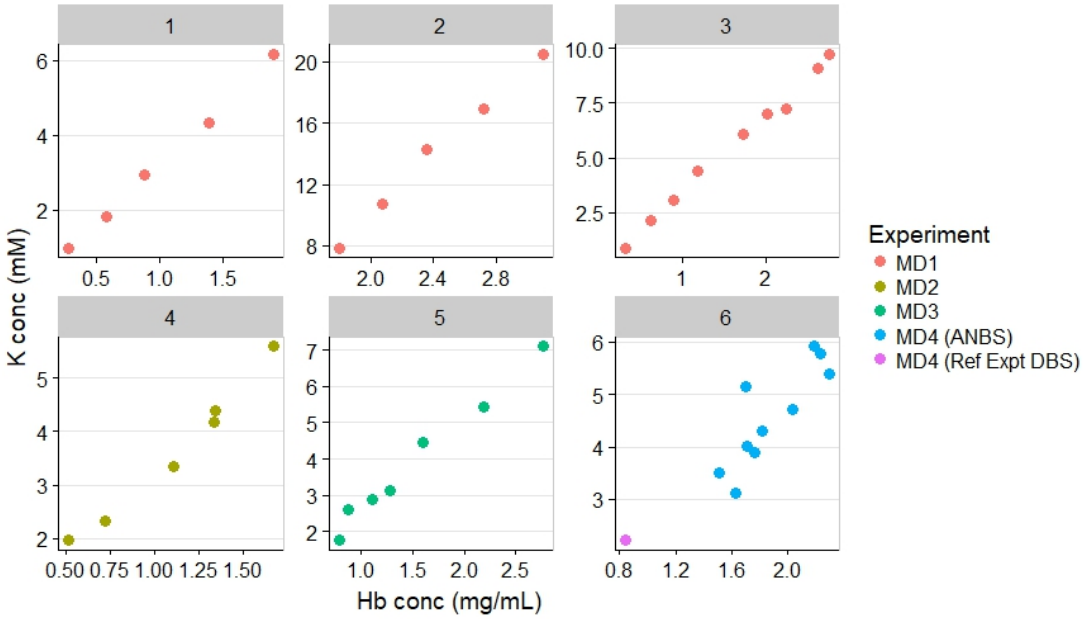



**Figure S2. 1** Plots of $K^+$ and Hb concentrations from the six batches of method-development experiments shown individually. MD1 was performed in three batches: (batch 1) punch sizes 2.0 – 6.0-mm, (batch 2) 6.0 – 7.8-mm, and (batch 3) 2.0 – 7.8-mm. MD2, MD3, and MD4 were performed in batch 4, 5, and 6, respectively. Colors represent the different types of method-development experiments: size (MD1), extraction volume (MD2), adjusted Hct (MD3), ANBS (MD4), and experimental DBS analyzed with ANBS as reference points (Ref Expt DBS, MD4).

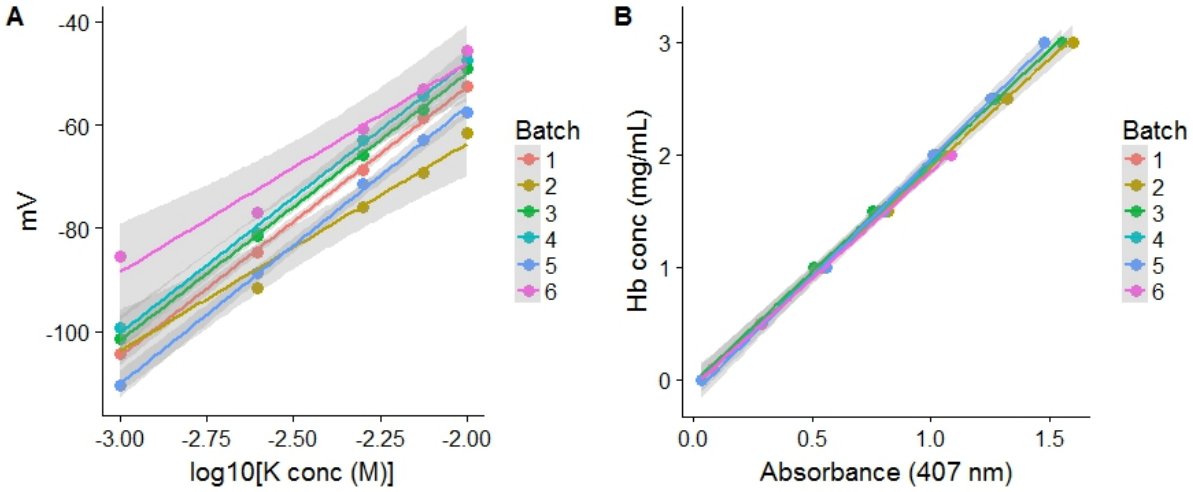



**Figure S2. 2** Calibration curves used for the six method-development experiments. Linear regression lines with pointwise 95% confidence bands (shaded) are shown. Batches represent independent experiments: MD1 (batch 1 – 3), MD2 (batch 4), MD3 (batch 5), and MD4 (batch 6). (A) $\mathbf{K^+}$ calibration curves were used to convert mV measured with an ion-selective micro-electrode in the DBS extracts to $\mathbf{K^+}$ concentrations in M (in log scale). The slope did not meet the ~55 mV value suggested by the manufacturer for batches 2 and 6. (B) Hb calibration curves were used to convert absorbance measurements at 407 nm to Hb concentrations in the DBS extract in mg/mL.

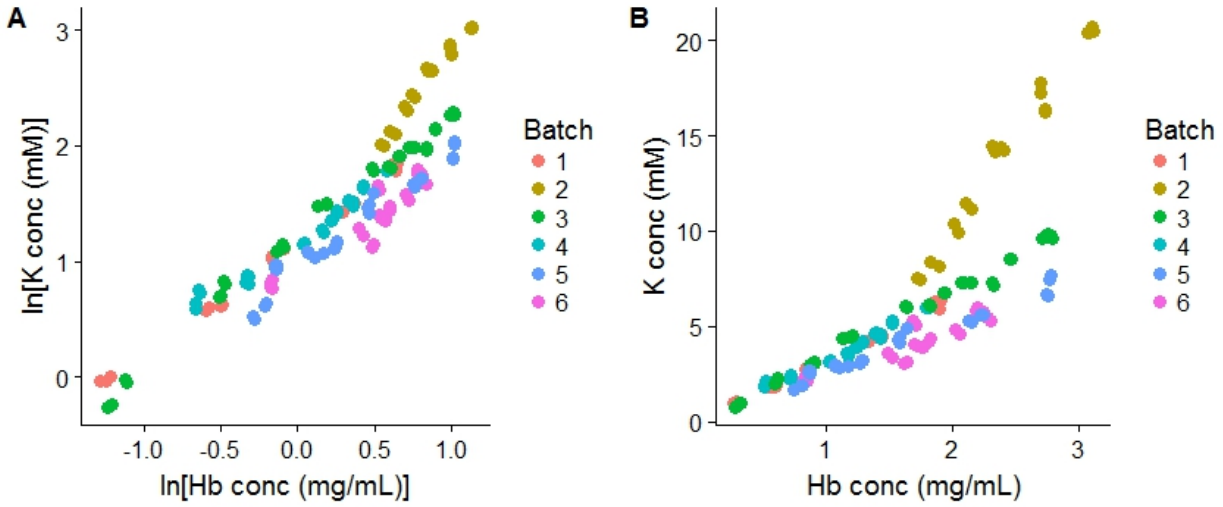



**Figure S2. 3** Unadjusted $\mathbf{K}^+$ and Hb concentrations measured in DBS extracts across all method-development experiments, colored by batch. Batches represent independent experiments: MD1 (batch 1 – 3), MD2 (batch 4), MD3 (batch 5), and MD4 (batch 6). Plots of $\mathbf{K}^+$(mM) and Hb (mg/mL) concentrations from (A) log-transformed data and (B) untransformed data.

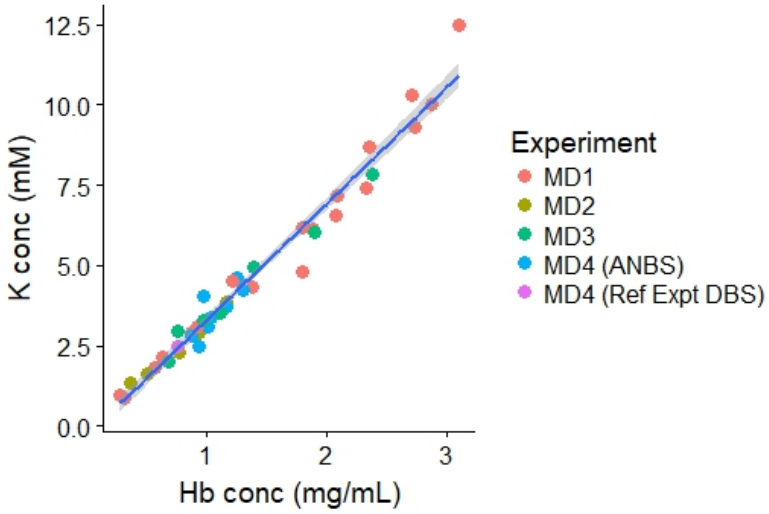



**Figure S2. 4** Linear relationship between $\mathbf{K}^+$ (mM) and Hb (mg/mL) concentrations in DBS extracts (untransformed data). Colors represent the different types of method-development experiments: size (MD1), extraction volume (MD2), adjusted Hct MD3), ANBS (MD4), and experimental DBS analyzed with ANBS as reference points (Ref Expt DBS, MD4). The regression line (blue) and the pointwise 95% confidence band (shaded) are shown. The slope of the line was 3.63 (95 % CI: 3.43, 3.83) and the intercept was -0.34 (95 % CI: -0.65, -0.03).

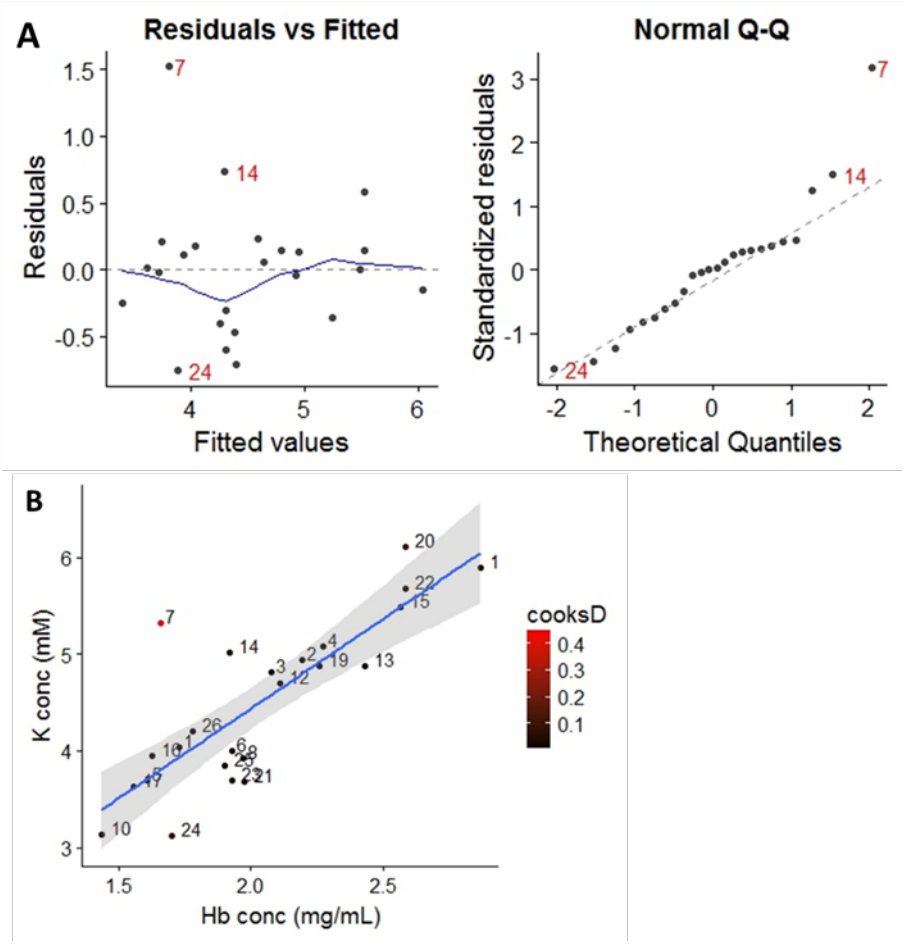



**Figure S2. 5** Model diagnostics from linear regression analysis of predicted and observed $\mathbf{K^+}$ concentrations from the 24 ANBS from the CCLS in V1: (A) Plot of residuals against fitted values (left) and a normal Q-Q plot of the residuals (right). Sample 7 deviates from the assumptions of linearity and normality. (B) Scatter plot of observed (x-axis) and predicted (y-axis) $\mathbf{K^+}$ concentrations in these ANBS with colors representing low (black) to high (red) Cook's D values. Sample 7 (red point) was identified as an influential outlier due to its relatively high Cook's D value. The linear regression line and the pointwise 95% confidence band (shaded) are shown.

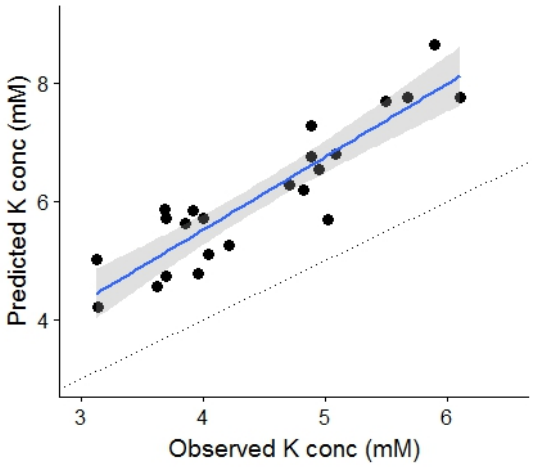



**Figure S2. 6** Plot of observed (x-axis) and predicted (y-axis) $\mathbf{K^+}$ concentrations (mM) of the 24 ANBS from CCLS analyzed in V1. The linear regression line (blue) and the pointwise 95% confidence band (shaded) are shown. The dotted line represents the line of equality (y = x) representing perfect agreement between the two measurements.

# Chapter 3.

# Untargeted Adductomics of Cys34 Modifications to Human Serum Albumin in Newborn Dried Blood Spots

Yukiko Yano,[1] Hasmik Grigoryan,[1] Courtney Schiffman,[2] William Edmands,[1] Lauren Petrick,[3] Katie Hall,[1] Todd Whitehead,[4,5] Catherine Metayer,[4,5] Sandrine Dudoit,[2] and Stephen Rappaport[1,5]

1. Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA 94720 USA
2. Division of Biostatistics, School of Public Health, University of California, Berkeley, CA 94720 USA
3. The Senator Frank R. Lautenberg Environmental Health Sciences Laboratory, Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029 USA
4. Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720 USA
5. Center for Integrative Research on Childhood Leukemia and the Environment, University of California, Berkeley, CA 94720 USA

## 3.1 Abstract

Metabolism of chemicals from the diet, exposures to xenobiotics, the microbiome, and lifestyle factors (e.g., smoking, alcohol intake) produce electrophiles that react with nucleophilic sites in circulating proteins, notably Cys34 of human serum albumin (HSA). To discover potential risk factors resulting from *in utero* exposures, we are investigating HSA-Cys34 adducts in archived newborn dried blood spots (DBS) that reflect systemic exposures during the last month of gestation. The workflow includes extraction of proteins from DBS, measurement of hemoglobin (Hb) to normalize for blood volume, addition of methanol to enrich HSA by precipitation of Hb and other interfering proteins, digestion with trypsin, and detection of HSA-Cys34 adducts via nanoflow liquid chromatography-high resolution mass spectrometry. As proof-of-principle, we applied the method to 49 archived DBS collected from newborns whose mothers either actively smoked during pregnancy or were nonsmokers. Twenty-six HSA-Cys34 adducts were detected, including Cys34 oxidation products, mixed disulfides with low-molecular-weight thiols (*e.g.*, cysteine, homocysteine, glutathione, cysteinylglycine, etc.), and other modifications. Data were normalized with a novel method ('scone') to remove unwanted technical variation arising from: HSA digestion, blood volume, DBS age, mass spectrometry analysis, and batch effects. Using an ensemble of linear and nonlinear models, the Cys34 adduct of cyanide was found to consistently discriminate between newborns of smoking and

nonsmoking mothers with a mean fold change (smoking/nonsmoking) of 1.31. These results indicate that DBS-adductomics is suitable for investigating *in utero* exposures to reactive chemicals and metabolites that may influence disease risks later in life.

## 3.2 Introduction

Many carcinogens are reactive electrophiles that are generated through metabolism of chemicals from the diet, environmental exposures, the microbiome, and lifestyle factors such as smoking and alcohol consumption. Although these reactive intermediates are short-lived *in vivo*, they can be quantified by measuring their reaction products (adducts) with circulating proteins, such as hemoglobin (Hb) and human serum albumin (HSA).[1,2] We have focused on HSA adducts bound to the highly nucleophilic sulfhydryl group at Cys34, which is a powerful antioxidant and scavenger of reactive electrophiles in the interstitial space.[3] Whereas targeted assays are limited to measurement of particular HSA-Cys34 adducts known *a priori*, our adductomics methodology motivates discovery and quantitation of unknown HSA modifications of potential health significance.[4]

In our Cys34 adductomics pipeline, HSA is isolated from plasma/serum, digested with trypsin, and analyzed via nanoflow liquid chromatography-high resolution mass spectrometry (nLC-HRMS) to find and quantitate modifications at the third largest tryptic peptide (T3) ($^{21}$ALVLIAFAQYLQQC$^{34}$PFEDHVK$^{41}$).[5] In four previous studies, we applied this adductomics pipeline to plasma/serum from healthy smokers and nonsmokers in the U.S.,[5] nonsmoking women in China exposed to indoor combustion products and local controls,[6] nonsmoking British patients with lung and heart disease and local controls,[7] and nonsmoking Chinese workers exposed to benzene and local controls.[8] Here, we extended the adductomics assay to measure Cys34 adducts in newborn dried blood spots (DBS).

Because newborn DBS have been routinely collected at birth to screen for inborn errors of metabolism in the U.S. and worldwide,[9] analysis of archived newborn DBS provides an avenue for investigating the etiologies of diseases initiated *in utero*. Retrospective investigations of chromosomal translocations in DNA from newborn DBS provide direct evidence of the prenatal origin of childhood leukemia, the most common childhood cancer.[10–12] Chronic diseases in adult life, such as type 2 diabetes mellitus, cardiovascular disease, and the metabolic syndrome, can also have fetal origins.[13] Since HSA has a residence time of 28 days,[4] measuring Cys34 adducts in newborn DBS would allow us to investigate exposures to reactive and potentially carcinogenic electrophiles during the last month of gestation.

Here, we describe untargeted measurements of HSA-Cys34 adducts in newborn DBS. A major challenge involved interfering species from the sample matrix, particularly lysed red blood cells and Hb that are not abundant in serum or plasma.[14,15] Indeed, Hb is present at a 7-fold higher concentration than HSA in whole blood[16] and interferes with tryptic digestion that releases the T3 peptide and its modifications for analysis.[17] Using controlled addition of methanol, we effectively removed Hb from DBS extracts prior to digestion and detected HSA-Cys34 adducts via nLC-HRMS in archived DBS from 49 newborns with mothers who either actively smoked during pregnancy or were nonsmokers. After using a novel method to normalize the data, we applied an ensemble of statistical methods to select adducts that discriminated between newborns of smoking and nonsmoking mothers.

## 3.3 Materials and methods

### *Chemicals and Reagents*

Acetonitrile (Ultra Chromasolv, LCMS grade), triethylammonium bicarbonate (TEAB) buffer (1 M), ethylenediamine-tetraacetic acid (EDTA, anhydrous), HSA (lyophilized powder, 97-99%), and porcine trypsin were from Sigma-Aldrich (St. Louis, MO). Methanol (Optima, LCMS grade), formic acid (Optima, LCMS grade), and iodoacetamide (IAA) were from Fisher Scientific (Pittsburgh, PA). Purified human Hb was purchased from MP Biomedicals, LLC (Santa Ana, CA). Isotopically labeled T3 (iT3) with sequence AL-[$^{15}$N,$^{13}$C-Val]-LIAFAQYLQQCPFEDH-[$^{15}$N,$^{13}$C-Val]-K was custom-made (>95%, BioMer Technology, Pleasanton, CA), and the carbamidomethylated iT3 (IAA-iT3)[18] was used as an internal standard to monitor retention time and mass drifts. Water was prepared with a PureLab purification system (18.2 mΩ cm resistivity at 25 ℃; Elga LabWater, Woodridge, IL).

### *Preparation of Capillary DBS for Method Development*

For method development, capillary DBS were collected with informed consent from adult volunteers by finger prick with a sterile safety lancet (Fisher HealthCare, Houston, TX). The first drop of blood was discarded and subsequent drops were collected on Whatman 903 Protein Saver cards (GE Healthcare, Cardiff, UK). Blood spots were air dried for a minimum of 4 days and stored at -20 ℃ in glassine envelopes (GE Healthcare, Cardiff, UK) prior to use. Punches of 5 and 6-mm diameter were obtained from these DBS with a Biopunch (Ted Pella Inc., Redding, CA).

### *Archived Newborn DBS*

Newborn DBS were obtained with informed consent for 49 healthy control children from the California Childhood Leukemia Study (CCLS).[19] These newborn DBS had been archived by the California Department of Public Health[20] at -20 ℃ for 14 to 32 y prior to analysis in the current investigation. Twenty-three of these participants had mothers who actively smoked during pregnancy and the remaining 26 had nonsmoking mothers. Interviews with the biological mother were conducted to collect data on the child's sex, race, and mother's smoking status during pregnancy. A total of 23 smoking/nonsmoking pairs were matched on sex and child's birth year. Smoking/nonsmoking pairs of newborn samples were analyzed together to minimize technical variation.

Our methodology was developed for 4.7-mm punches from DBS. Because the archived newborn DBS for the present investigation were remnants from previous analyses, they consisted of areas of filter media equivalent to 4.7-mm punches based on size and weight.

### *Extraction of Proteins and Measurements of Hb and Total Protein*

DBS punches were placed in microcentrifuge tubes and extracted with 55 $\mu$L of water at room temperature for 15 min with constant agitation at 1400 rpm (Thermomixer, Eppendorf, Hamburg, Germany). Samples were then centrifuged for 10 s and 5-$\mu$L aliquots were diluted with 45 $\mu$L of water to measure Hb concentrations (for normalization of blood volumes) with a Cytation 5 microplate spectrophotometer (BioTek Instruments, Winooski, VT) at room temperature. The absorbance of duplicate 2.5-$\mu$L sample aliquots was measured at 407 nm, which was the experimentally-determined absorbance maximum corresponding to heme in the *ex*

*vivo* oxidation state of Hb.[21,22] Absorbance readings were converted into Hb concentrations with five-point linear calibration curves using Hb standard solutions ranging from 0.5 to 5.0 mg/mL.

Absorbance measurements at 280 nm were used to calculate total protein concentrations in DBS extracts with correction for nucleic acid interferences at 260 nm and background correction at 320 nm.[23]

***Sample Preparation for Adductomics***

Various extraction protocols were tested and the method described below was found to be optimal for isolating HSA from DBS while removing Hb and other proteins from the extract (Results and Discussion provides further details). After Hb measurement, 41 $\mu$L of methanol was added to the remaining 50 $\mu$L of DBS extract (resulting in 45% methanol), vortexed, and mixed at 37 ℃ for 30 min with agitation at 1400 rpm (Thermomixer, Eppendorf, Hamburg, Germany). Samples were then stored at 4 ℃ for 30 min and centrifuged at 14,000×g for 10 min to remove precipitates and cellulose fibers. A 55-$\mu$L aliquot of the supernatant was diluted with 95 $\mu$L of digestion buffer (50 mM TEAB, 1 mM EDTA, pH 8.0), and the solution was loaded into a Costar Spin-X centrifuge tube filter (0.22 $\mu$m cellulose acetate, Corning Incorporated, Corning, NY) and centrifuged at 10,000×g for 10 min. A 130-$\mu$L aliquot of the filtered solution (containing ~17% methanol to enhance trypsin activity)[5] was transferred into BaroFlex 8-well strips (Pressure Biosciences Inc., South Easton, MA) to which 1 $\mu$L of 10-$\mu$g/$\mu$L trypsin was added (~1:10 enzyme-to-protein ratio). Pressure-assisted proteolytic digestion was performed with a Barozyme HT48 (Pressure Biosciences Inc., South Easton, MA) instrument, which cycled between ambient pressure (30 s) and 1,380 bar (20 kpsi, 90 s) for 32 min. After digestion, 3 $\mu$L of 10% formic acid was added to denature trypsin. Digests were transferred to new tubes and centrifuged for 2 min at 10,000×g. A 100-$\mu$L aliquot of the digest was transferred to a 300-$\mu$L silanized glass vial (Micosolv Technology Corporation, Leland, NC), and 1 $\mu$L of the isotopically labeled internal standard (IAA-iT3, 20 pmol/$\mu$L) was added. Samples were stored in liquid nitrogen prior to nLC-HRMS. The 49 newborn DBS were processed daily in four batches of 12 or 13 samples.

***nLC-HRMS Analysis***

Digests were analyzed by an Orbitrap Elite HRMS coupled to a Dionex Ultimate 3000 nanoflow LC system via a Flex Ion nanoelectrospray ionization source (Thermo Fisher Scientific, Waltham, MA), as described previously.[5] A Dionex PepSwitft monolithic column (100 $\mu$m i.d. × 25 cm) (Thermo Scientific, Sunnyvale, CA) was used with mobile phases consisting of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B), with gradient elution (2-35 % B, 26 min) at a flow rate of 750 nL/min. Two 1- $\mu$L aliquots were injected for each sample. Full scan mass spectra were acquired in the positive ion mode with a resolution of 120,000 at $m/z$ 400 in the $m/z$ = 350 – 1500 mass range using the Orbitrap. The MS was operated in data-dependent mode to collect tandem MS (MS2) spectra in the linear ion trap. Additional details of nLC-HRMS analysis are available in Supporting Information.

***Identification, Quantitation, and Annotation of Putative Adducts***

HSA adducts were pinpointed using the adductomics pipeline, as described previously.[5] Adducts were grouped by monoisotopic mass (MIM) within 10 ppm and retention time (RT) within 1.5 min. The means of MIMs and RTs were calculated for each adduct across all samples. Masses added to the thiolate ion of the triply charged T3 peptide (Cys34-$S^-$) were calculated and

annotated as described previously[5] (see Supporting Information for details). Mass accuracies were estimated by the difference between theoretical and observed MIMs.

Peak picking and integration were performed using the Xcalibur Processing Method (version 3.0, Thermo Fisher Scientific, Waltham, MA) based on the average MIMs (5 ppm mass accuracy) and RTs of the putative adducts. Peaks were integrated with the Genesis algorithm after normalizing the RTs using the internal standard (IAA-iT3) and using a RT window of 60 s.

***Statistical Analysis***

All statistical analyses were performed using the R statistical programming environment.[24] Peak areas were log-transformed and the means of duplicate measurements were used after removal of four obvious outliers. Missing values were imputed using *k*-nearest-neighbor imputation (see Supporting Information for details).[25]

Data normalization was optimized using the Bioconductor R package 'scone,'[26] which explored different scaling methods and used a regression-based strategy to adjust for unwanted factors of variation.[27] Normalization workflows were ranked by a variety of metrics to reduce unwanted technical variation while preserving the biological variation of interest.[28] The top-ranking normalization scheme according to 'scone' used DESeq scaling and removed unwanted variation due to the following factors: digested HSA, blood volume, DBS age, instrument performance, and batch effects. Here, 'digested HSA' was quantified by the abundance of the tryptic housekeeping peptide adjacent to T3 with sequence $^{42}$LVNEVTEFAK$^{51}$ (doubly charged precursor ion at $m/z$ = 575.3111).[5] 'Blood volume' was indicated by measurement of Hb in DBS extracts. 'DBS age' (*i.e.*, 2017 – child birth year) was used to adjust for differences in the extraction efficiency due to the age of the DBS.[29] 'Instrument performance' was indicated by the drift in the abundance of the internal standard over time. 'Batch effect' was used to adjust for differences in the four subsets of samples that were prepared on different days.

To ensure robust associations with mothers' smoking status, a combination of linear and nonlinear models were used to identify discriminating adducts. First, the following multivariable regression model was applied:

$$Y_i = \beta_0 + \beta_1 X_{Smoke} + \beta_2 X_{Sex} + \beta_3 X_{Race} + \beta_4 X_{Batch} + \beta_5 X_{HK} + \beta_6 X_{IS} + \beta_3 X_{Hb} + \beta_3 X_{DBS\ Age} + \varepsilon_i, \quad (1)$$

where $Y_i$ is a vector of logged, DESeq scaled abundances of the *i*th adduct, $X_{Sex}$, and $X_{Race}$ (0 = other, 1 = white) are binary vectors, $X_{Batch}$ is a four-level categorical variable indicating batch, $X_{HK}$ is the vector of housekeeping peptide abundances, $X_{IS}$ is the vector of internal standard abundances, $X_{Hb}$ is the vector of Hb measurements, and $X_{DBS\ Age}$ is a vector of DBS sample ages. The coefficient $\beta_1$ and estimated $p$-values were used to rank adducts by their associations with mothers' smoking status. Significance levels were adjusted for multiple testing with the Benjamini-Hochberg method to control the false discovery rate (FDR) at $\alpha = 0.05$.[30] The mean fold change (smoking/nonsmoking) in adduct intensities between newborns of smoking and nonsmoking mothers was calculated as $\exp(\beta_1)$.

Next, a logistic least absolute shrinkage and selection operator (lasso)[31] model was fitted to the logged, scaled, and normalized adduct abundances from 'scone' along with the matching variables (*i.e.*, sex, birth year), to select a subset of adducts that best predicted the mothers' smoking status. To increase stability, the logistic lasso regression was performed on 500 bootstrapped data sets.[32] The percentage of times each adduct was selected by the lasso model out of the 500 iterations was used as a measure of variable importance. This process was

performed for a range of values of the lasso penalty parameter (lambda range: 0.12 ~ 0.20) to ensure that the final variable selections were robust.

Finally, adducts were also ranked by their associations with the mothers' smoking status using random forest variable importance.[33] A random forest with 500 trees was used to predict mothers' smoking status with the logged, scaled, and normalized adduct abundances in addition to matching variables. Adducts were ranked by the mean decrease in Gini index, which indicates the total decrease in node impurity (as measured by the Gini index when splitting on the adducts within the decision tree averaged over all trees).[34]

To evaluate the importance of selected adducts that discriminated between the smoking and nonsmoking mothers based on all three statistical approaches, we obtained 10-fold cross-validated area under the estimated receiver operating characteristic (ROC) curve (AUC) by fitting a logistic regression model on smoking status with discriminating adducts as predictors. Cross-validated AUC estimates were obtained with the 'cvAUC' R package.[35]

## 3.4 Results and discussion

### *Measurement of Hb in Archived DBS*

Our analysis was performed exclusively with newborn DBS that had been maintained in freezers at -20 ℃ for 14 to 32 y prior to analysis. Quantitative analysis of Hb in DBS stored at room temperature can be problematic because of oxidation of Hb.[36] Indeed, when DBS were stored at room temperature for several months compared to storage at -20 ℃, we observed that the color changed from deep red to dark brown indicating oxidation of Hb,[37] and decreased water solubility of Hb (data not shown). The absorption spectra in the 250 – 750 nm range for the Hb calibration curve (Figure S3.1A) and Hb measured from extracts of 10 randomly selected 4.7-mm punches from archived newborn DBS (Figure S3.1B) both showed maximum absorbance at 407 nm. The Hb calibration curves measured for each of the four batches of newborn DBS showed a strong linear relationship between Hb concentrations and absorbance measurements at 407 nm ($r^2 > 0.99$, Figure S3.1C). Details on oxidation of Hb is provided in Supporting Information.

### *Adductomics Analysis of DBS*

In preparing DBS for adductomics, HSA must first be extracted from the filter paper and isolated from whole blood. Previous analyses of proteins extracted from DBS have used mixed aqueous-organic solutions to selectively precipitate proteins in solution.[38,39] Since Hb is one of the most prominent interfering proteins in whole blood, we tested various mixtures of organic solvents (ethanol, methanol, acetonitrile, and 1-propanol) to precipitate Hb while retaining HSA in solution (data not shown), and found ethanol and methanol to be most effective. When the concentration of ethanol and methanol were gradually increased from 30 to 60% (v/v), HSA remained in solution at concentrations less than 40% ethanol or 45% methanol and increasingly precipitated at concentrations up to 60% for both ethanol and methanol (Figure S3.2). Methanol was more effective at precipitating Hb with a ~95% decrease in concentration at 45% methanol when compared to DBS extracted with water (Figure S3.2).

The recovery of HSA was also influenced by the total protein concentration of the DBS extract, where higher total protein concentrations led to a lower recovery of HSA after precipitation (Figure S3.3). Based on preliminary analysis of ten 4.7-mm punches from archived newborn DBS, we found that the average total protein concentration for newborns (4.98 mg/mL)

was approximately equivalent to a 6-mm punch from an adult DBS. The observed higher total protein concentration of newborn blood reflects the higher Hb concentrations in newborns compared to adults.[40] Therefore, we used 6-mm punches from adult DBS to find the optimal concentration of methanol in the extraction mixture to isolate HSA. In addition, the isoelectric points (pI) of fetal and adult Hb differ (fetal Hb: pI 6.98, adult Hb: pI 6.87),[41] and fetal Hb precipitates more readily at neutral pH. When comparing 40, 45, 48, and 50% methanol, we observed that Hb did not precipitate with 40% methanol and that there was a ~40% loss of HSA when the methanol concentration was increased to 50% (Figure S3.4). Based on this result, we chose 45% methanol to isolate HSA in the DBS extract. We also found that incubating the samples at 37 °C (as opposed to room temperature) after addition of methanol to the aqueous DBS extract was essential for denaturing and precipitating Hb. When we tested extraction with 45% methanol on four 4.7-mm punches from archived newborn DBS, there was no loss of HSA and the residual Hb concentration was ~0.02 mg/mL (1.2% of the initial concentration) (Figure S3.5).

Digestion of HSA was optimized by testing various digestion programs using the pressure-cycling technology and by adjusting the proteolytic enzyme-to-protein ratio (E:P). While conventional proteomics approaches perform reduction and alkylation of proteins prior to digestion,[42] we did not apply these techniques in order to preserve Cys34 disulfides and to prevent the formation of artifacts. The digestion time was tested at 16, 32 and 64 min to determine the optimal time needed for a high yield of digestion. Both the total ion chromatogram and base peak chromatogram were examined for the presence of undigested proteins and yield of peptides.[17] While chromatograms from 32 and 64 min digests were comparable, there were fewer peptides and a more prominent peak for undigested proteins with the 16 min digestion, suggesting that 32 min was sufficient (Figure S3.6). Undigested protein was observed despite longer digestions, and probably reflects the lack of denaturation and reduction of disulfide bonds. It may also be due to the presence of residual non-HSA proteins, including Hb, which increase competition for trypsin cleavage sites and thereby interfere with the digestion of HSA.[42] We tested various 30 min digestion programs consisting of shorter and longer cycles at high pressure, but there was little difference in the resulting chromatograms (data not shown). The E:P was optimized to ensure an amount of trypsin that was sufficient for digestion while preventing autolysis.[43] When the E:P was increased from 1:18 to 1:3, trypsin activity showed a plateau at about 1:10, after which a further increase in trypsin did not improve the digestion (Figure S3.7). Increasing trypsin to a ratio of 1:5 resulted in incomplete digestion and more trypsin autolysis products, which could lead to ion suppression during MS detection.[44]

### *Analysis of Archived Newborn DBS*

Normalization of blood volumes based on Hb measurements proved to be effective in removing unwanted variation in adduct abundances as indicated by 'scone.' In addition, there was a relatively strong negative correlation (Pearson's $r$ = -0.56) between Hb concentrations and the second estimated factor of unwanted variation (Figure S3.8) using the RUVg method of Risso et al., a factor analysis method that estimates factors of unwanted variation via decomposition of a subset of the data consisting of negative control adducts.[27] This shows that Hb is indeed a factor of unwanted variation in DBS analysis. Furthermore, the weights of the DBS and Hb concentrations were highly correlated (Pearson's $r$ = 0.93), suggesting that Hb is a good predictor of blood volume in newborn DBS.

Relative log abundance (RLA) plots,[45] which were obtained by standardizing each adduct by the median abundance across samples and logging the resulting ratio, were used to visually inspect the reproducibility of replicate measurements and to assess the normalization scheme. Two subjects (of nonsmoking mothers) were removed from the analysis and only one measurement was used for one subject due to high variation in adduct abundances based on the RLA plot of duplicate measurements (Figure S3.9). Three adducts were removed from the analysis since they were missing in over half of the samples (Figure S3.10). This left 47 subjects and 26 putative Cys34 adducts for analysis.

Newborns of smoking and nonsmoking mothers, matched by sex and birth year, were similar with respect to race, DBS weight, and Hb concentrations (Table S3.1). The distribution of adduct peak areas in each sample before and after normalization for unwanted factors of variation (*i.e.*, Hb concentration, DBS age, housekeeping peptide, internal standard, and batch effects) is shown in Figure 3.1. By comparing the RLA plots from before (Figure 3.1A) and after (Figure 3.1B) normalization, it can be seen that this scheme effectively removed unwanted variation.

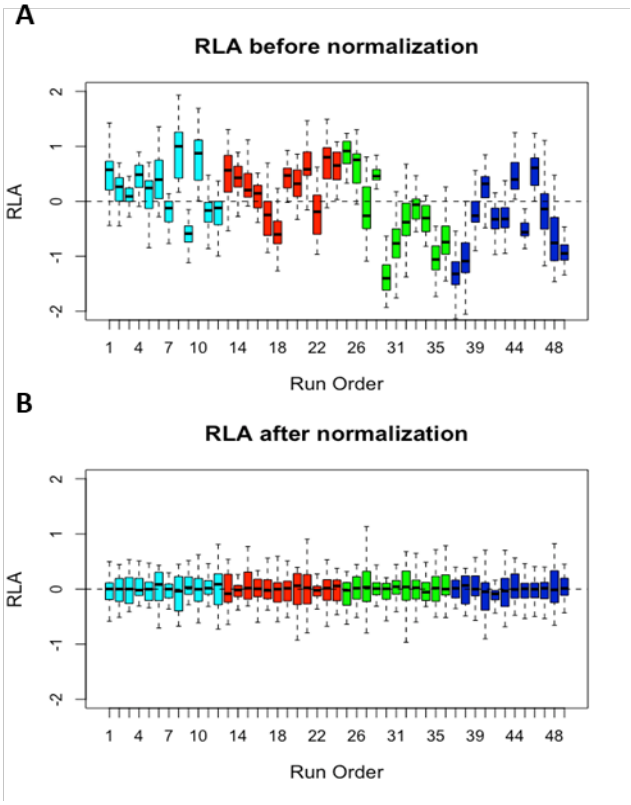



**Figure 3. 1** RLA plots from (A) before and (B) after normalization with 'scone.' Colors represent sample batches.

**Table 3. 1** Adducts detected in 47 archived DBS from newborns whose mothers smoked during pregnancy and those who did not.

| Adduct | RT (min) | MIM observed (*m/z*, 3+) | MIM theoretical (*m/z*, 3+) | ΔMass (ppm) | Added mass (Da) | Elemental composition of added mass | Annotation |
|---|---|---|---|---|---|---|---|
| 796.43[b,c,d] | 27.08 | 796.4304 | 796.4301 | 0.39 | -45.9854 | $-CH_2S$ | Cys34→Gly |
| 805.76[b,c,d] | 26.73 | 805.7616 | 805.7618 | -0.22 | -17.9919 | $-SH_2$, +O | Cys34→Oxoalanine |
| 810.43 | 26.85 | 810.4335 | | | -3.9761 | | Unknown |
| 811.09 | 26.60 | 811.0915 | 811.0875 | 4.98 | -2.0020 | $-H_2$ | Cys34 Sulfenamide |
| 811.76[a,b,c,d] | 27.67 | 811.7589 | 811.7594 | -0.61 | 1.0078 | +H | Unmodified T3[e] |
| 815.76 | 27.26 | 815.7593 | 815.7594 | -0.07 | 12.0013 | $+CH_2O$, $-H_2O$ | $CH_2$ crosslink |
| 816.42[a,b,c,d] | 25.43 | 816.4197 | 816.4191 | 0.72 | 13.9825 | $-H_2$, +O | Cys34 Sulfinamide[e] |
| 816.43[a,b,c,d] | 27.53 | 816.4312 | 816.4312 | -0.05 | 15.0248 | $+CH_3$ | Methylation (not at Cys34) |
| 819.09 | 27.59 | 819.0865 | 819.0867 | -0.23 | 22.9907 | +Na | S-Sodiation |
| 820.09[a,b,d] | 27.72 | 820.0914 | 820.0911 | 0.31 | 26.0053 | +CN | S-Cyanylation |
| 821.75[a] | 27.33 | 821.7506 | 821.7507 | -0.08 | 30.9831 | $-H_2$, $+O_2$ | Cys34 Sulfonamide |
| 822.42[a,b,c,d] | 26.70 | 822.4224 | 822.4226 | -0.28 | 32.9984 | $+HO_2$ | Cys34 Sulfinic acid[e] |
| 827.09[a,b] | 27.25 | 827.0947 | 827.0945 | 0.24 | 47.0153 | $+CH_3O_2$ | S-(O)-O-$CH_3$ |
| 827.75[a,b,c,d] | 27.00 | 827.7542 | 827.7543 | -0.08 | 48.9938 | $+HO_3$ | Cys34 Sulfonic acid[e] |
| 830.43[b] | 27.20 | 830.4342 | 830.4348 | -0.74 | 57.0337 | $+C_3H_5O$ | Unknown (possible S-addition of acrolein) |
| 832.43 | 27.43 | 832.4260 | 832.4262 | -0.24 | 63.0091 | $+CH_3O_3$ | Unknown |
| 835.11[b,d] | 27.29 | 835.1064 | 835.1066 | -0.35 | 71.0503 | $+C_4H_7O$ | S-Addition of crotonaldehyde[e] |
| 839.78[b] | 27.58 | 839.7777 | 839.7785 | -0.99 | 85.0643 | $+C_5H_9O$ | S-Addition of tiglic aldehyde[e] |
| 845.11 | 27.16 | 845.1110 | 845.1102 | 0.98 | 101.0642 | $+C_5H_9O_2$ | Unknown |
| 850.10[b,c,d] | 27.51 | 850.0950 | 850.0958 | -0.88 | 116.0162 | $+C_4H_6NOS$ | S-Addition of hCys (-H2O) |
| 851.42[a,b,c,d] | 26.12 | 851.4249 | 851.4274 | -2.90 | 120.0060 | $+C_3H_6NO_2S$ | S-Addition of Cys[e] |
| 856.10[a,b,c,d] | 26.46 | 856.0990 | 856.0993 | -0.32 | 134.0282 | $+C_4H_8NO_2S$ | S-Addition of hCys[e] |
| 857.44 | 27.67 | 857.4418 | 857.4419 | -0.10 | 138.0567 | $+C_7H_8NO_2$ | Unknown |
| 870.43[a,b,c,d] | 25.75 | 870.4346 | 870.4345 | 0.11 | 177.0351 | $+C_5H_9N_2O_3S$ | S-Addition of CysGly[e] |
| 913.45[a,b,c,d] | 26.06 | 913.4487 | 913.4487 | -0.06 | 306.0773 | $+C_{10}H_{16}N_3O_6S$ | S-Addition of GSH[e] |
| 918.12[a] | 26.39 | 918.1205 | 918.1206 | -0.16 | 320.0926 | $+C_{11}H_{18}N_3O_6S$ | Unknown (likely S-addition of GSH and $CH_2$ crosslink) |

Cys, cysteine; hCys, homocysteine; CysGly, cysteinylglycine; GSH, glutathione; MIM, monoisotopic mass; RT, retention time.

[a]Also detected by Liu et al. 2018.[7]

[b]Also detected by Grigoryan et al. 2018.[8]

[c]Also detected by Lu et al. 2017.[6]

[d]Also detected by Grigoryan et al. 2016.[5]

[e]Annotation confirmed with a synthetic standard

Table 3.1 summarizes the 26 adducts that were detected in the DBS of newborns from smoking and nonsmoking mothers. Representative MS2 spectra of the adducts are shown in Figure S3.11. Nineteen of the 26 adducts have been previously reported, including truncations, unmodified T3, methylated T3, Cys34 sulfoxidation products (*e.g.*, sulfinic and sulfonic acids), a cyanide adduct, and Cys34 disulfides of low-molecular-weight thiols.[5,6,8] Only three of the remaining 7 adducts had putative annotations, *i.e.*, the Cys34 sulfenamide (811.09), a $CH_2$ crosslink (815.76), and the sodium adduct (819.09). Aside from the unmodified T3 peptide (811.76), the Cys34 sulfinic acid (822.42) and the *S*-glutathione (GSH) disulfide (913.45) were the most abundant adducts across all samples (Figure S3.12).

In studying reaction pathways leading to Cys34 sulfoxidation products, Grigoryan et al. reported an intramolecular cyclic sulfinamide adduct (816.42) with the added mass ($+O, -H_2$), which results from the formation of a cross-link between Cys34 and the amide group of the adjacent Gln33.[18] Two different pathways were proposed for the formation of the sulfinamide adduct: (1) from dehydration of Cys34 sulfenic acid (Cys34-SOH) resulting in the sulfenamide adduct (mass difference [$-H_2$]) with the Cys34-Gln33 cross-link, which is then oxidized to the sulfinamide adduct; (2) from oxidation of the sulfenic acid to the sulfinic acid (822.42, Cys34-$SO_2H$), from which loss of water results in the sulfinamide adduct (Figure 3.2). The second reaction pathway appeared to be more likely because the intermediate sulfenamide adduct had not been detected in plasma/serum samples.[5,6,18] However, in newborn DBS we detected both the sulfenamide (811.09) and sulfinamide (816.42) adducts, suggesting that formation of the sulfinamide adduct is possible via both pathways. In addition, we detected the sulfonamide adduct (821.75, added mass [$+O_2, -H_2$]), which results from oxidation of the sulfinamide adduct (Figure 3.2).[46] The formation of these intramolecular cyclic adducts may have been promoted by the drying of DBS with subsequent dehydration of sulfenic, sulfinic, and sulfonic acids to produce the sulfenamide, sulfinamide, and sulfonamide adducts, respectively (Figure 3.2). It is also possible that these intramolecular cyclic adducts (particularly sulfenamide) were detected in the present analysis due to an increased stability of these analytes in DBS compared to plasma and serum. Analytes in DBS are typically less reactive than in liquid blood because they are stabilized through adsorption onto a solid cellulose matrix (*i.e.*, the filter paper).[47] Proteins commonly degrade in aqueous solutions due to aggregation, oxidation, and hydrolysis[48] that appear to be minimized during long-term storage of DBS in a freezer. In fact, we did not observe the T3 dimer in our analysis of newborn DBS although this dimer is routinely detected in plasma/serum samples.[5–8] Furthermore, adducts in DBS may be less susceptible to formation of artifacts because HSA is immobilized by the filter paper and less likely to interact with other molecules during storage.

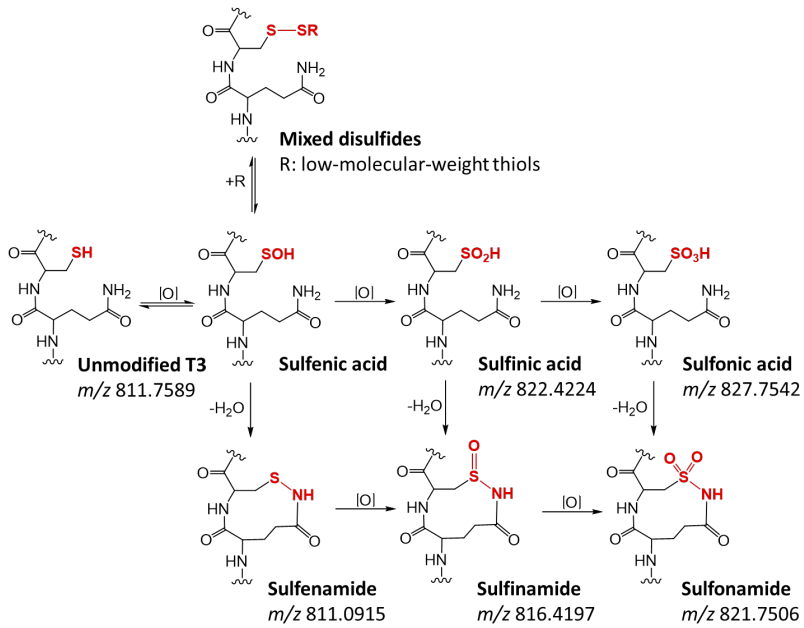



**Figure 3. 2** Reaction pathways proposed for the formation of Cys34 oxidation products (Adapted from Grigoryan et al. 2012).[18] Observed monoisotopic masses are shown.

### *Adducts that Discriminated Newborns of Smoking and Nonsmoking Mothers*

We had anticipated that Cys34 adducts of two toxic contaminants of cigarette smoke, *i.e.*, ethylene oxide and acrylonitrile, might be more abundant in newborns of smoking mothers given our previous detection of these adducts in plasma from adult smokers and their absence in plasma from nonsmokers.[5] However, these adducts were not detected in the newborn DBS, possibly due to low concentrations of the precursor molecules in newborn blood. Mothers may have stopped smoking during the third trimester or before the last month of pregnancy, and this may explain why we did not see all of the expected adducts in the present analysis.

We used three different statistical approaches to select adducts that discriminated newborns of smoking and nonsmoking mothers. The volcano plot (Figure 3.3A) shows the relationship between the smoker/nonsmoker fold change of a given adduct and the statistical significance of the difference in adduct abundance obtained from Equation 1. Discriminating adducts were also identified using lasso on bootstrapped data, which ranked adducts by the percentage of times each adduct was introduced into the model out of 500 iterations (Figure 3.3B) and by random forest (Figure 3.3C).

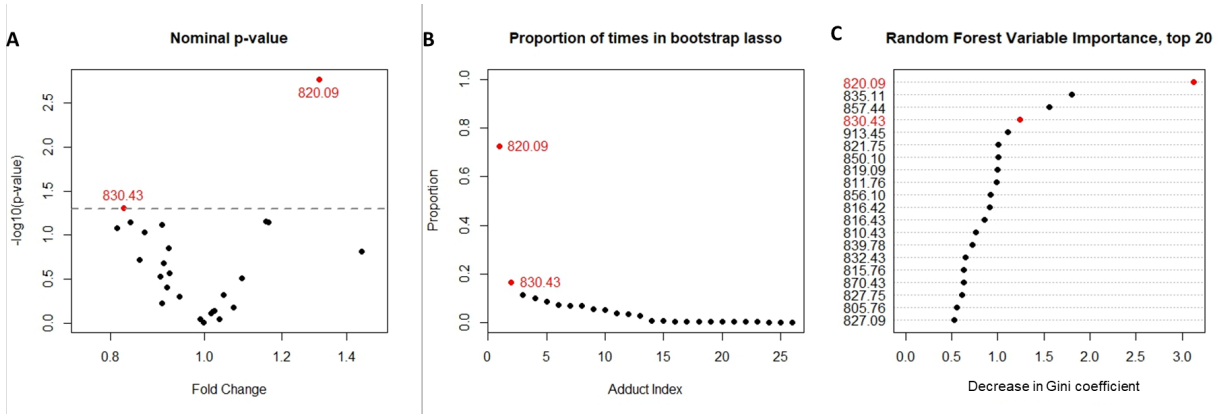



**Figure 3. 3** Selection of adducts that discriminate newborns by maternal smoking status. Points represent adducts and the two highest-ranked adducts are shown in red (820.09 is the cyanide adduct and 830.43 is an unknown adduct, possibly S-addition of acrolein). (A) Volcano plot showing the mean fold change for each adduct between newborns with smoking/nonsmoking mothers, and the corresponding statistical significance). The dashed line represents a nominal p-value of 0.05. (B) Proportion of times each adduct was selected into the lasso model out of 500 iterations. (C) The top 20 adducts ranked by random forest variable importance.

Of the 26 adducts that were tested, the Cys34 adduct of cyanide (820.09) was ranked the highest by all three statistical methods. As seen from the volcano plot (Figure 3.3A), the cyanide adduct showed a marked difference between newborns of smoking and nonsmoking mothers, with a smoker/nonsmoker mean fold change of 1.31 (nominal *p*-value = 0.0017; FDR-adjusted *p*-value = 0.044). The cyanide adduct was also top-ranked by both the lasso model (Figure 3.3B) and random forest (Figure 3.3C). Based on *p*-values from Equation 1, the second-ranked adduct (830.43) was an unknown with added mass corresponding to $+C_3H_5O$ (possibly *S*-addition of acrolein) with a fold change of 0.82 (nominal *p*-value = 0.049; adjusted *p*-value = 0.302) (Figure 3.3A). This unknown adduct was also ranked second by lasso (Figure 3.3B) and fourth by random forest (Figure 3.3C).

To test whether the cyanide adduct could be used to distinguish between newborns based on maternal smoking status, we performed a ROC analysis using logistic regression with the cyanide adduct as the predictor. The cyanide adduct alone provided a cross-validated AUC of 0.79 (95% CI: 0.65, 0.93). Although the AUC estimate is likely to be optimistic, because we did not have an independent test set for the ROC analysis, this indicates good discrimination between newborns of smoking and nonsmoking mothers.

The elevated levels of the cyanide adduct among newborns of smoking mothers are consistent with inhalation of hydrogen cyanide from tobacco smoke.[49] The half-life of cyanide in blood is less than one hour, which makes it difficult to obtain accurate measurements of cyanide exposure from the direct analysis of blood from smokers.[50] While more stable metabolites of cyanide, such as thiocyanate, are often used as surrogate measures of cyanide exposure, pairwise correlations are small between such metabolites and cyanide exposures.[50,51] Since the residence time of HSA is ~1 month,[4] the cyanide adduct of Cys34 arguably represents a more accurate measure of chronic low-level exposure to cyanide.[52]

### *Using Adductomics to Discover Biomarkers of in utero Exposures*

Adducts of HSA represent biomarkers of *in utero* exposures during the last month of gestation. A good example of such exposures is maternal smoking during pregnancy, which has

been consistently associated with increased risks of adverse birth outcomes (*e.g.*, low birth weight, preterm birth)[53] and has also been suggested to increase the risk of diseases later in life, including various types of cancer.[54–57] However, the long term effects of *in utero* tobacco-smoke exposures on the risk of childhood cancer have been inconsistent, with many studies reporting null associations.[58] One limitation of these epidemiological investigations has been reliance on maternal self-reports to retrospectively characterize fetal exposures to tobacco smoke.[58,59] Exposure misclassification due to recall and reporting bias is a particular concern among pregnant women, who may feel uncomfortable discussing their smoking histories during pregnancy, and can result in underestimation of fetal health effects from smoking mothers.[60]

Biomarkers complement interview-based exposure assessment by providing objective measures of exposure that are not susceptible to recall bias. Nicotine and its metabolite cotinine are commonly measured in biological fluids (*e.g.*, urine, blood, saliva) to assess tobacco smoke exposures.[60] For retrospective analyses of fetal exposures, archived newborn DBS are particularly attractive because they are readily available in California's repository that is maintained at -20 ℃. In addition, newborn DBS enable direct measurement of fetal exposures that can accumulate in the placenta and exceed those of the mother.[61]

Metabolites of nicotine, mainly cotinine, have been measured in newborn DBS to improve smoking surveillance among pregnant women.[62,63] However, the half-life of cotinine is only about 28 h in infants, and cotinine may only be detected in newborns of heavy smokers who smoke throughout pregnancy.[64] Since the residence time of HSA is 28 days,[4] Cys34 adducts detected in newborn DBS represent exposures received during the last month of gestation and are only marginally affected by the day-to-day variability in exposure.[65] In the present study, the Cys34 cyanide adduct discriminated between mothers who reported smoking during pregnancy vs. those who did not, suggesting that maternal self-reported smoking was reasonable in the 47 subjects tested.

## 3.5 Conclusion

By substantially modifying an untargeted HSA-Cys34 adductomics pipeline,[5] we were able to analyze Cys34 adducts extracted from newborn DBS. We then validated the methodology with archived DBS from newborns of smoking and nonsmoking mothers. Despite the small sample size and limited blood volume derived from 4.7-mm DBS punches, we were able to detect significantly higher levels of the Cys34 cyanide adduct among newborns whose mothers smoked during pregnancy. Since archived newborn DBS represent snapshots of *in utero* exposures, our adductomics method can be applied to investigate early-life exposures associated with childhood diseases. We are currently applying this methodology to archived newborn DBS from the CCLS to discover HSA-Cys34 adducts associated with childhood leukemia.

## 3.6 Compliance with ethical standards

***Conflict of interest***
The authors declare they have no conflict of interest.

***Ethics approval***
The CCLS and Center for Integrative Research on Childhood Leukemia and the Environment were approved by the University of California Committee for the Protection of Human Subjects, the California Health and Human Services Agency Committee for the Protection of Human Subjects, and the institutional review boards of all participating hospitals, as appropriate.

***Informed consent***
Written informed consent was obtained from all adult volunteer subjects and parents of all participating subjects in the CCLS.

## 3.7 Acknowledgements

This work was supported by the National Institute for Environmental Health Sciences of the U.S. National Institutes of Health (NIEHS grants P01 ES018172, P50ES018172, R01ES009137 and P42ES0470518), by the U.S. Environmental Protection Agency (USEPA grants RD83451101 and RD83615901), and by a grant for a pilot project from Children with Cancer, a registered Charity in the U.K.

The biospecimens used in this study were obtained from the California Biobank Program, (SIS request number 26), in accordance with Section 6555(b), 17 CCR. The NIEHS, USEPA, and California Department of Public Health are not responsible for the results or conclusions drawn by the authors of this publication.

## 3.8 References


(1) Rubino, F. M.; Pitton, M.; Di Fabio, D.; Colombi, A. Toward an "Omic " Physiopathology of Reactive Chemicals: Thirty Years of Mass Spectrometric Study of the Protein Adducts with Endogenous and Xenobiotic Compounds. *Mass Spectrom. Rev.* **2009**, *28* (5), 725–784.

(2) Törnqvist, M.; Fred, C.; Haglund, J.; Helleberg, H.; Paulsson, B.; Rydberg, P. Protein Adducts: Quantitative and Qualitative Aspects of Their Formation, Analysis and Applications. *J. Chromatogr. B* **2002**, *778* (1–2), 279–308.

(3) Aldini, G.; Vistoli, G.; Regazzoni, L.; Gamberoni, L.; Facino, R. M.; Yamaguchi, S.; Uchida, K.; Carini, M. Albumin Is the Main Nucleophilic Target of Human Plasma: A Protective Role Against Pro-Atherogenic Electrophilic Reactive Carbonyl Species? *Chem. Res. Toxicol.* **2008**, *21* (4), 824–835.

(4) Rappaport, S. M.; Li, H.; Grigoryan, H.; Funk, W. E.; Williams, E. R. Adductomics: Characterizing Exposures to Reactive Electrophiles. *Toxicol. Lett.* **2012**, *213* (1), 83–90.

(5) Grigoryan, H.; Edmands, W.; Lu, S. S.; Yano, Y.; Regazzoni, L.; Iavarone, A. T.; Williams, E. R.; Rappaport, S. M. Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin. *Anal. Chem.* **2016**, *88* (21), 10504–10512.

(6) Lu, S. S.; Grigoryan, H.; Edmands, W. M. B.; Hu, W.; Iavarone, A. T.; Hubbard, A.; Rothman, N.; Vermeulen, R.; Lan, Q.; Rappaport, S. M. Profiling the Serum Albumin Cys34 Adductome of Solid Fuel Users in Xuanwei and Fuyuan, China. *Environ. Sci. Technol.* **2017**, *51* (1), 46–57.

(7) Liu, S.; Grigoryan, H.; Edmands, W. M. B.; Dagnino, S.; Sinharay, R.; Cullinan, P.; Collins, P.; Chung, K. F.; Barratt, B.; Kelly, F. J.; et al. Cys34 Adductomes Differ between Patients with Chronic Lung or Heart Disease and Healthy Controls in Central London. *Environ. Sci. Technol.* **2018**, *52* (4), 2307–2313.

(8) Grigoryan, H.; Edmands, W. M. B.; Lan, Q.; Carlsson, H.; Vermeulen, R.; Zhang, L.; Yin, S.-N.; Li, G.-L.; Smith, M. T.; Rothman, N.; et al. Adductomic Signatures of Benzene Exposure Provide Insights into Cancer Induction. *Carcinogenesis* **2018**, *39* (5), 661–668.

(9) Therrell, B. L.; Padilla, C. D.; Loeber, J. G.; Kneisser, I.; Saadallah, A.; Borrajo, G. J. C.; Adams, J. Current Status of Newborn Screening Worldwide: 2015. *Semin. Perinatol.* **2015**, *39* (3), 171–187.

(10) Gale, K. B.; Ford, A. M.; Repp, R.; Borkhardt, A.; Keller, C.; Eden, O. B.; Greaves, M. F. Backtracking Leukemia to Birth: Identification of Clonotypic Gene Fusion Sequences in Neonatal Blood Spots. *Proc. Natl. Acad. Sci.* **1997**, *94* (25), 13950–13954.

(11) Greaves, M. In Utero Origins of Childhood Leukaemia. *Early Hum. Dev.* **2005**, *81* (1), 123–129.

(12) Wiemels, J. L.; Walsh, K. M.; de Smith, A. J.; Metayer, C.; Gonseth, S.; Hansen, H. M.; Francis, S. S.; Ojha, J.; Smirnov, I.; Barcellos, L.; et al. GWAS in Childhood Acute Lymphoblastic Leukemia Reveals Novel Genetic Associations at Chromosomes 17q12 and 8q24.21. *Nat. Commun.* **2018**, *9* (1), 286.

(13) Gluckman, P. D.; Hanson, M. A.; Cooper, C.; Thornburg, K. L. Effect of In Utero and Early-Life Conditions on Adult Health and Disease. *N. Engl. J. Med.* **2008**, *359* (1), 61–73.

(14) Chambers, A. G.; Percy, A. J.; Hardie, D. B.; Borchers, C. H. Comparison of Proteins in

Whole Blood and Dried Blood Spot Samples by LC/MS/MS. *J. Am. Soc. Mass Spectrom.* **2013**, *24* (9), 1338–1345.
(15) Martin, N. J.; Cooper, H. J. Challenges and Opportunities in Mass Spectrometric Analysis of Proteins from Dried Blood Spots. *Expert Rev. Proteomics* **2014**, *11* (6), 685–695.
(16) Chambers, A. G.; Percy, A. J.; Yang, J.; Borchers, C. H. Multiple Reaction Monitoring Enables Precise Quantification of 97 Proteins in Dried Blood Spots. *Mol. Cell. Proteomics* **2015**, *14* (11), 3094–3104.
(17) Hustoft, H. K.; Reubsaet, L.; Greibrokk, T.; Lundanes, E.; Malerod, H. Critical Assessment of Accelerating Trypsination Methods. *J. Pharm. Biomed. Anal.* **2011**, *56* (5), 1069–1078.
(18) Grigoryan, H.; Li, H.; Iavarone, A. T.; Williams, E. R.; Rappaport, S. M. Cys34 Adducts of Reactive Oxygen Species in Human Serum Albumin. *Chem. Res. Toxicol.* **2012**, *25* (8), 1633–1642.
(19) Metayer, C.; Zhang, L.; Wiemels, J. L.; Bartley, K.; Schiffman, J.; Ma, X.; Aldrich, M. C.; Chang, J. S.; Selvin, S.; Fu, C. H.; et al. Tobacco Smoke Exposure and the Risk of Childhood Acute Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer Epidemiol. Biomarkers Prev.* **2013**, *22* (9), 1600–1611.
(20) California Department of Public Health (CDPH). Newborn Screening Specimens Use and Storage https://www.cdph.ca.gov/Programs/CFH/DGDS/Pages/nbs/NBSDBS-Storage.aspx (accessed May 11, 2018).
(21) Waterman, M. R. [48] Spectral Characterization of Human Hemoglobin and Its Derivatives. In *Methods in Enzymology*; 1978; Vol. 703, pp 456–463.
(22) Meng, F.; Alayash, A. I. Determination of Extinction Coefficients of Human Hemoglobin in Various Redox States. *Anal. Biochem.* **2017**, *521*, 11–19.
(23) BioTek Instruments. *Multi-Volume Based Protein Quantification*; 2012.
(24) R Development Core Team, (2016). R: A Language and Environment for Statistical Computing. https://www.r-project.org/ (accessed Mar 8, 2017).
(25) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* **2001**, *17* (6), 520–525.
(26) Cole, M.; Risso, D. Scone: Single Cell Overview of Normalized Expression Data. *R Packag. version 1.2.0* **2018**.
(27) Risso, D.; Ngai, J.; Speed, T. P.; Dudoit, S. Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples. *Nat. Biotechnol.* **2014**, *32* (9), 896–902.
(28) Cole, M. B.; Risso, D.; Wagner, A.; DeTomaso, D.; Ngai, J.; Purdom, E.; Dudoit, S.; Yosef, N. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *bioRxiv* **2017**, 1–30.
(29) Petrick, L.; Edmands, W.; Schiffman, C.; Grigoryan, H.; Perttula, K.; Yano, Y.; Dudoit, S.; Whitehead, T.; Metayer, C.; Rappaport, S. An Untargeted Metabolomics Method for Archived Newborn Dried Blood Spots in Epidemiologic Studies. *Metabolomics* **2017**, *13* (3), 27.
(30) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57* (1), 289–300.
(31) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58* (1), 267–288.
(32) Bach, F. R. Bolasso: Model Consistent Lasso Estimation through the Bootstrap. In

*Proceedings of the 25th International Conference on Machine Learning*; ACM Press: New York, New York, USA, 2008; pp 33–40.
(33) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2* (3), 18–22.
(34) Calle, M. L.; Urrea, V. Letter to the Editor: Stability of Random Forest Importance Measures. *Brief. Bioinform.* **2011**, *12* (1), 86–89.
(35) LeDell, E.; Petersen, M.; Laan, M. Van Der. Cross-Validated Area Under the ROC Curve Confidence Intervals. *R Packag. version 1.1.0* **2016**.
(36) Bremmer, R. H.; de Bruin, D. M.; de Joode, M.; Buma, W. J.; van Leeuwen, T. G.; Aalders, M. C. G. Biphasic Oxidation of Oxy-Hemoglobin in Bloodstains. *PLoS One* **2011**, *6* (7), e21845.
(37) Bremmer, R. H.; Nadort, A.; van Leeuwen, T. G.; van Gemert, M. J. C.; Aalders, M. C. G. Age Estimation of Blood Stains by Hemoglobin Derivative Determination Using Reflectance Spectroscopy. *Forensic Sci. Int.* **2011**, *206* (1–3), 166–171.
(38) Funk, W. E.; Waidyanatha, S.; Chaing, S. H.; Rappaport, S. M. Hemoglobin Adducts of Benzene Oxide in Neonatal and Adult Dried Blood Spots. *Cancer Epidemiol. Biomarkers Prev.* **2008**, *17* (8), 1896–1901.
(39) DeWilde, A.; Sadilkova, K.; Sadilek, M.; Vasta, V.; Hahn, S. H. Tryptic Peptide Analysis of Ceruloplasmin in Dried Blood Spots Using Liquid Chromatography-Tandem Mass Spectrometry: Application to Newborn Screening. *Clin. Chem.* **2008**, *54* (12), 1961–1968.
(40) Andropoulos, D. B. Appendix B: Pediatric Normal Laboratory Values. In *Gregory's Pediatric Anesthesia*; Wiley-Blackwell: Oxford, UK, 2011; pp 1300–1314.
(41) 3. Fetal Hemoglobin: Structure and Function. *Scand. J. Clin. Lab. Invest.* **1982**, *42* (sup160), 32–37.
(42) Hustoft, H. K.; Malerod, H.; Wilson, S. R.; Reubsaet, L.; Lundanes, E.; Greibrokk, T. *A Critical Review of Trypsin Digestion for LC-MS Based Proteomics*; Leung, H.-C., Ed.; InTech, 2012.
(43) Klammer, A. A.; MacCoss, M. J. Effects of Modified Digestion Schemes on the Identification of Proteins from Complex Mixtures. *J. Proteome Res.* **2006**, *5* (3), 695–700.
(44) van den Broek, I.; Niessen, W. M. A.; van Dongen, W. D. Bioanalytical LC–MS/MS of Protein-Based Biopharmaceuticals. *J. Chromatogr. B* **2013**, *929*, 161–179.
(45) De Livera, A. M.; Dias, D. A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P. Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **2012**, *84* (24), 10768–10776.
(46) Go, Y.-M.; Chandler, J. D.; Jones, D. P. The Cysteine Proteome. *Free Radic. Biol. Med.* **2015**, *84*, 227–245.
(47) Wagner, M.; Tonoli, D.; Varesio, E.; Hopfgartner, G. The Use of Mass Spectrometry to Analyze Dried Blood Spots. *Mass Spectrom. Rev.* **2016**, *35* (3), 361–438.
(48) Björkesten, J.; Enroth, S.; Shen, Q.; Wik, L.; Hougaard, D. M.; Cohen, A. S.; Sörensen, L.; Giedraitis, V.; Ingelsson, M.; Larsson, A.; et al. Stability of Proteins in Dried Blood Spot Biobanks. *Mol. Cell. Proteomics* **2017**, *16* (7), 1286–1296.
(49) Talhout, R.; Schulz, T.; Florek, E.; Van Benthem, J.; Wester, P.; Opperhuizen, A. Hazardous Compounds in Tobacco Smoke. *Int. J. Environ. Res. Public Health* **2011**, *8* (2), 613–628.
(50) Vinnakota, C. V.; Peetha, N. S.; Perrizo, M. G.; Ferris, D. G.; Oda, R. P.; Rockwood, G. A.; Logue, B. A. Comparison of Cyanide Exposure Markers in the Biofluids of Smokers

and Non-Smokers. *Biomarkers* **2012**, *17* (7), 625–633.
(51) Fasco, M. J.; Stack, R. F.; Lu, S.; Hauer, C. R.; Schneider, E.; Dailey, M.; Aldous, K. M. Unique Cyanide Adduct in Human Serum Albumin: Potential as a Surrogate Exposure Marker. *Chem. Res. Toxicol.* **2011**, *24* (4), 505–514.
(52) Fasco, M. J.; Hauer, C. R.; Stack, R. F.; O'Hehir, C.; Barr, J. R.; Eadon, G. A. Cyanide Adducts with Human Plasma Proteins: Albumin as a Potential Exposure Surrogate. *Chem. Res. Toxicol.* **2007**, *20* (4), 677–684.
(53) DiFranza, J. R.; Aligne, C. A.; Weitzman, M. Prenatal and Postnatal Environmental Tobacco Smoke Exposure and Children's Health. *Pediatrics* **2004**, *113* (4), 1007–1015.
(54) Knopik, V. S.; Maccani, M. A.; Francazio, S.; McGeary, J. E. The Epigenetics of Maternal Cigarette Smoking during Pregnancy and Effects on Child Development. *Dev. Psychopathol.* **2012**, *24* (4), 1377–1390.
(55) de Smith, A. J.; Kaur, M.; Gonseth, S.; Endicott, A.; Selvin, S.; Zhang, L.; Roy, R.; Shao, X.; Hansen, H. M.; Kang, A. Y.; et al. Correlates of Prenatal and Early-Life Tobacco Smoke Exposure and Frequency of Common Gene Deletions in Childhood Acute Lymphoblastic Leukemia. *Cancer Res.* **2017**, *77* (7), 1674–1683.
(56) Metayer, C.; Petridou, E.; Aranguré, J. M. M.; Roman, E.; Schüz, J.; Magnani, C.; Mora, A. M.; Mueller, B. A.; de Oliveira, M. S. P.; Dockerty, J. D.; et al. Parental Tobacco Smoking and Acute Myeloid Leukemia. *Am. J. Epidemiol.* **2016**, *184* (4), 261–273.
(57) Gonseth, S.; de Smith, A. J.; Roy, R.; Zhou, M.; Lee, S.-T.; Shao, X.; Ohja, J.; Wrensch, M. R.; Walsh, K. M.; Metayer, C.; et al. Genetic Contribution to Variation in DNA Methylation at Maternal Smoking-Sensitive Loci in Exposed Neonates. *Epigenetics* **2016**, *11* (9), 664–673.
(58) Momen, N. C.; Olsen, J.; Gissler, M.; Li, J. Exposure to Maternal Smoking during Pregnancy and Risk of Childhood Cancer: A Study Using the Danish National Registers. *Cancer Causes Control* **2016**, *27* (3), 341–349.
(59) Whitehead, T. P.; Metayer, C.; Wiemels, J. L.; Singer, A. W.; Miller, M. D. Childhood Leukemia and Primary Prevention. *Curr. Probl. Pediatr. Adolesc. Health Care* **2016**, *46* (10), 317–352.
(60) Florescu, A.; Ferrence, R.; Einarson, T.; Selby, P.; Soldin, O.; Koren, G. Methods for Quantification of Exposure to Cigarette Smoking and Environmental Tobacco Smoke: Focus on Developmental Toxicology. *Ther. Drug Monit.* **2009**, *31* (1), 14–30.
(61) Holbrook, B. D. The Effects of Nicotine on Human Fetal Development. *Birth Defects Res. Part C* **2016**, *108* (2), 181–192.
(62) Spector, L. G.; Murphy, S. E.; Wickham, K. M.; Lindgren, B.; Joseph, A. M. Prenatal Tobacco Exposure and Cotinine in Newborn Dried Blood Spots. *Pediatrics* **2014**, *133* (6), e1632–e1638.
(63) Yang, J.; Pearl, M.; Jacob, P.; DeLorenze, G. N.; Benowitz, N. L.; Yu, L.; Havel, C.; Kharrazi, M. Levels of Cotinine in Dried Blood Specimens from Newborns as a Biomarker of Maternal Smoking Close to the Time of Delivery. *Am. J. Epidemiol.* **2013**, *178* (11), 1648–1654.
(64) Spector, L. G.; Hecht, S. S.; Ognjanovic, S.; Carmella, S. G.; Ross, J. A. Detection of Cotinine in Newborn Dried Blood Spots. *Cancer Epidemiol. Biomarkers Prev.* **2007**, *16* (9), 1902–1905.
(65) Rappaport, S. M.; Waidyanatha, S.; Qu, Q.; Shore, R.; Jin, X.; Cohen, B.; Chen, L. C.; Melikian, A. a.; Li, G.; Yin, S.; et al. Albumin Adducts of Benzene Oxide and 1,4-

Benzoquinone as Measures of Human Benzene Metabolism. *Cancer Res.* **2002**, *62* (5), 1330–1337.

## 3.9 Supporting information

### 3.9.1 Supplemental Methods

***Chemicals and Reagents***

HSA (lyophilized powder, 97-99%), acetic acid (LCMS grade), dimethyl sulfoxide, were from Sigma-Aldrich (St. Louis, MO). Tris base was from Fisher Scientific (Pittsburgh, PA). Glycine and sodium dodecyl sulfate (SDS) were from Bio-Rad (Hercules, CA). Dithiothreitol was from Promega (Madison, WI). Purified human Hb was purchased from MP Biomedicals, LLC (Santa Ana, CA).

***Gel Electrophoresis***

One-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was used during method development to optimize the isolation of HSA in DBS extracts. Aliquots of capillary DBS extracts (containing about 10 $\mu$g total protein) were mixed with an equal volume of Laemmli sample buffer (Bio-Rad, Hercules, CA) containing 0.35 M dithiothreitol and denatured at 95 ℃ for 10 min. SDS-PAGE was performed with ~5 $\mu$g protein/well on 10% precast polyacrylamide gels (Bio-Rad, Hercules, CA) in Tris-glycine SDS buffer (pH = 8.3) at 40 V for 10 min and subsequently at 120 V for ~60 min. Proteins were fixed with 40% methanol containing 10% acetic acid (30 min) and stained with Coomassie Brilliant Blue R-250 (Bio-Rad, Hercules, CA) for 40 min. Gels were destained overnight in 40% methanol containing 10% acetic acid and images of the gels were acquired with an AlphaImager HP system (Alpha Innotech, San Leandro, CA). Gel band densities were quantified using ImageJ software (NIH, https://imagej.nih.gov/ij/).

***nLC-HRMS analysis***

After each sample was injected in duplicate, the column was washed after every two samples by injecting 1 $\mu$L of a solution consisting of 80% acetonitrile, 10% acetic acid, 5% dimethyl sulfoxide, and 5% water.

The MS was operated in data-dependent mode, in which the fifteen most intense ions exceeding 12,000 counts in the full scan mass spectrum were subjected to tandem MS (MS2) using collision-induced dissociation (CID) in the linear ion trap with a normalized collision energy of 28%.

The lock mass option was enabled to provide a real-time internal mass calibration using a reference list of 3 identified background ions.[1] Dynamic exclusion was enabled with the following parameters: repeat count 2, repeat duration 20 s, exclusion list size 500, exclusion duration 90 s, and exclusion precursor ion width $\pm$20 ppm. Charge state screening was enabled to only select doubly and triply charged precursor ions in MS1 for CID fragmentation.

***k-nearest-neighbor imputation***

Pairwise distances between adducts were calculated using the Euclidean distance based on all non-missing values. To choose the value of the parameter *k*, 12 low abundance adducts were randomly chosen to be made missing in 10 samples, their values were imputed using several values of *k* ranging from 1 to 15, and the mean squared error was calculated for each *k*. The *k* with the smallest mean squared error, $k = 5$, was chosen.

***Adduct annotation***

The mass added to the thiolate ion of the triply charged T3 peptide (Cys34-$S^-$) was calculated by subtracting the observed MIM of the unmodified T3 peptide from the MIM of the putative adduct, multiplying the difference by 3 (charge state) and adding the MIM of one hydrogen atom ($m/z$ = 1.0078).

Putative adducts were annotated by obtaining the elemental compositions of the added mass using the Molecular Weight Calculator (version 6.50, www.alchemistmatt.com), Elemental Composition Finder from the Xcalibur Qual Browser (version 3.0, Thermo Fisher Scientific), ChemCalc Molecular Formula Finder (http://www.chemcalc.org/mf_finder), and UNIMOD (http://www.unimod.org/). Using the proposed empirical formula of a given putative adduct, the isotope distribution was simulated using the Xcalibur software, and the theoretical MIM was compared with the observed MIM to calculate mass accuracies. Database searches using the empirical formula were performed using the PubChem Compound database (https://pubchem.ncbi.nlm.nih.gov/) and the Human Metabolome Database (http://www.hmdb.ca/).

***Measurement of Hb in Archived DBS***

Upon exposure to oxygen, Hb is converted to oxyhemoglobin (oxyHb), which is further oxidized to methemoglobin (metHb) and denatured to hemichrome (HC).[2] The transition from oxyHb to metHb and HC is accompanied by the change in color from red to brown, resulting in spectral differences primarily in the ~500 – 650 nm region.[3] Various methods have previously been proposed to determine the sum of oxyHb, metHb, and HC (*i.e.*, 'total Hb') to quantify all Hb derivatives in DBS as a proxy for hematocrit.[2,4–6] While such methods may be important for the analysis of highly oxidized DBS, this was unnecessary for our analysis of newborn DBS archived at -20 ℃. Freezing temperatures have been shown to slow the oxidation process of Hb,[7] and we found that the conversion of oxyHb to metHb and HC was minimal in DBS maintained at -20 ℃, despite long-term storage.

The absorption spectra in the 250 – 750 nm range for the Hb calibration curve and Hb measured from extracts of 10 randomly selected 4.7-mm punches from archived newborn DBS are shown in Figure S-2. Absorption in the 450 – 750 nm range show spectral characteristics indicating that oxyHb is the predominant Hb form in both the Hb standards and the newborn DBS extracts, although there is some evidence for the presence of metHb at 630 nm (inset for Figure S-2A and S-2B).[3,6,8]

### 3.9.2 Supplemental Figures

**Figure S3. 1** Absorption spectra in the 250 – 750 nm range with insets showing the 450 – 750 nm range for the (A) Hb calibration curve and (B) ten 4.7-mm punches from archived newborn DBS. (C) Hb calibration curves used for determining Hb concentrations in DBS extracts based on the absorbance at 407 nm. The calibration curve was measured for each of the four batches of newborn.

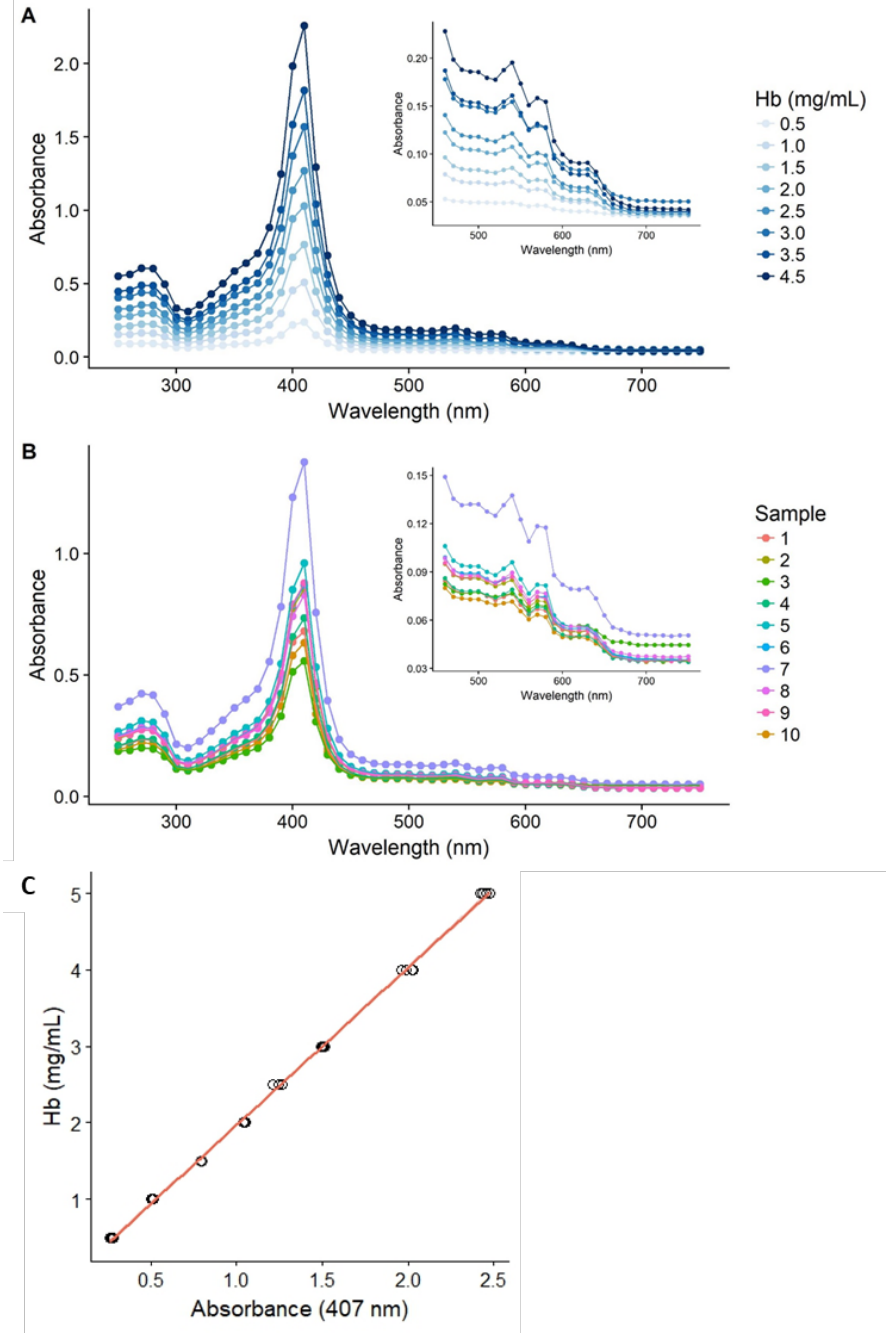
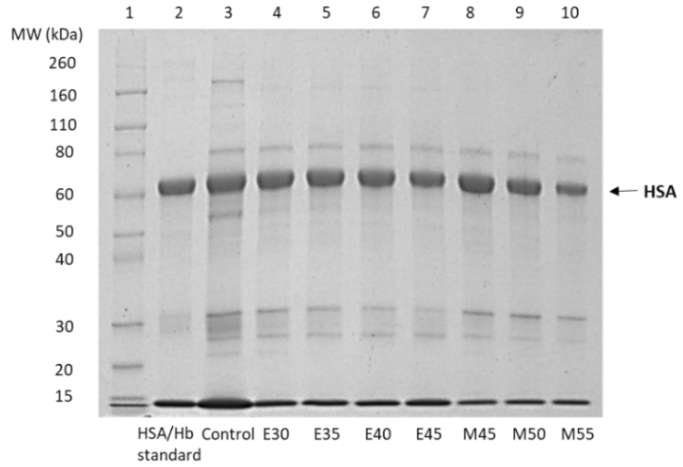


**Figure S3. 2** Isolation of HSA in DBS extracts with aqueous mixtures of ethanol (EtOH) and methanol (MeOH) using one-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis. Lanes 4-10 show the recovery of HSA after precipitation with 30% EtOH (E30), 35% EtOH (E35), 40% EtOH (E40), 45% EtOH (E45), 45% MeOH (M45), 50% MeOH (M50), 55% MeOH (M55). HSA and Hb standards were mixed together in Lane 2. The table on the right shows the percentage of HSA recovered and percentage of Hb removed after precipitation with EtOH and MeOH. DBS extracted in water was used as the control. HSA recovery was calculated from the gel by dividing band intensities of HSA after extraction with EtOH and MeOH by the band intensity of HSA in the control. Hb removal was calculated relative to the control based on the absorbance at 407 nm using UV-Vis spectroscopy.




| Sample | Hb (mg/mL) | Band intensity (HSA) | HSA recovery (%) | Hb removal (%) |
|---|---|---|---|---|
| Control | 2.37 | 3489.4 | | |
| EtOH 30% | 0.24 | 3183.7 | 91.2 | 10.2 |
| EtOH 35% | 0.20 | 3173.3 | 90.9 | 8.4 |
| EtOH 40% | 0.18 | 3166.7 | 90.8 | 7.5 |
| EtOH 45% | 0.17 | 2584.1 | 74.1 | 7.2 |
| MeOH 45% | 0.13 | 3218.7 | 92.2 | 5.4 |
| MeOH 50% | 0.10 | 2444.9 | 70.1 | 4.4 |
| MeOH 55% | 0.11 | 1494.0 | 42.8 | 4.5 |

**Figure S3. 3** Recovery of HSA relative to the total protein concentration in the DBS extract. DBS punches of increasing size were extracted to assess the effect of the total protein concentration on the recovery of HSA after precipitation with 50% methanol. (A) Extraction of DBS punches ranging from 2.0 – 5.0 mm. (B) Extraction of DBS punches ranging from 6.0 – 7.2 mm. HSA recovery before and after precipitation are shown side-by-side for each punch size in the gels. (C) Plot of HSA recovery in relation the total protein concentration before addition of methanol to precipitate proteins. HSA recovery was estimated by taking the ratio of post-precipitation band intensity at 60 kDa divided by the pre-precipitation band intensity at 60 kDa for each punch size. The plot shows a decrease in HSA recovery with increasing total protein concentration.

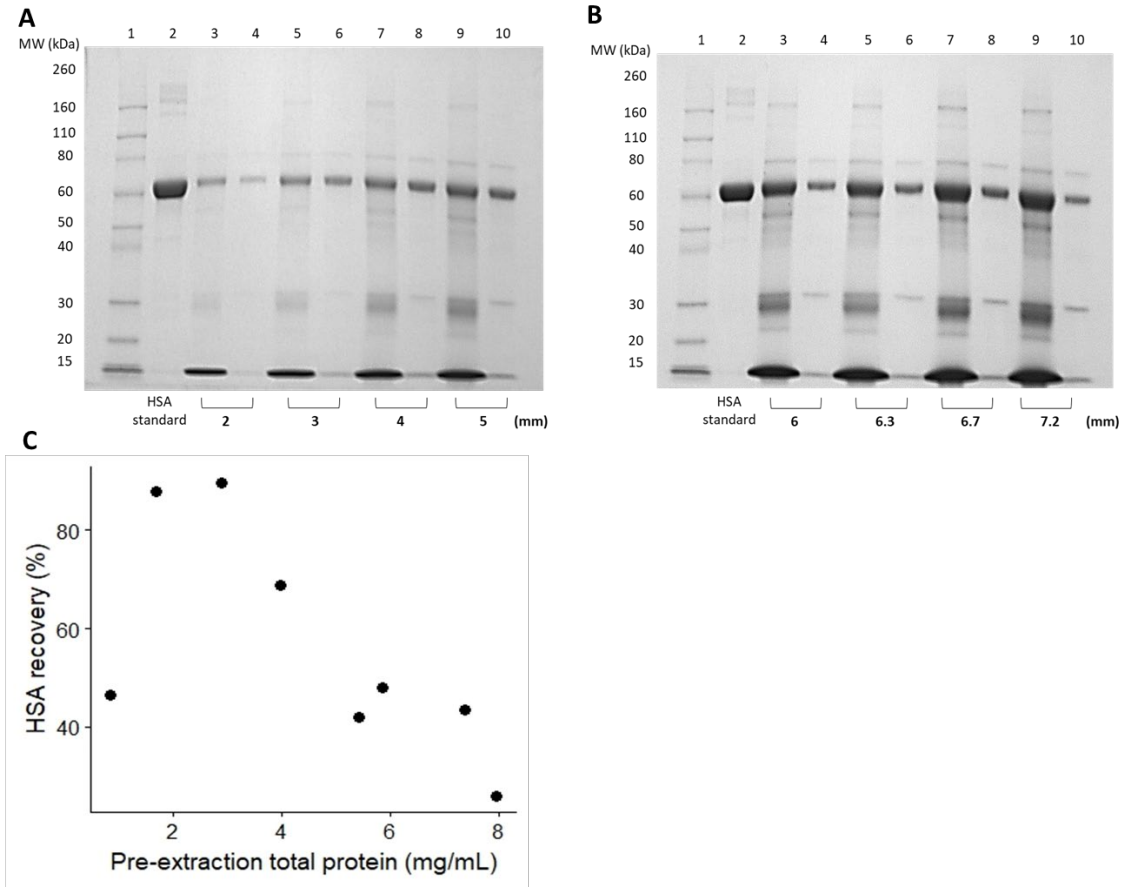
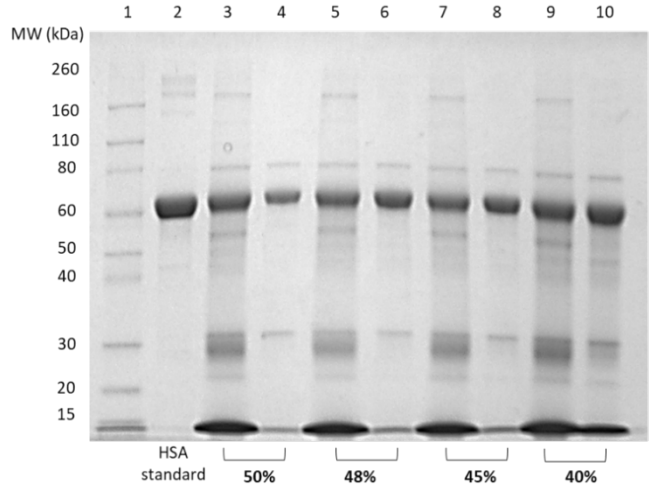
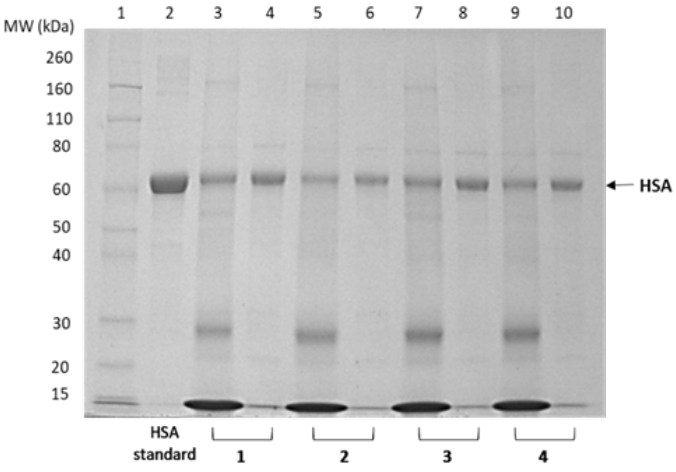


**Figure S3. 4** Recovery of HSA with increasing amounts of methanol in the extraction mixture. Gel on the left shows HSA recovery after precipitation of proteins in the DBS extracts from 6-mm punches using 40, 45, 48, and 50% methanol. HSA recovery before and after precipitation are shown side-by-side for each methanol mixture. The table on the right shows the percentage of HSA recovered after precipitation. HSA recovery was estimated by taking the ratio of post-precipitation band intensity at 60 kDa divided by the pre-precipitation band intensity at 60 kDa.




| Sample | Band intensity (HSA) | HSA recovery (%) |
|---|---|---|
| 50% pre | 4227.8 | |
| 50% post | 2795.7 | 66.1 |
| 48% pre | 4338.7 | |
| 48% post | 4105.1 | 94.6 |
| 45% pre | 3981.6 | |
| 45% post | 3856.9 | 96.9 |
| 40% pre | 4571.7 | |
| 40% post | 4806.1 | 105.1 |

**Figure S3. 5** Extraction of four 4.7-mm punches from archived newborn DBS using 45% methanol. The gel shows HSA recovery after precipitation of proteins in the DBS extracts. HSA recovery before (left lane) and after (right lane) precipitation are shown side-by-side. The table shows the percentage of Hb removed based on absorbance measurements at 407 nm (ratio of Hb concentrations post-precipitation:pre-precipitation).




| ANBS | Hb (mg/mL) | Hb removal (%) |
|---|---|---|
| 1 pre | 1.41 | |
| 1 post | 0.04 | 97.2 |
| 2 pre | 1.50 | |
| 2 post | 0.01 | 99.3 |
| 3 pre | 1.64 | |
| 3 post | 0.01 | 99.4 |
| 4 pre | 1.43 | |
| 4 post | 0.01 | 99.3 |

**Figure S3. 6** Comparison of HSA digestion between 16, 32, and 64 min digestion programs. Total ion chromatograms of HSA isolated from DBS extracts digested for 16 min (red), 32 min (green), and 64 min (blue) are shown. The largest peak at ~30 min represents undigested protein.

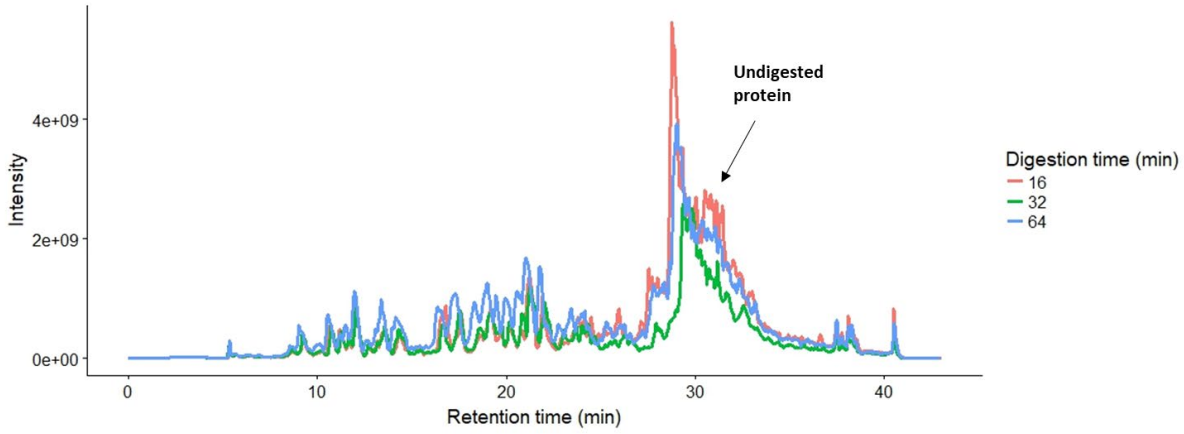


**Figure S3. 7** Digestion of HSA with increasing amounts of trypsin. (A) Gel shows HSA isolated from DBS extracts were digested with 0 – 35 $\mu$g of trypsin (enzyme to protein ratio ranging from 1:18 to 1:3). There were three main peaks associated with undigested protein (Band 1 – 3). (B) Undigested proteins in the gel were quantified using ImageJ and plotted against the amount of trypsin used.

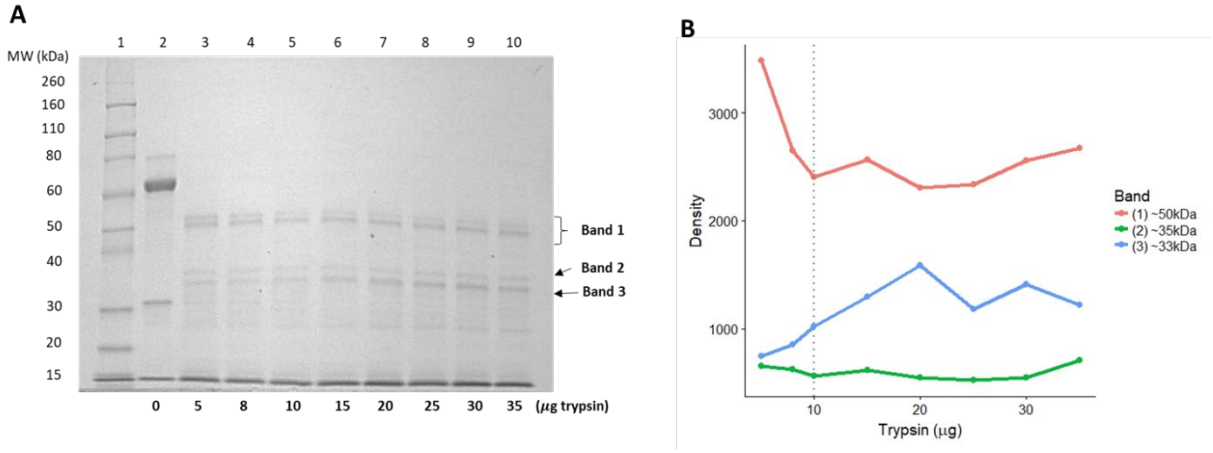



**Figure S3. 8** Correlation between Hb concentrations and the second estimated factor of unwanted variation.

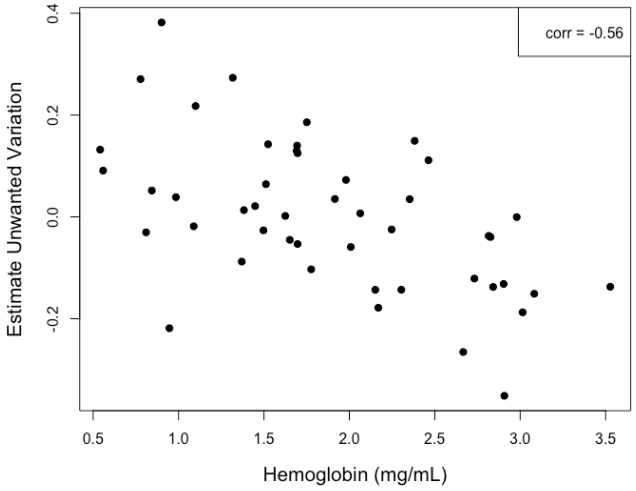


**Figure S3. 9** Relative log abundance (RLA) plot of duplicate measurements for each sample. The RLA plot was obtained by standardizing each adduct by the median abundance across samples and logging the resulting ratio. Duplicate measurements should ideally have similar distributions with a median of 0 (dotted line) and low variation around the median. The x-axis shows the sample number and the measurement number. Measurements shown in red were removed from the analysis due to high variance and poor reproducibility. Samples 30 and 41 were removed due to poor reproducibility of duplicate measurements, and the second measurement of sample 33 was removed due to high variance.

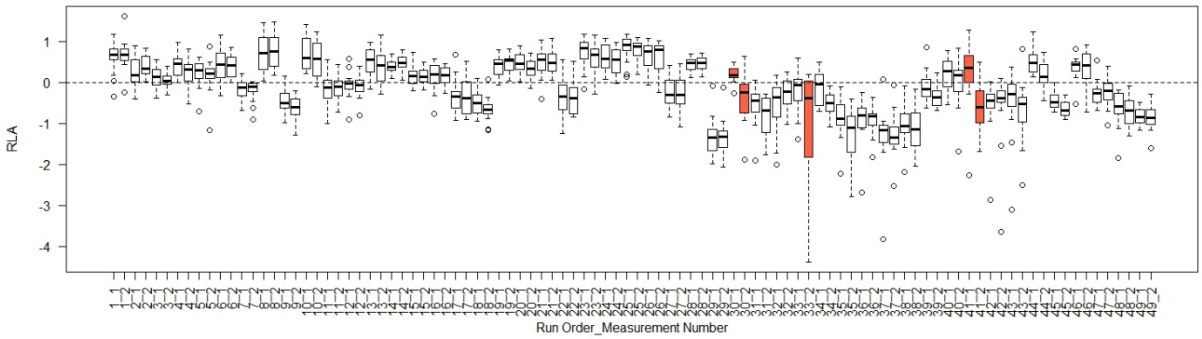



**Figure S3. 10** Percentage of samples with missing values for each adduct. Three adducts missing in over 50% (dotted line) of the samples were removed from the analysis. (MIM, monoisotopic mass).

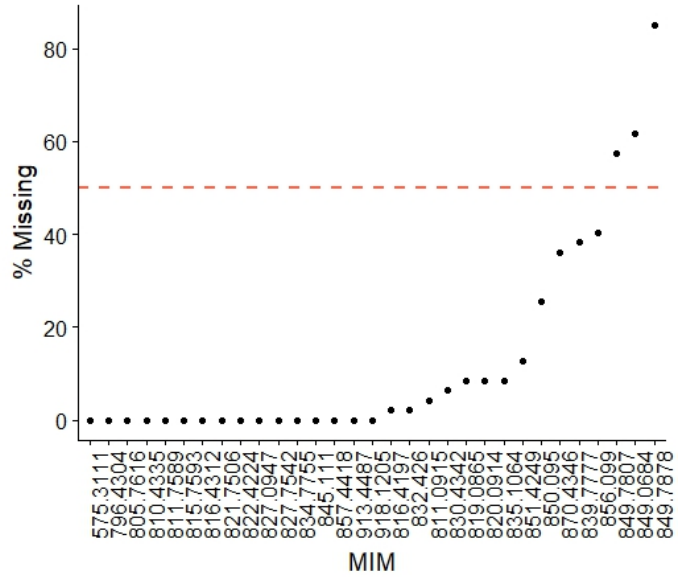
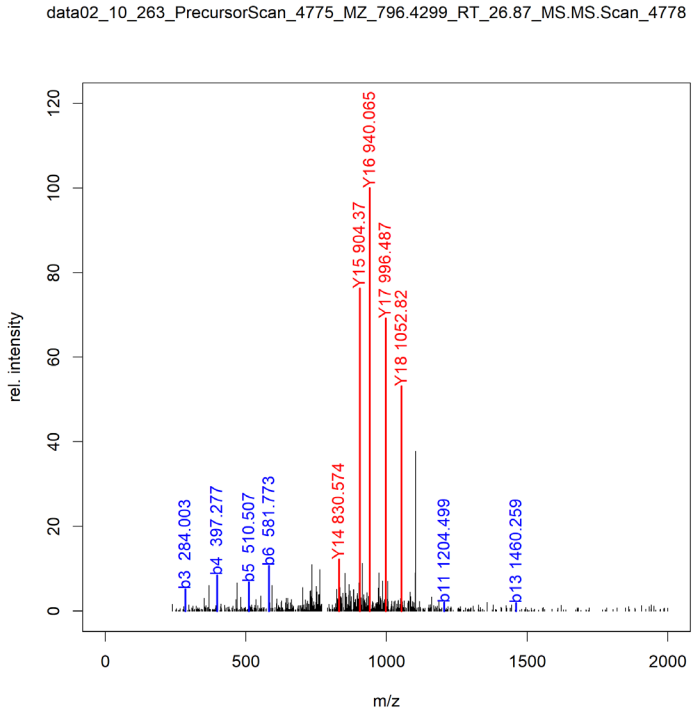
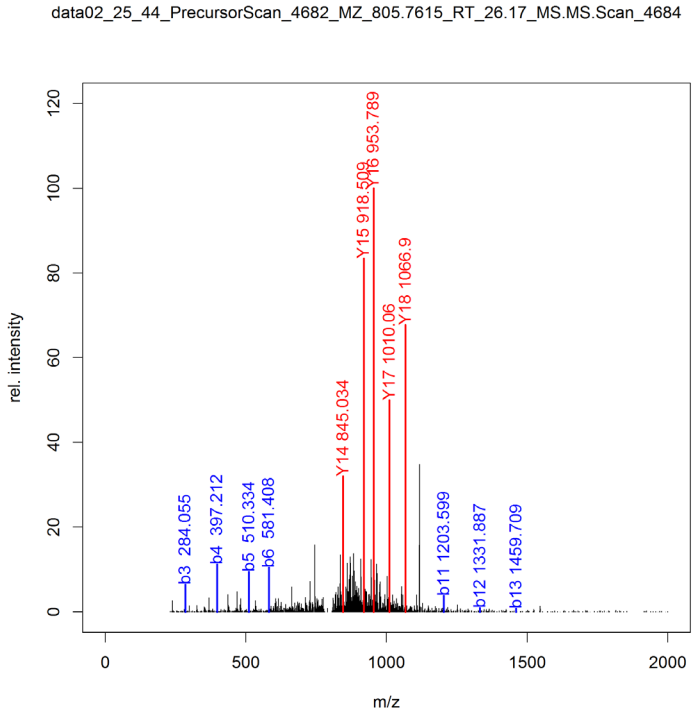


**Figure S3. 11** Representative MS2 spectra of the 26 adducts detected from archived newborn DBS.

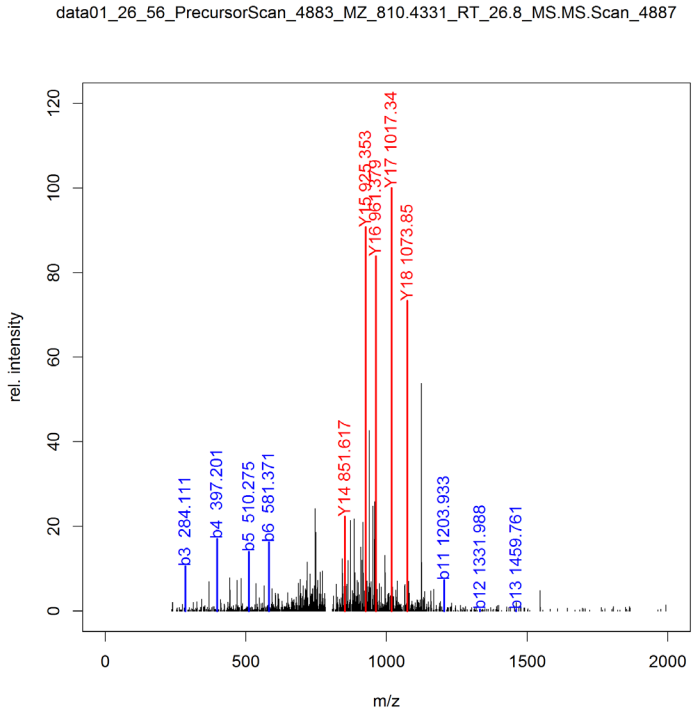
***Adduct 796.43***



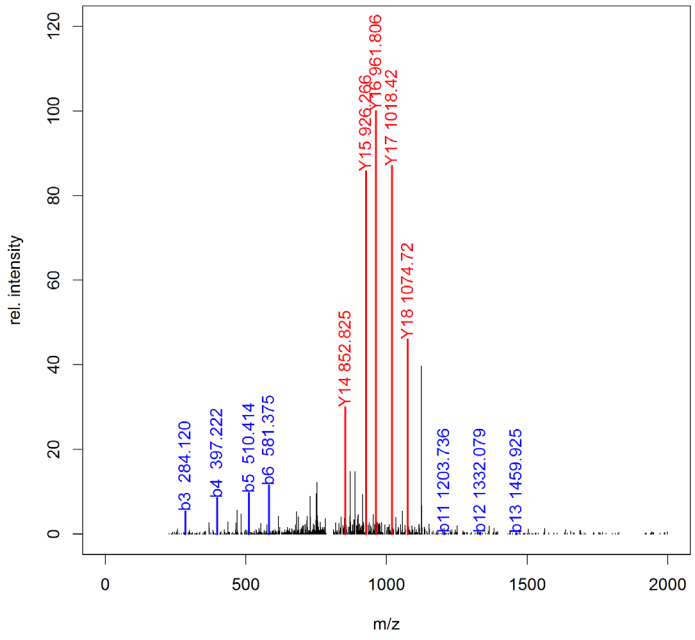

***Adduct 805.76***



***Adduct 810.43***

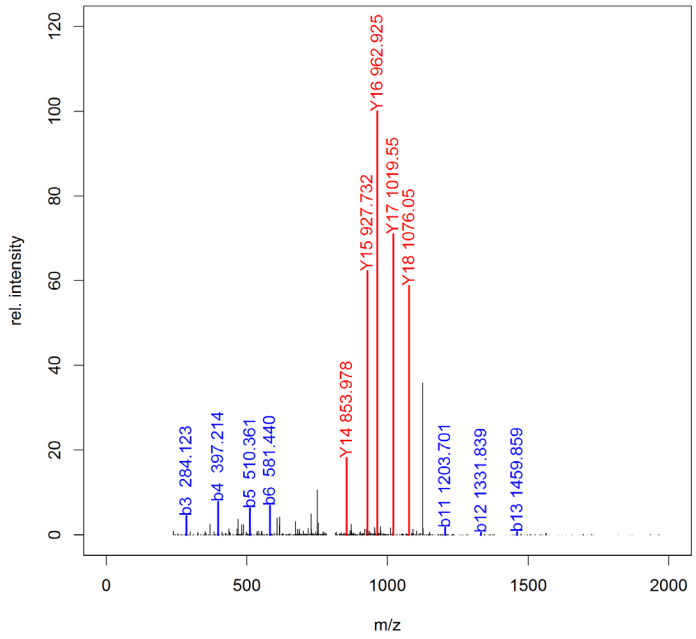



***Adduct 811.09***



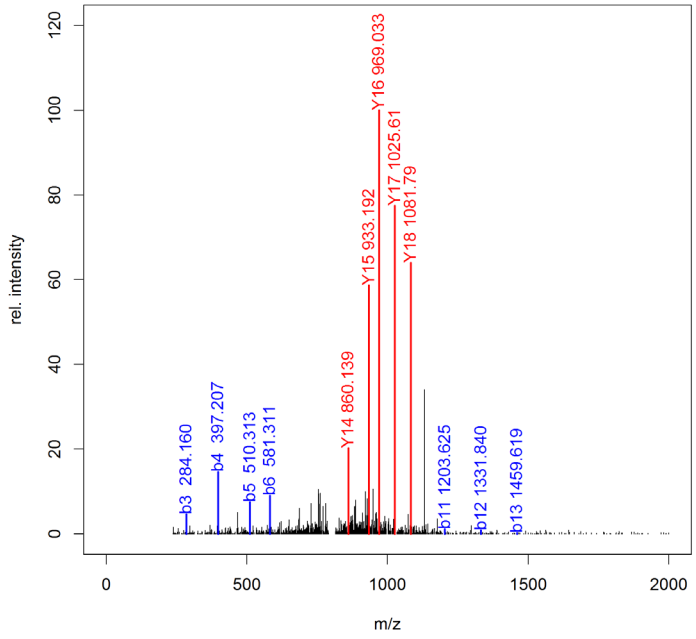
***Adduct 811.76***

data01_25_58_PrecursorScan_4693_MZ_811.7585_RT_26.83_MS.MS.Scan_4694



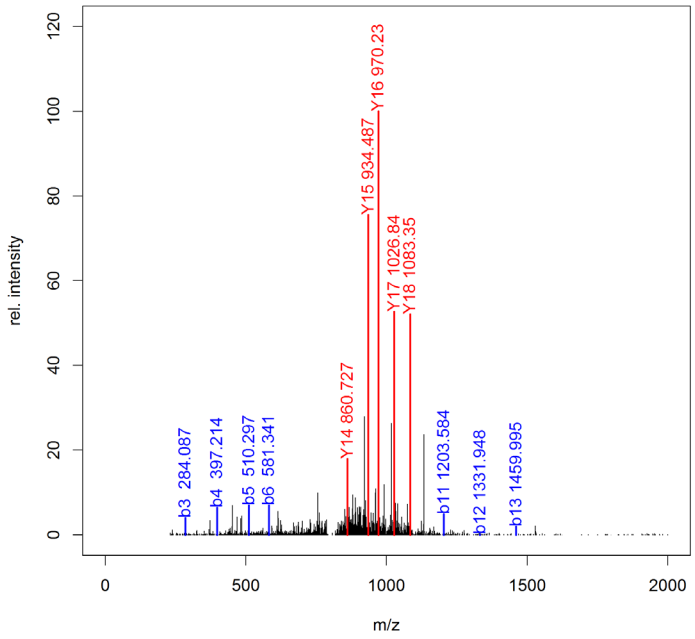

***Adduct 815.76***

data01_11_40_PrecursorScan_5015_MZ_815.7604_RT_27.35_MS.MS.Scan_5016



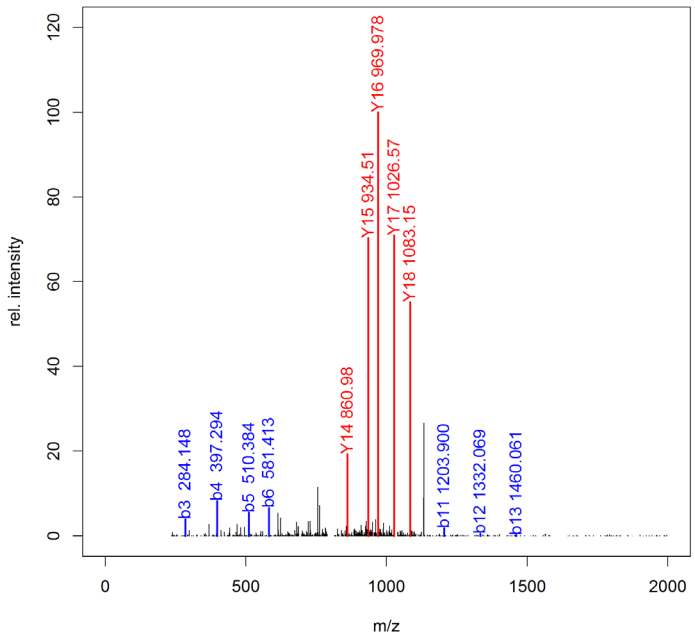
***Adduct 816.42***

data01_23_31_PrecursorScan_4419_MZ_816.4185_RT_25.13_MS.MS.Scan_4422



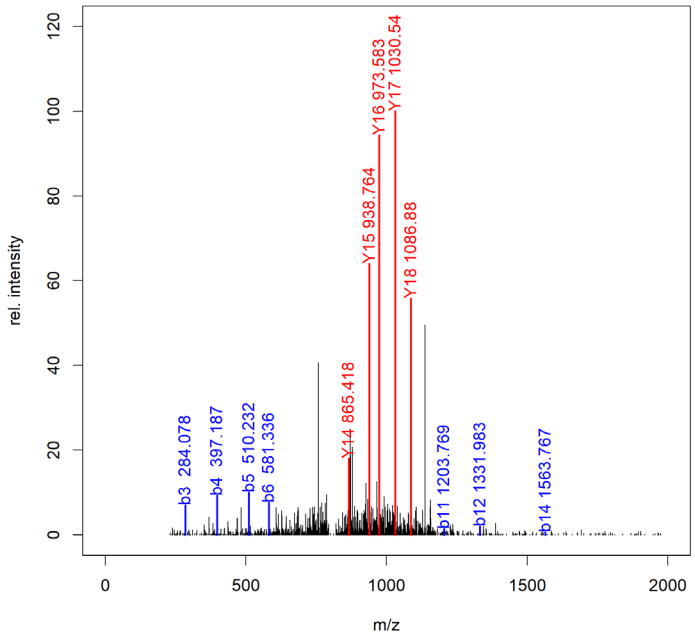

***Adduct 816.43***

data01_19_39_PrecursorScan_4578_MZ_816.4312_RT_27.59_MS.MS.Scan_4579



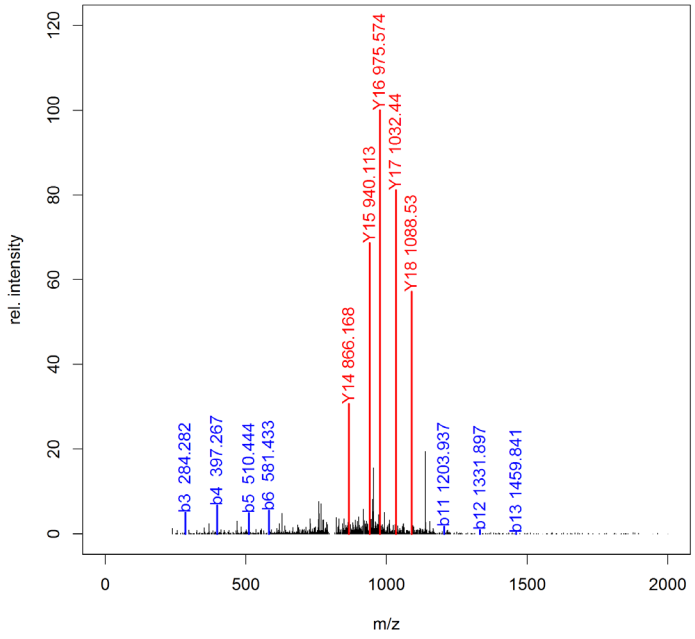
***Adduct 819.09***

ıta02_39_170727004242_245_PrecursorScan_4979_MZ_819.0856_RT_27.32_MS.MS.Scan_49



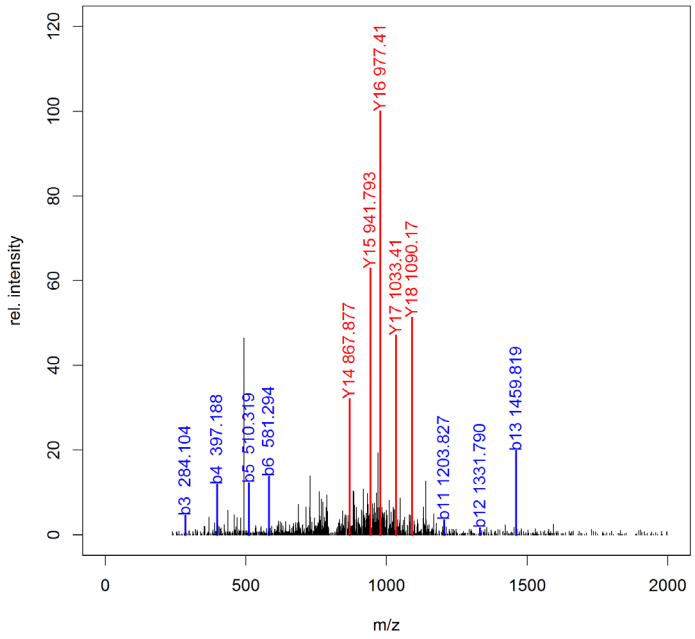

***Adduct 820.09***

data01_21_61_PrecursorScan_4951_MZ_820.0905_RT_27.35_MS.MS.Scan_4953



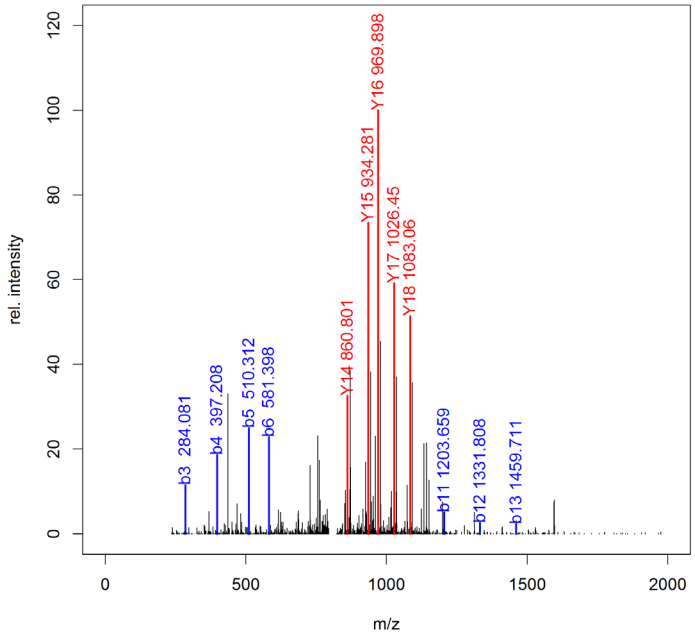
***Adduct 821.75***

data01_47_20_PrecursorScan_4242_MZ_821.7497_RT_27.55_MS.MS.Scan_4244



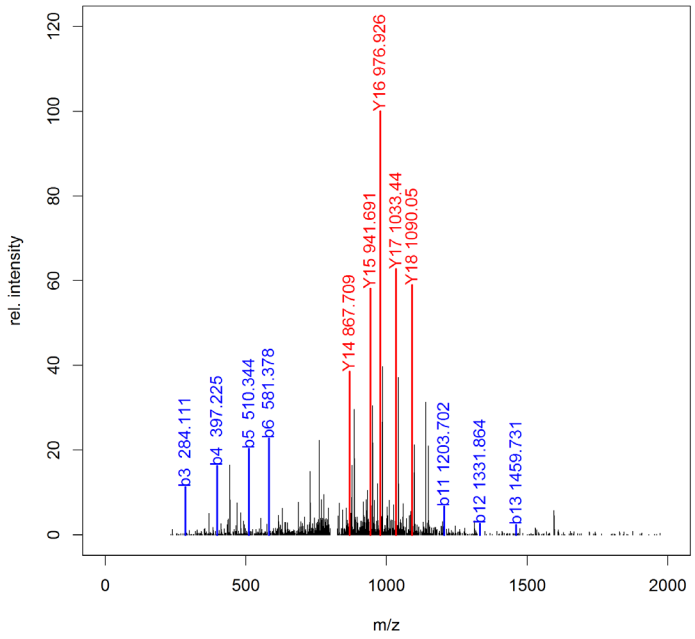

***Adduct 822.42***

data01_1_90_PrecursorScan_5493_MZ_822.4216_RT_26.94_MS.MS.Scan_5494



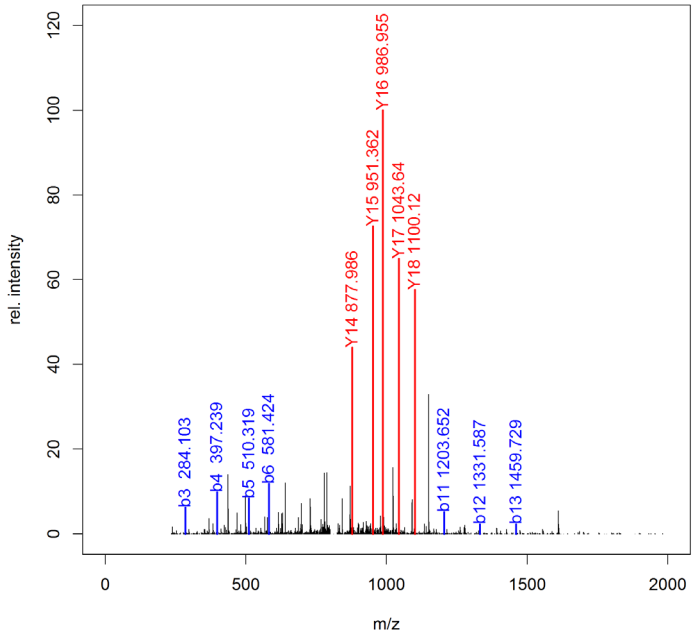
***Adduct 827.09***

data01_6_54_PrecursorScan_4725_MZ_827.0957_RT_26.95_MS.MS.Scan_4726



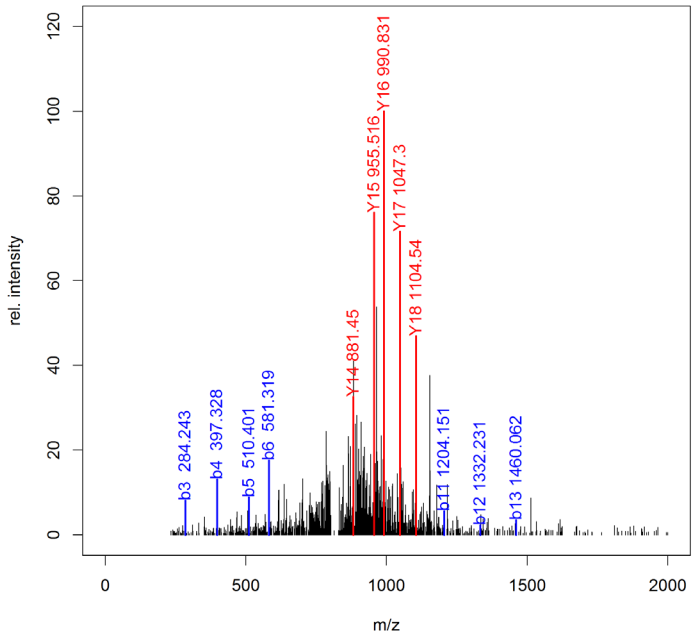

***Adduct 827.75***

data01_16_42_PrecursorScan_4654_MZ_827.7546_RT_27.34_MS.MS.Scan_4655



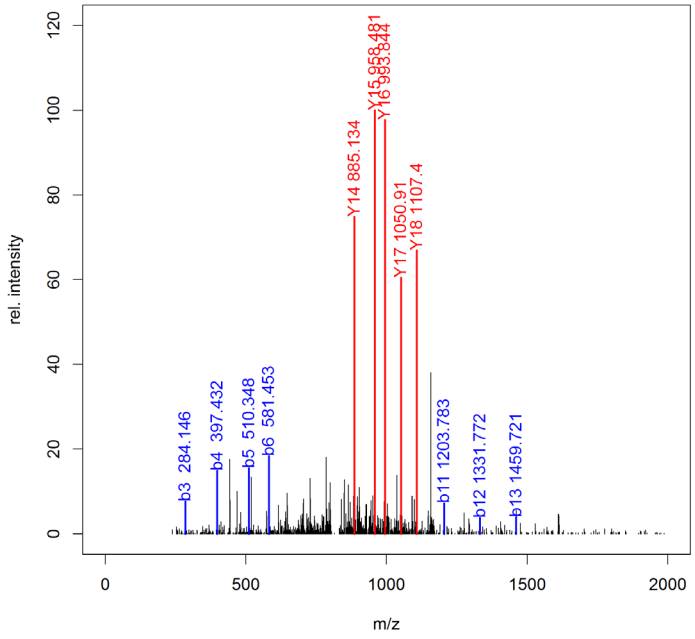
***Adduct 830.43***

data02_21_51_PrecursorScan_4579_MZ_830.4344_RT_27.04_MS.MS.Scan_4583



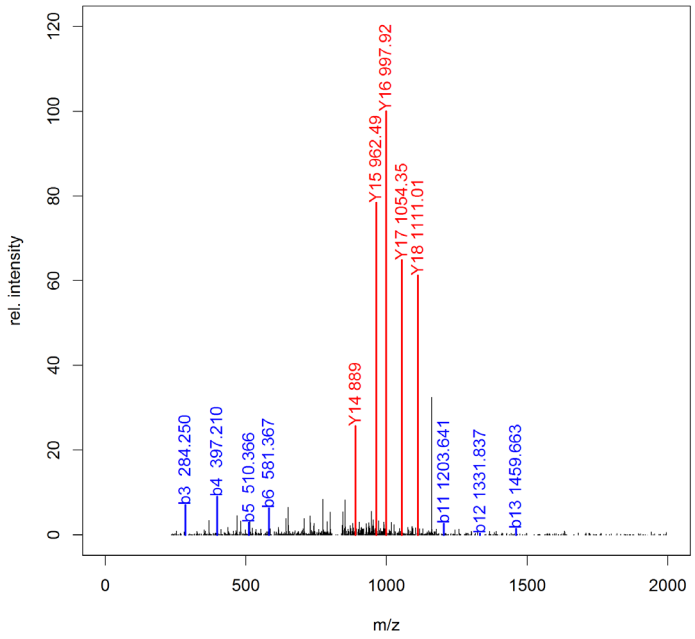

***Adduct 832.43***

data01_24_54_PrecursorScan_4691_MZ_832.4244_RT_27.32_MS.MS.Scan_4692



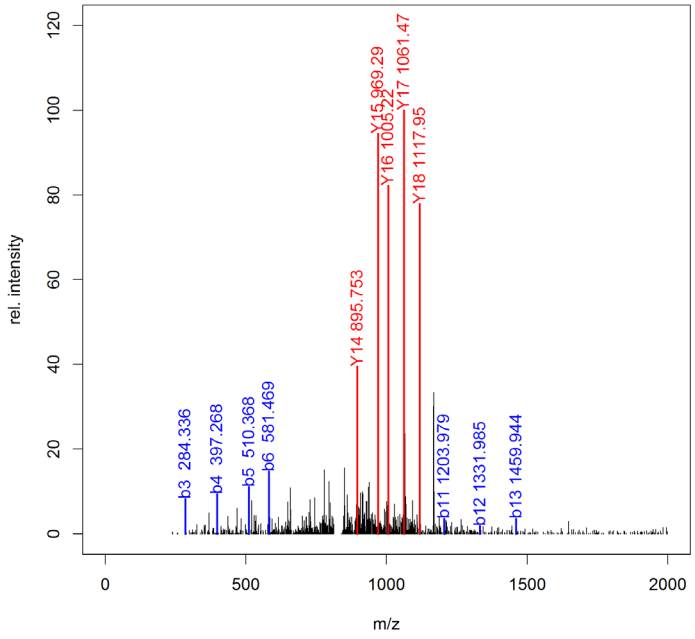
***Adduct 835.11***

data02_10_70_PrecursorScan_4885_MZ_835.1063_RT_27.45_MS.MS.Scan_4886



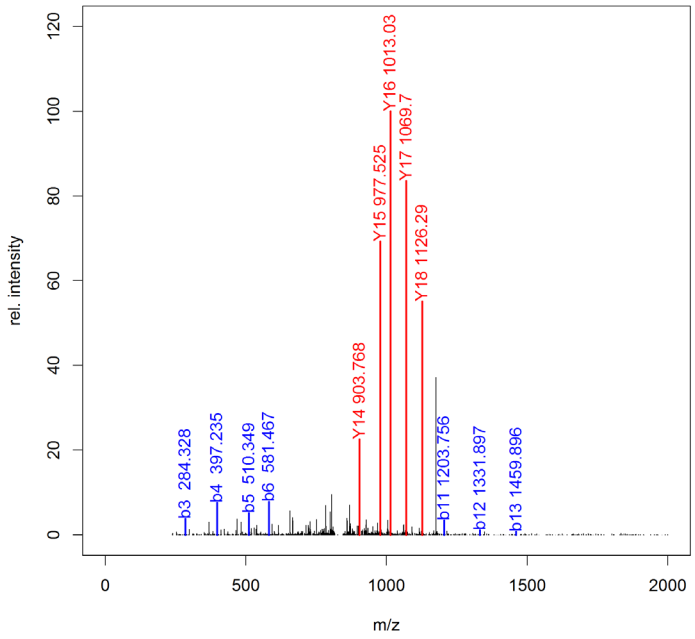

***Adduct 839.78***

data01_39_33_PrecursorScan_4108_MZ_839.7736_RT_27.53_MS.MS.Scan_4109



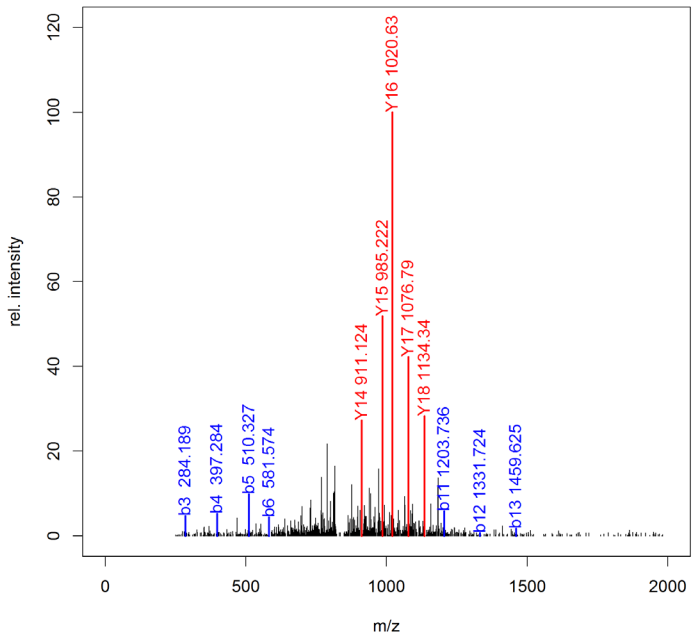
***Adduct 845.11***

data02_17_26_PrecursorScan_4323_MZ_845.1103_RT_27.54_MS.MS.Scan_4324



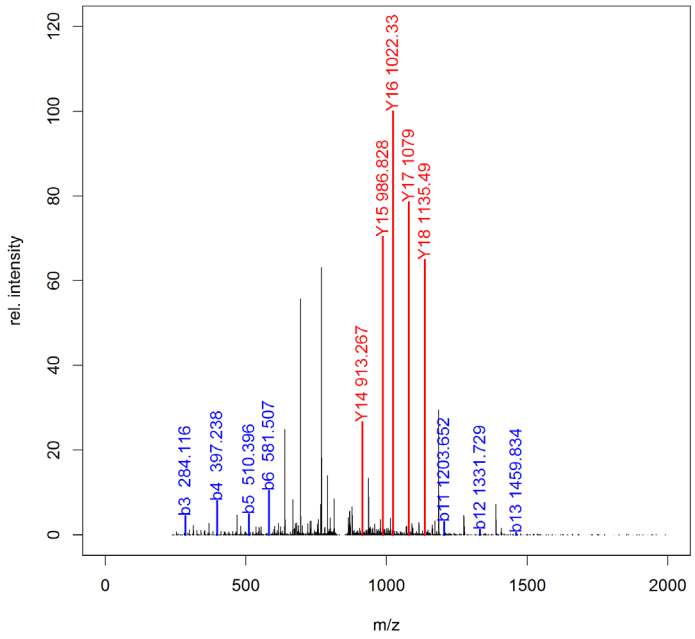

***Adduct 850.10***

data01_22_24_PrecursorScan_4596_MZ_850.0957_RT_28.14_MS.MS.Scan_4598



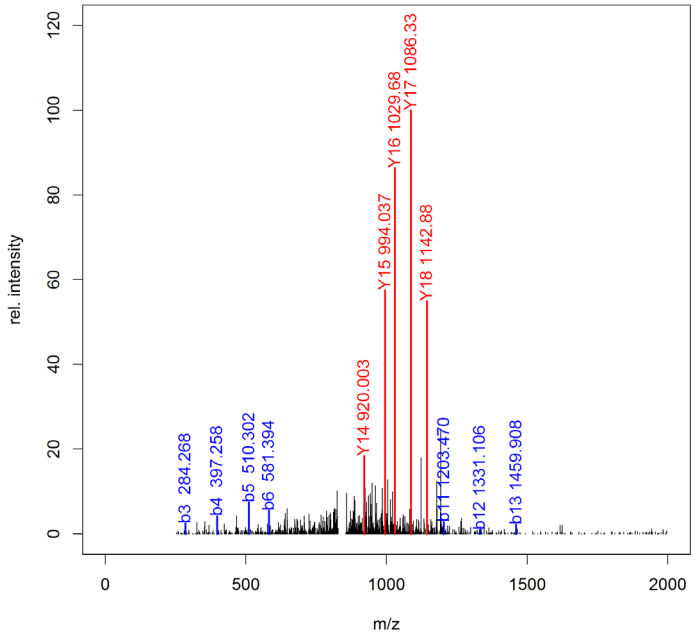
***Adduct 851.42***

ata02_44_170724164505_22_PrecursorScan_4357_MZ_851.4245_RT_26.19_MS.MS.Scan_436



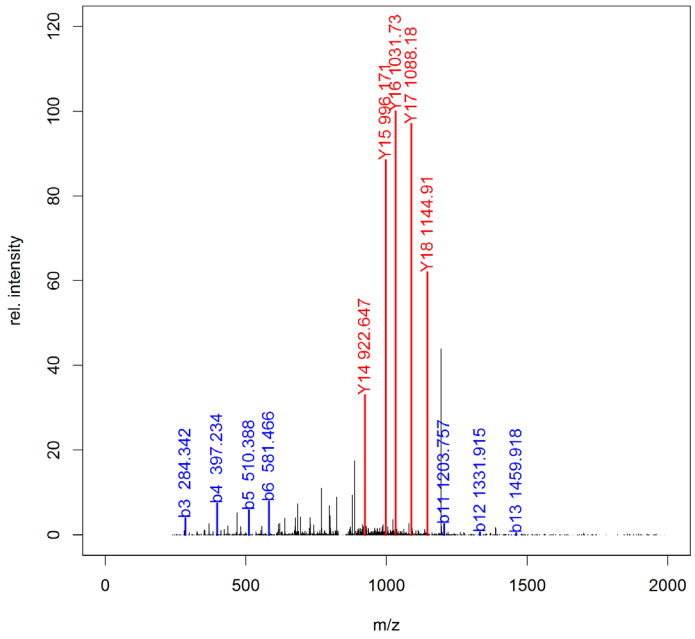

***Adduct 856.10***

data01_27_28_PrecursorScan_4181_MZ_856.0991_RT_26.56_MS.MS.Scan_4183



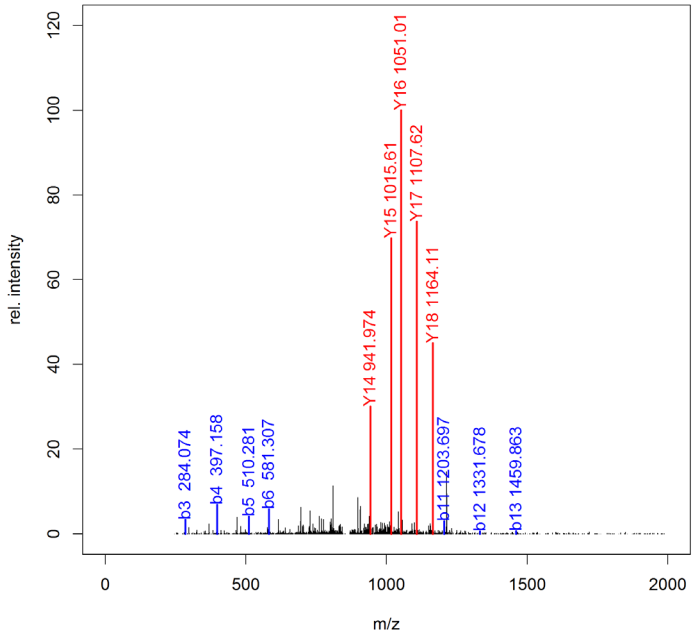
***Adduct 857.44***

data02_23_58_PrecursorScan_4714_MZ_857.4417_RT_26.77_MS.MS.Scan_4715



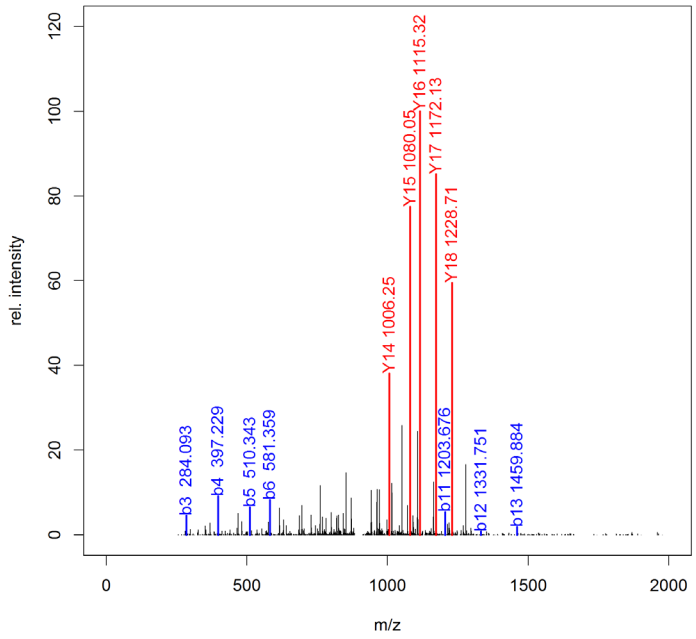

***Adduct 870.43***

data02_47_19_PrecursorScan_3829_MZ_870.4338_RT_25.82_MS.MS.Scan_3830



***Adduct 913.45***

data02_12_34_PrecursorScan_4375_MZ_913.4484_RT_26.13_MS.MS.Scan_4376




***Adduct 918.12***

data01_27_27_PrecursorScan_4181_MZ_918.1204_RT_26.56_MS.MS.Scan_4182

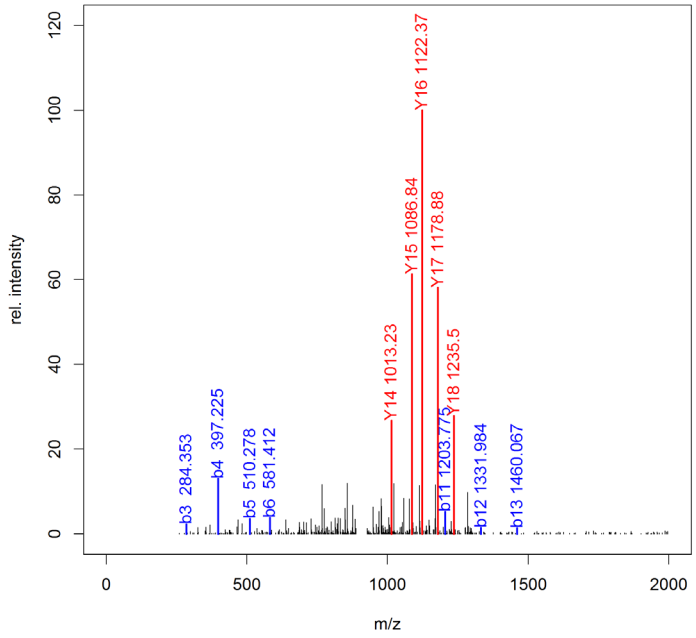
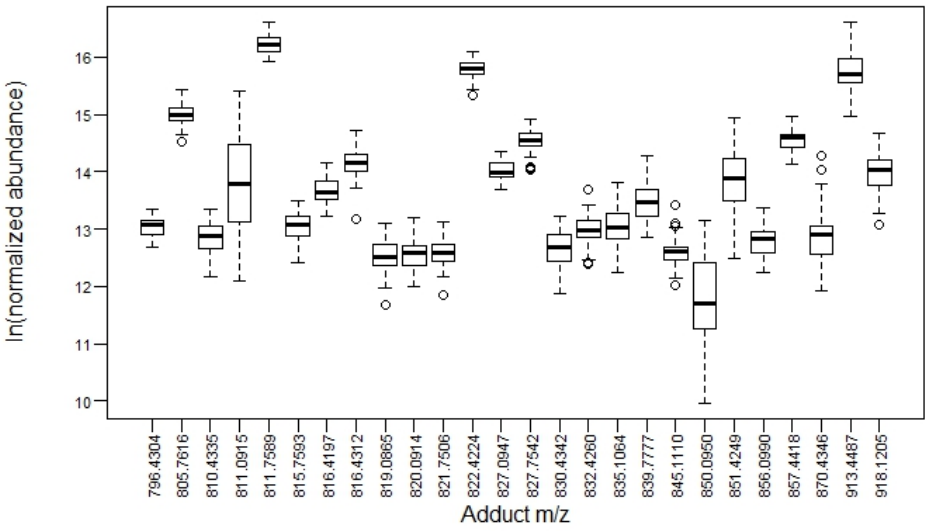


**Figure S3. 12** Box plot of adduct abundances across all samples. The x-axis shows the observed adduct *m/z*, and the y-axis shows the logged, normalized abundances.



### 3.9.3 Supplemental Table

**Table S3. 1** Summary characteristics of the 47 subjects included in the analysis.

| | Smoker | Nonsmoker |
|---|---|---|
| **Number of samples** | 23 | 24 |
| **Child's race** | | |
| White | 18 | 17 |
| Other | 5 | 7 |
| **Child's sex** | | |
| Male | 13 | 14 |
| Female | 10 | 10 |
| **Child's birth year** | | |
| Mean (SD) | 1995 (5) | 1996 (4) |
| Min | 1985 | 1985 |
| Max | 2003 | 2003 |
| **Hb (mg/mL)** | | |
| Mean (SD) | 1.80 (0.72) | 1.97 (0.80) |
| Min | 0.56 | 0.54 |
| Max | 3.08 | 3.53 |
| **DBS weight (mg)** | | |
| Mean (SD) | 6.64 (2.16) | 6.62 (2.08) |
| Min | 2.64 | 2.56 |
| Max | 11.07 | 9.45 |

### 3.9.4 Supplemental References


(1) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. Interferences and Contaminants Encountered in Modern Mass Spectrometry. *Anal. Chim. Acta* **2008**, *627* (1), 71–81.

(2) Capiau, S.; Wilk, L. S.; Aalders, M. C. G.; Stove, C. P. A Novel, Nondestructive, Dried Blood Spot-Based Hematocrit Prediction Method Using Noncontact Diffuse Reflectance Spectroscopy. *Anal. Chem.* **2016**, *88* (12), 6538–6546.

(3) Bremmer, R. H.; Nadort, A.; van Leeuwen, T. G.; van Gemert, M. J. C.; Aalders, M. C. G. Age Estimation of Blood Stains by Hemoglobin Derivative Determination Using Reflectance Spectroscopy. *Forensic Sci. Int.* **2011**, *206* (1–3), 166–171.

(4) Kerti, A.; Morlin, Z. M.; Pajor, F.; Bardos, L. Determination of Hemoglobin Content in Whole Blood and in Dried Blood Spots in Lambs. *Anim. Welfare, Ethol. Hous. Syst.* **2013**, 513–516.

(5) Miller IV, J. H.; Poston, P. A.; Rutan, S. C.; Karnes H, T. An On-Card Approach for Assessment of Hematocrit on Dried Blood Spots Which Allows for Correction of Sample Volume. *J. Anal. Bioanal. Tech.* **2013**, *04* (01), 1–8.

(6) Meng, F.; Alayash, A. I. Determination of Extinction Coefficients of Human Hemoglobin in Various Redox States. *Anal. Biochem.* **2017**, *521*, 11–19.

(7) Bremmer, R. H.; de Bruin, D. M.; de Joode, M.; Buma, W. J.; van Leeuwen, T. G.; Aalders, M. C. G. Biphasic Oxidation of Oxy-Hemoglobin in Bloodstains. *PLoS One* **2011**, *6* (7), e21845.

(8) Zijlstra, W. G.; Buursma, A.; Meeuwsen-van der Roest, W. P. Absorption Spectra of Human Fetal and Adult Oxyhemoglobin, De-Oxyhemoglobin, and Methemoglobin. *Clin. Chem.* **1991**, *37* (9), 1633–1638.

# Chapter 4.

# Untargeted Adductomics of Newborn Dried Blood Spots to Identify Cys34 Modifications to Human Serum Albumin Associated with Childhood Leukemia

Yukiko Yano,[1] Courtney Schiffman,[2] Josie Hayes,[1] Hasmik Grigoryan,[1] William Edmands,[1] Lauren Petrick,[3] Todd Whitehead,[4,5] Catherine Metayer,[4,5] Sandrine Dudoit,[2] and Stephen Rappaport[1,5]

1. Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA 94720 USA
2. Division of Biostatistics, School of Public Health, University of California, Berkeley, CA 94720 USA
3. The Senator Frank R. Lautenberg Environmental Health Sciences Laboratory, Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029 USA
4. Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720 USA
5. Center for Integrative Research on Childhood Leukemia and the Environment, University of California, Berkeley, CA 94720 USA

## 4.1 Abstract

Early life exposures, including those occurring *in utero*, are likely to play an important role in the etiology of childhood leukemia. However, few risk factors have been established, and the underlying disease mechanisms remain elusive. The developing fetus is exposed to various chemicals, originating from both endogenous (*e.g.*, maternal, fetal, microbial metabolism) and exogenous sources (*e.g.*, maternal exposures from the diet, xenobiotics, and lifestyle factors such as smoking and alcohol intake). Metabolism of these chemicals generate electrophiles that react with nucleophilic sites in circulating proteins, particularly Cys34 of human serum albumin (HSA), to produce adducts. To discover potential risk factors for childhood leukemia resulting from *in utero* exposures, we used untargeted adductomics to measure HSA-Cys34 adducts in 783 archived newborn dried blood spots (DBS) collected from incident cases of childhood acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) and matched population-based controls. A total of 28 HSA-Cys34 adducts were measured, including Cys34 adducts with reactive oxygen and carbonyl species, mixed disulfides with low-molecular-weight thiols (*e.g.*, cysteine, homocysteine, glutathione, etc.), and other modifications. After data preprocessing and normalization to adjust for unwanted technical variation, an ensemble of variable selection methods including both linear and nonlinear models was used to identify Cys34 adducts discriminating between cases and controls. Although we found no differences in Cys34 adduct abundances between childhood leukemia cases and controls overall, particular subtypes (*i.e.*, B-

cell ALL with t(12;21) and T-cell ALL) had higher abundances of adducts of reactive carbonyl species, suggestive of oxidative stress and lipid peroxidation as potentially etiologic factors. In addition, a Cys34 adduct of homocysteine (with loss of $H_2O$), showed consistent discrimination between AML cases and controls with a fold change (cases/controls) of 0.66. Since homocysteine is an important intermediate in folate-methionine metabolism, this may point to alterations in one-carbon metabolism and epigenetic changes as predictors of AML. Discriminating adducts warrant replication in future studies with larger sample sizes of specific subtypes, particularly AML and T-cell ALL.

## 4.2 Introduction

Acute leukemia is the most common type of cancer among children under the age of 15, representing about one-third of all cases.[1] Childhood leukemia is a biologically heterogeneous disease, which includes subtypes defined by cell lineage and cytogenetic characteristics, such as chromosome translocations and changes in chromosome number.[2] The two main types of childhood leukemia are acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). ALL is more common and accounts for approximately 80% of childhood leukemia cases, while AML comprises 15% of cases.[3] In the United States (U.S.), incidence rates have been steadily increasing by about 1% per year, and approximately 3,800 children are diagnosed with ALL or AML each year.[3] Prenatal and early-life exposures are likely to play important roles in the etiology of childhood leukemia.[1,4] The growing incidence, particularly among affluent countries, such as the U.S. and Europe,[5] suggest that exposures may be contributing to this upwards trend.[1,6] However, the only established causal exposure for childhood leukemia is ionizing radiation,[4,6] and the underlying disease mechanisms remain elusive.

It has been confirmed that the majority of childhood leukemia cases originate *in utero*.[7] This was shown by backtracking leukemic translocations in archived newborn dried blood spots (DBS) or stored cord blood of  leukemia cases up to 14 years before the onset of disease.[8,9] In fact, it has been shown that about 1% of the normal population harbors preleukemic clones, and that about 1% of these eventually develop overt leukemia.[10] Thus, the etiology of childhood leukemia appears to follow a two-hit model, where chromosome translocations occurring *in utero* initiate leukemogenesis, but additional postnatal genetic or epigenetic changes are required to lead to overt leukemia.[11] Furthermore, it is possible that exposures occurring during fetal development contribute to initial genetic aberrations involved in the first hit, and these can increase the risk of developing childhood leukemia later in life.

The fetus receives chemical exposures from both endogenous (*e.g.*, maternal, fetal, microbial metabolism) and exogenous sources (*e.g.*, maternal exposures from the diet, xenobiotics, and lifestyle factors including smoking and alcohol consumption) during gestation. Metabolism of these chemicals generate electrophiles that react with nucleophilic sites in circulating proteins to produce adducts. Although these electrophilic species are too reactive to measure in blood, their adducts with circulating proteins can be used to investigate the corresponding exposures. Since the sulfhydryl group at Cys34 of human serum albumin (HSA) is a particularly powerful scavenger of reactive electrophiles in the interstitial space,[12] we developed an untargeted adductomics method to characterize Cys34 modifications in archived newborn DBS.[13] Our adductomics assay uses nanoflow liquid chromatography-high resolution mass spectrometry (nLC-HRMS) to detect and quantitate Cys34 adducts.[13] Because HSA has a

residence time of 28 days,[14] Cys34 adducts in archived DBS capture systemic exposures occurring during the last month of gestation.

Our DBS-adductomics methodology was validated with 49 archived DBS collected from newborns whose mothers either actively smoked during pregnancy or were nonsmokers.[13] The sample workflow included extraction of proteins from DBS, measurement of hemoglobin to normalize for blood volume, addition of methanol to enrich HSA, digestion with trypsin, and detection of HSA-Cys34 adducts via nLC-HRMS. Data normalization was performed to remove unwanted technical variation arising from HSA digestion, blood volume, DBS age, mass spectrometry, and batch effects. Twenty-six Cys34 adducts were detected, which primarily consisted of Cys34 oxidation products and mixed disulfides with low-molecular-weight thiols. Using an ensemble of variable selection methods, including both linear and nonlinear models, the Cys34 adduct of cyanide was found to consistently discriminate between newborns of smoking and nonsmoking mothers with a fold change (smoking/nonsmoking) of 1.31 and a cross-validated area under the estimated receiver operating characteristic curve of 0.79. Indeed, hydrogen cyanide is a component of cigarette smoke, and these results indicated that DBS-based adductomics is suitable for investigating *in utero* exposures to reactive electrophiles that may influence disease risks later in life.

Here, we performed an untargeted investigation of HSA-Cys34 adducts in archived newborn DBS collected from 783 participants of a population-based case-control study to discover potential risk factors for childhood leukemia resulting from *in utero* exposures. After data preprocessing and normalization an ensemble of variable selection methods was used to identify adducts that discriminated between cases of ALL and AML and matched controls.

## 4.3 Materials and methods

### *Chemicals and Reagents*

Acetonitrile (Ultra Chromasolv, LCMS grade), triethylammonium bicarbonate (TEAB) buffer (1 M), ethylenediamine-tetraacetic acid (EDTA, anhydrous), and porcine trypsin were from Sigma-Aldrich (St. Louis, MO). Methanol (Optima, LCMS grade), formic acid (Optima, LCMS grade), and iodoacetamide (IAA) were from Fisher Scientific (Pittsburgh, PA). Purified human hemoglobin was from MP Biomedicals, LLC (Santa Ana, CA). Isotopically labeled T3 (iT3) with sequence AL-[$^{15}$N,$^{13}$C-Val]-LIAFAQYLQQCPFEDH-[$^{15}$N,$^{13}$C-Val]-K was custom-made (>95%, BioMer Technology, Pleasanton, CA), and the carbamidomethylated iT3 (IAA-iT3)[15] was used as an internal standard to monitor retention time and mass drifts, as well as drift in instrument performance. Water was prepared with a PureLab purification system (18.2 mΩ cm resistivity at 25 ℃; Elga LabWater, Woodridge, IL).

### *Study Subjects and Specimens*

Participants of the California Childhood Leukemia Study (CCLS), a previously described population-based case-control study,[16] were selected with informed consent obtained from parents of the children. Incident cases of newly diagnosed ALL and AML among children under age 15 were ascertained from major clinical centers in northern and central California. Controls were randomly selected from birth certificates obtained through the California Office of Vital Records. Cases and controls were individually matched on child's month and year of birth, sex, Hispanic ethnicity based on birth certificates, and maternal race.[16] Parents were interviewed to collect information on socio-demographic characteristics.

Archived newborn DBS were retrospectively obtained from the California Department of Public Health, Genetic Diseases Screening Branch.[17] The DBS were collected at birth and had been archived at -20 ℃ for 12 to 33 years prior to the present investigation. We used single, 4.7-mm DBS punches (equivalent to 5 – 8 $\mu$L of whole blood) from a total of 783 subjects; including 387 cases (of which 339 were ALL cases and 45 were AML cases) and 396 controls. Figure 4.1 shows the number of cases and controls for each childhood leukemia subtype.

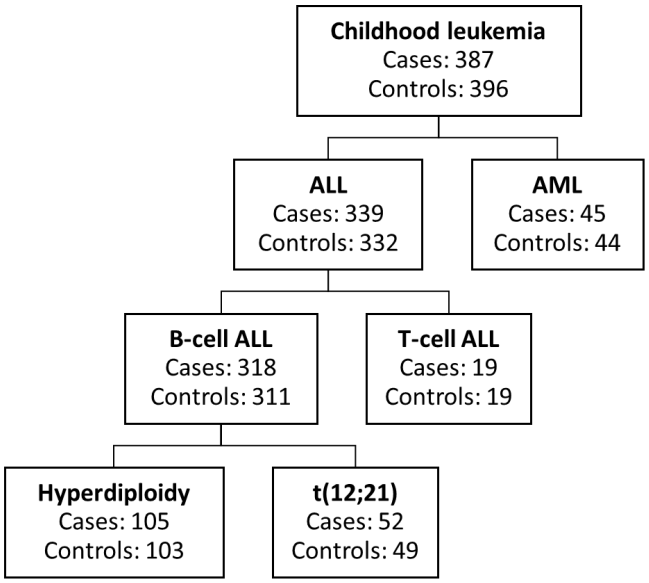



**Figure 4. 1** Number of cases and controls in each subtype of childhood leukemia of the 783 subjects included in the analysis.

### *Sample Preparation and nLC-HRMS analysis*

For each subject, one 4.7-mm DBS punch was analyzed to detect and quantify HSA-Cys34 adducts as described previously.[13] Briefly, proteins were extracted from DBS, hemoglobin (Hb) concentrations were measured in the DBS extracts to normalize for blood volume using UV-Visible absorption spectroscopy. Methanol was added to enrich HSA by precipitating Hb and other interfering proteins. Upon digestion with trypsin and pressure cycling, the digests were analyzed by an Orbitrap Elite HRMS coupled to a Dionex Ultimate 3000 nanoflow LC system with a nanoelectrospray ionization source (Thermo Fisher Scientific, Waltham, MA). The MS was operated in data-dependent mode, and tandem MS (MS2) were acquired in the linear ion trap. Samples were analyzed in four batches of ~200 samples, and duplicate injections were made for each sample.

### *Adduct Identification, Quantification, and Annotation*

HSA-Cys34 adducts were identified, quantified, and annotated as described previously in detail.[18] Briefly, Cys34 adducts were pinpointed from the MS2 spectra as modifications to the third largest tryptic peptide (T3) with the sequence $^{21}$ALVLIAFAQYLQQC$^{34}$PFEDHVK$^{41}$ (*m/z* = 811.7594), which contains the Cys34 site of modification. Triply charged precursor ions with *m/z* = 811.7594 $\pm$ $\Delta$ *m/z*, where $\Delta$ *m/z* represents the mass added to the T3-thiolate ion (Cys34-

$S^-$), were identified as putative Cys34 adducts. The tryptic peptide adjacent to T3 with sequence $^{42}$LVNEVTEFAK$^{51}$ (doubly charged precursor ion at $m/z$ = 575.3111) was used as a housekeeping peptide to adjust for the amount of digested HSA in each sample. An isotopically labeled and carbamidomethylated T3 peptide (IAA-iT3)[15] was used as an internal standard to monitor retention time, mass drifts, and instrument performance. Adduct abundances were obtained by peak picking and integration using the Xcalibur Processing Method (version 3.0, Thermo Fisher Scientific, Waltham, MA) based on the average MIMs (5 ppm mass accuracy) and RTs of the putative adducts, as described previously.[13] Putative adducts were annotated with database searches based on the proposed empirical formula.[13] Mass accuracies were estimated by differences between theoretical and observed MIMs.

***Data Preprocessing and Normalization***

All statistical analyses were performed using the R statistical programming environment.[19] Adduct abundances were log-transformed and the means of duplicate measurements were calculated. Missing values were imputed using *k*-nearest neighbor imputation (see Supplemental Methods in Supporting Information for details).[20] Data normalization was optimized with the Bioconductor R package 'scone'[20,21] to select an appropriate scheme for removing unwanted variation ('scone' explores different scaling methods and uses regression models to remove unwanted technical variation while preserving variation due to case-control status). The optimal normalization scheme used DESeq scaling and adjusted for the following sources of unwanted variation: batch effects, instrument performance, digested HSA, DBS age, and blood volume. Here, 'batch effects' represents the four batches of samples analyzed at different times and 'instrument performance' was indicated by the abundance of the internal standard in each sample. As mentioned earlier, 'digested HSA' was quantified by the abundance of the housekeeping peptide in each sample. 'DBS age' was calculated as 2018 – child's birth year to account for differences in the extraction efficiency of analytes due to the age of the DBS.[22] Variations in 'blood volume' across samples were adjusted based on the Hb measurements of the DBS extracts.[13]

***Variable Selection: Identification of Discriminating Adducts***

Variable selection was performed to identify adducts that discriminated childhood leukemia cases from matched controls overall and separately for ALL and AML. For ALL, we performed additional analyses stratified by the major subgroups, which included B-cell ALL, B-cell ALL with high-hyperdiploidy (51-67 chromosomes), B-cell ALL with t(12;21) chromosome translocation, and T-cell ALL (Figure 4.1).

To ensure robust associations between adduct abundances and case-control status, a combination of linear and nonlinear models were used to find discriminating adducts.[13] First, the following multivariable linear regression model was applied:

$$Y_i = \beta_0 + \beta_1 X_{case} + \beta_2 X_{Sex} + \beta_3 X_{Ethnicity} + \beta_4 X_{Batch} + \beta_5 X_{HK} + \beta_6 X_{IS} + \beta_3 X_{Hb} + \beta_3 X_{DBS\ Age} + \varepsilon_i, \quad (1)$$

where $Y_i$ is a vector of logged, DESeq scaled abundances of the *i*th adduct; $X_{Sex}$ and $X_{Ethnicity}$ are the binary, matching variables; $X_{Batch}$ is a categorical variable indicating the four analytical batches; $X_{HK}$ is a vector of housekeeping peptide abundances representing the amount of digested HSA in each sample; $X_{IS}$ is a vector of internal standard abundances indicating instrument performance; $X_{Hb}$ is a vector of Hb measurements for blood volume normalization; and $X_{DBS\ Age}$ is a vector of DBS sample ages. Adducts were ranked by the nominal *p*-values of

the coefficient $\beta_1$, and the case-control fold change (FC) in adduct abundances were calculated as $\exp(\beta_1)$. Significance levels were adjusted for multiple testing by controlling the false discovery rate $\alpha = 0.05$ using the Benjamini-Hochberg method.[23]

Second, a logistic least absolute shrinkage and selection operator (lasso)[24] regression model was fit with case-control status regressed on the logged, scaled, normalized adduct abundances of the 28 Cys34 adducts along with the matching variables (*i.e.*, sex, ethnicity). Since lasso can be inconsistent in the presence of correlated variables and some adducts were strongly correlated, the logistic lasso regression was performed on 500 bootstrapped datasets to increase the stability of the results.[25] Adducts were ranked by the proportion of times each adduct was selected into the model out of 500 bootstrap iterations. To ensure robust results, this process was repeated with a range of lasso penalty parameter values, which were selected using 10-fold cross validation.

Lastly, adducts were ranked by their nonlinear associations with case-control status using random forest variable importance.[26] A random forest with 500 trees was constructed to predict the case-control status with the logged, scaled, normalized abundances of the 28 Cys34 adducts and the matching variables. The mean decrease in Gini index was used to rank the adducts by their importance in the random forest classifier.[27]

Adducts that were consistently top-ranked in all three statistical approaches, or were highly ranked in the two linear models and within the top-10 ranking adducts in random forest were selected for further investigation. To evaluate the importance of the selected adducts in discriminating leukemia cases and controls, we obtained 10-fold cross-validated area under the receiver operating characteristic curve (cvAUC) estimates by fitting a logistic regression model on case-control status with discriminating adducts as predictors using the 'cvAUC' R package.[28] For the T-cell ALL subgroup with the small sample size, the cvAUC was computed with a 5-fold cross-validation. It should be noted that the cvAUC estimates are likely to be optimistic since we did not have independent test sets.

## 4.4 Results

### *Data Preprocessing and Normalization*

The reproducibility of duplicate measurements was visually assessed by obtaining relative log abundance (RLA) plots[29] with each sample measurement shown side-by-side (Figure S4.1A) and with boxplots showing the difference in duplicate measurements for each sample (Figure S4.1B). To obtain RLA plots, the log-ratios of the adduct abundances of each sample to the median adduct abundance across samples were computed, and sample-wise boxplots of this standardized data were produced. There were no obvious outliers from the boxplots of the difference in duplicate measurements (Figure S4.1B) and RLA plots (Figure S4.1A) so all samples were kept for downstream analyses. Four adducts that were missing in over half of the samples were removed from the analysis (Figure S4.2). The remaining missing values were imputed using *k*-nearest neighbor, with *k*=3 (Figure S4.3). This left a total of 783 subjects and 28 Cys34 adducts for inclusion in downstream statistical analyses.

Figure 4.2A shows the RLA plot of the dataset before normalization. As can be seen, there were noticeable batch effects with the second batch (shown in green in Figure 4.2A) showing the largest deviation. Figure 4.3 shows variation in the other four known unwanted factors of variation (*i.e.*, instrument performance, digested HSA, DBS age, blood volume) by the sample run order. DBS age and blood volume demonstrate consistent variation across all

samples. However, both the internal standard and the housekeeping peptide, representing instrument performance and digested HSA, respectively, indicate a drop in signal similar to what is seen in Figure 4.2A. Moreover, when removal of unwanted variation (RUVg)[30] was used to estimate the unwanted factors of variation based on negative control adducts, the first estimated unwanted factor of variation was highly correlated with the internal standard (Pearson's $r$ = -0.88, Figure S4.4A) and the housekeeping peptide (Pearson's $r$ = -0.76, Figure S4.4B). This suggests that the main source of unwanted variation was technical variation arising from the sample analysis. Figure 4.2B shows the RLA of the logged, DESeq scaled, normalized data which was adjusted for all known factors of unwanted variation. It can be seen that the optimized normalization scheme effectively removed unwanted variation, particularly due to batch effects, since the distribution is centered around zero with small variability.

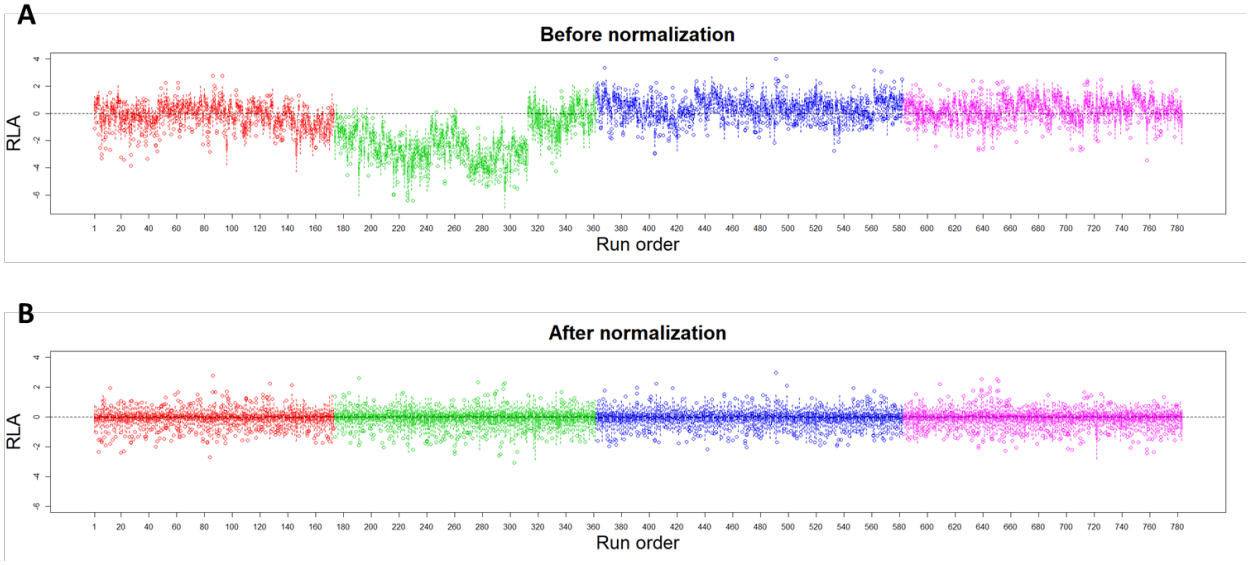



**Figure 4. 2** RLA plots from (A) before and (B) after normalization. Colors represent sample batches 1-4.

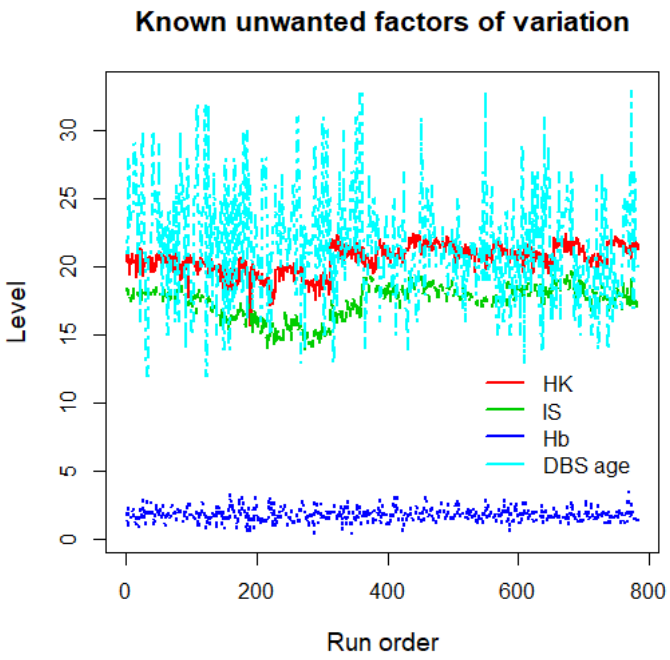



**Figure 4. 3** Variation across sample run order for the following four known unwanted factors of variation: instrument performance (IS), digested HSA (HK), DBS age, blood volume (Hb).

**Table 4. 1** Twenty-eight Cys34 adducts detected in archived newborn DBS from childhood leukemia cases and controls.

| Adduct | Median RT (min) | MIM observed (*m/z*, 3+) | MIM theoretical (*m/z*, 3+) | Δ Mass (ppm) | Added mass (Da) | Elemental composition of added mass | Annotation |
|---|---|---|---|---|---|---|---|
| 796.43[a,c,d,e] | 25.67 | 796.4297 | 796.4301 | -0.53 | -45.9900 | $-CH_2S$ | Cys34→Gly |
| 800.43[c,d,e] | 26.75 | 800.4296 | 800.4301 | -0.65 | -33.9900 | $-SH_2$ | Cys34→Dehydroalanine |
| 805.76[a,c,d,e] | 25.46 | 805.7615 | 805.7618 | -0.34 | -18.0000 | $-SH_2$, +O | Cys34→Oxoalanine |
| 810.43[a] | 25.73 | 810.4336 | | | -3.9900 | | Unknown |
| 811.09[a] | 25.24 | 811.0872 | 811.0875 | -0.38 | -2.0100 | $-H_2$ | Cys34 Sulfenamide |
| 811.76[a,b,c,d,e] | 26.36 | 811.7589 | 811.7594 | -0.52 | 1.0078 | +H | Unmodified T3[f] |
| 815.76[a] | 25.61 | 815.7591 | 815.7594 | -0.35 | 12.0000 | $+CH_2O$, $-H_2O$ | $CH_2$ crosslink |
| 816.42[a,b,c,d,e] | 23.97 | 816.4187 | 816.4191 | -0.51 | 13.9800 | $-H_2$, +O | Cys34 Sulfinamide[f] |
| 816.43[a,b,c,d,e] | 26.63 | 816.4310 | 816.4312 | -0.28 | 15.0178 | $+CH_3$ | Methylation (not at Cys34) |
| 819.09[a] | 26.70 | 819.0870 | 819.0867 | 0.34 | 22.9978 | +Na | *S*-Sodiation |
| 820.09[a,b,c,e] | 26.21 | 820.0911 | 820.0911 | -0.01 | 25.9978 | +CN | *S*-Cyanylation |
| 821.75[a,b] | 26.13 | 821.7504 | 821.7507 | -0.32 | 30.9778 | $-H_2$, $+O_2$ | Cys34 Sulfonamide |
| 822.42[a,b,c,d,e] | 25.66 | 822.4222 | 822.4226 | -0.53 | 32.9878 | $+HO_2$ | Cys34 Sulfinic acid[f] |
| 824.41[d] | 25.81 | 824.4082 | 824.4113 | -3.77 | 38.9578 | +K | *S*-Potassiation |
| 827.09[a,b,c] | 26.05 | 827.0940 | 827.0945 | -0.64 | 46.9978 | $+CH_3O_2$ | *S*-(O)-O-$CH_3$ |
| 827.75[a,b,c,d,e] | 25.97 | 827.7536 | 827.7543 | -0.79 | 48.9778 | $+HO_3$ | Cys34 Sulfonic acid[f] |
| 830.43[a,c] | 25.90 | 830.4343 | 830.4348 | -0.61 | 57.0178 | $+C_3H_5O$ | Unknown (possible *S*-addition of acrolein) |
| 832.43[a] | 26.32 | 832.4259 | 832.4262 | -0.34 | 63.0178 | $+CH_3O_3$ | Unknown |
| 835.11[a,c,e] | 26.44 | 835.1070 | 835.1066 | 0.47 | 71.0578 | $+C_4H_7O$ | *S*-Addition of crotonaldehyde[f] |
| 845.11[a] | 26.05 | 845.1092 | 845.1102 | -1.12 | 101.0578 | $+C_5H_9O_2$ | Unknown |
| 850.10[a,c,d,e] | 26.59 | 850.0957 | 850.0958 | -0.04 | 116.0278 | $+C_4H_6NOS$ | *S*-Addition of hCys ($-H_2O$) |
| 851.43[a,b,c,d,e] | 24.66 | 851.4269 | 851.4274 | -0.63 | 120.0178 | $+C_3H_6NO_2S$ | *S*-Addition of Cys[f] |
| 857.44[a] | 26.39 | 857.4417 | 857.4419 | -0.31 | 138.0478 | $+C_7H_8NO_2$ | Unknown |
| 862.11 | 26.89 | 862.1136 | 862.1096 | 4.63 | 152.0578 | $+C_7H_8N_2O_2$ | Unknown |
| 862.77 | 25.97 | 862.7735 | 862.7694 | 4.70 | 154.0378 | $+C_6H_6N_2O_3$ | Unknown |
| 870.43[a,b,c,d,e] | 24.14 | 870.4347 | 870.4345 | 0.16 | 177.0178 | $+C_5H_9N_2O_3S$ | *S*-Addition of CysGly[f] |
| 913.45[a,b,c,d,e] | 24.77 | 913.4485 | 913.4487 | -0.27 | 306.0778 | $+C_{10}H_{16}N_3O_6S$ | *S*-Addition of GSH[f] |
| 918.12[a,b] | 25.14 | 918.1204 | 918.1206 | -0.22 | 320.0878 | $+C_{11}H_{18}N_3O_6S$ | Unknown (likely *S*-addition of GSH and $CH_2$ crosslink) |

Cys, cysteine; hCys, homocysteine; CysGly, cysteinylglycine; GSH, glutathione; MIM, monoisotopic mass; RT, retention time.

[a]Also detected by Yano et al. 2018.[13]

[b]Also detected by Liu et al. 2018.[31]

[c]Also detected by Grigoryan et al. 2018.[32]

[d]Also detected by Lu et al. 2017.[33]

[e]Also detected by Grigoryan et al. 2016.[18]

[f]Annotation confirmed with a synthetic standard

### *Adducts Detected in Archived Newborn DBS*

Table 4.1 shows the 28 adducts that were detected in the archived newborn DBS. A variety of adducts was observed in addition to the unmodified T3; this includes Cys34 sulfoxidation products, Cys34 disulfides with low-molecular-weight thiols, and Cys34 adducts with reactive carbonyl species (RCS). Nineteen of the 28 are recurring adducts observed in previous analyses of plasma/serum samples.[18,31–33] Of the remaining 9 adducts, two were new to the present study (unknown adducts 862.11 and 862.77; representative MS2 spectra are shown in Figure S4.5), and seven were observed in a previous pilot analysis of archived DBS collected from newborns whose mothers either actively smoked during pregnancy or were nonsmokers.[13]

### *Adducts discriminating childhood leukemia cases and matched controls*

Characteristics of the 783 subjects included in this analysis are presented in Table 4.2. Stratified summary characteristics of ALL cases and matched controls, as well as AML cases and matched controls are also shown. As expected from the matched design, cases and controls were similar with respect to sex, ethnicity, birth year, and DBS age, both overall and for ALL and AML. Hb concentrations were also similar between cases and controls, meaning that the DBS punches had comparable blood volumes. The mean age at diagnosis was ~6 years for both ALL and AML cases.

None of the adducts discriminated between childhood leukemia cases and controls overall. Although adduct 815.76 ($CH_2$ crosslink) was highly-ranked in all three variable selection approaches (Figure S4.6, Table S4.1), this adduct was not statistically significant in the linear model and had a cvAUC of 0.54 (95% confidence interval [CI]: 0.50, 0.58), indicating low prediction. Variable selection results were inconsistent and null findings persisted for total ALL (Figure S4.7, Table S4.2) and total B-cell ALL (Figure S4.8, Table S4.3) subgroups. When the analysis was performed on the B-cell ALL with high-hyperdiploidy subgroup, adduct 824.41 (Cys34 adduct with potassium) was ranked second by the linear model and first by lasso and random forest (Figure S4.9, Table S4.4). Adduct 824.41 had a FC of 0.87 (nominal *p*-value = 0.16), and the cvAUC of this adduct was 0.61 (95% CI: 0.53, 0.69). Furthermore, for the B-cell ALL subgroup with the t(12;21) translocation, adduct 811.09 (Cys34 sulfenamide) was ranked first by the linear model (FC = 0.84, nominal *p*-value = 0.09), second by lasso, and sixth by random forest (Figure S4.10, Table S4.5). For this subgroup, adduct 835.11 (*S*-addition of crotonaldehyde) was first-ranked by lasso and random forest, but this adduct was ranked sixth by the linear model (FC = 1.13, nominal *p*-value = 0.34). The cvAUC of adducts 811.09 and 835.11 were 0.61 (95% CI: 0.50, 0.72) and 0.60 (95% CI: 0.48, 0.72), respectively. For the T-cell ALL subgroup, adducts 835.11 (FC = 1.44, nominal *p*-value = 0.03, adjusted *p*-value = 0.48), 830.43 (possible *S*-addition of acrolein; FC = 1.49, nominal *p*-value = 0.05, adjusted *p*-value = 0.48), and 851.43 (*S*-addition of cysteine; FC = 1.72, nominal *p*-value = 0.05, adjusted *p*-value = 0.48) were the top-three adducts ranked by the linear model (Figure S4.11). All three adducts were within the top-10 ranking adducts with lasso and random forest (Figure S4.11, Table S4.6). The 5-fold cvAUC for adducts 835.11, 830.43, and 851.43 were 0.60 (95% CI: 0.43, 0.78), 0.46 (95% CI: 0.22, 0.70), and 0.47 (95% CI: 0.22, 0.72), respectively.

The top-10 ranking adducts for AML, based on the three variable selection methods, are summarized in Table S4.7. The Cys34 homocysteine adduct with loss of $H_2O$ (adduct 850.10), showed clear discrimination between cases and controls (Figure 4.4). This homocysteine adduct was first-ranked by the linear model (Figure 4.4A) and by lasso (Figure 4.4B), and was ranked second by random forest (Figure 4.4C). AML cases had lower levels of the homocysteine adduct

compared to controls (FC = 0.66, nominal *p*-value = 0.01, adjusted *p*-value = 0.39). The homocysteine adduct alone provided a cvAUC of 0.69 (95% CI: 0.58, 0.81). Although adduct 851.43 (*S*-addition of cysteine) was second-ranked by the linear model with a nominal *p*-value of 0.05 (FC = 0.65, adjusted *p*-value = 0.63), this adduct was not highly ranked by lasso or random forest.

**Table 4. 2** Summary characteristics of all subjects overall, and stratified by ALL and AML.

| | **Overall** | | **ALL** | | **AML** | |
|---|---|---|---|---|---|---|
| | **Cases** | **Controls** | **Cases** | **Controls** | **Cases** | **Controls** |
| **Number of samples** | 387 | 396 | 339 | 332 | 45 | 44 |
| **Child's sex** | | | | | | |
| Male | 222 | 227 | 191 | 188 | 28 | 27 |
| Female | 165 | 169 | 148 | 144 | 17 | 17 |
| **Child's ethnicity** | | | | | | |
| Hispanic | 180 | 185 | 158 | 156 | 22 | 22 |
| Non-Hispanic | 207 | 211 | 181 | 176 | 23 | 22 |
| **Child's birth year** | | | | | | |
| Mean (SD) | 1996 (4) | 1996 (4) | 1996 (4) | 1996 (4) | 1997 (5) | 1997 (5) |
| Min | 1985 | 1985 | 1985 | 1985 | 1986 | 1986 |
| Max | 2006 | 2006 | 2005 | 2005 | 2006 | 2006 |
| **Child's age at diagnosis (yr)** | | | | | | |
| Mean (SD) | 5.68 (3.56) | | 5.69 (3.45) | | 5.64 (4.37) | |
| Min | 0.00 | | 0.00 | | 0.00 | |
| Max | 14.76 | | 14.76 | | 14.06 | |
| **DBS age (yr)** | | | | | | |
| Mean (SD) | 22 (4) | 22 (4) | 22 (4) | 22 (4) | 21 (5) | 21 (5) |
| Min | 12 | 12 | 13 | 13 | 12 | 12 |
| Max | 33 | 33 | 33 | 33 | 32 | 32 |
| **Hb (mg/mL)** | | | | | | |
| Mean (SD) | 1.83 (0.51) | 1.86 (0.46) | 1.85 (0.50) | 1.86 (0.46) | 1.68 (0.52) | 1.96 (0.45) |
| Min | 0.48 | 0.41 | 0.48 | 0.41 | 0.74 | 0.71 |
| Max | 3.56 | 3.17 | 3.56 | 3.17 | 3.51 | 3.14 |

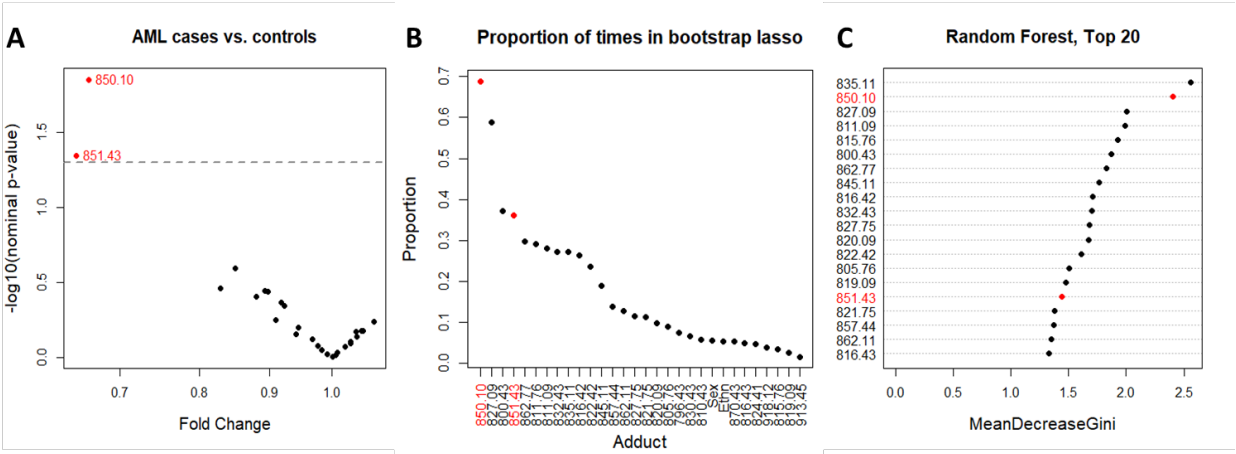



**Figure 4. 4** Variable selection for AML cases and controls using (A) multivariable linear regression, (B) logistic lasso regression on bootstrapped data, and (C) random forest. (A) Volcano plot showing the case/control fold change and the corresponding nominal *p*-value for each adduct. The dotted line represents a nominal *p*-value of 0.05. (B) Adducts ranked by the proportion of times each adduct was selected into the lasso model out of 500 bootstrap iterations. (C) The top 20 adducts ranked by random forest variable importance (*i.e.*, mean decrease in Gini index).

## 4.5 Discussion

The purpose of this study was to perform untargeted adductomics, using archived newborn DBS to characterize *in utero* exposures to reactive electrophiles that are associated with childhood leukemia. In our previous adductomics study of DBS from newborns of smoking and nonsmoking mothers, we had shown the importance of preprocessing and normalization steps to remove unwanted technical variation prior to statistical analyses.[13] Results from the current study reinforce this finding. Indeed, over the several months required for analysis of 783 samples in this study, normalization for run order proved to be essential in removing unwanted variation in LC-HRMS performance (Figures 4.2 and 4.3). Such technical variation can easily obscure detection of meaningful biological variability in the data (*e.g.*, case-control differences).

Statistical analyses of case-control differences in incidences of ALL and AML led to different results. While no clear associations were detected between adduct abundances and ALL, discrepancies in the results across the different ALL subtypes may point to the biological heterogeneity and distinct etiologies of this disease.[34] In fact, there were suggestions that Cys34 modifications by RCS were associated with B-cell ALL with t(12;21) and T-cell ALL. Upon lipid peroxidation, various reactive oxygen species (ROS) can oxidize membrane lipids to generate highly reactive α, β-unsaturated aldehydes such as acrolein and crotonaldehyde.[35] Despite small sample sizes, Cys34 adducts of acrolein (adduct 830.43) and crotonaldehyde (adduct 835.11) were both elevated in T-cell ALL cases compared to controls (FC of 1.49 and 1.44, respectively). The crotonaldehyde adduct was also slightly more abundant in B-cell ALL cases with t(12;21) compared to controls, with a FC of 1.13. Although these findings need to be replicated with larger sample sizes, elevated levels of these Cys34 adducts of RCS suggest that oxidative stress may be involved in ALL. Indeed, increased levels of oxidative damage to proteins, as measured by protein carbonylation, has been observed in ALL cases.[36] Moreover, elevated levels of carbonylated HSA has been observed in lung and liver cancer as well as other chronic diseases and conditions.[37] Since protein carbonylation is irreversible,[38] increases in

Cys34 adducts of RCS may be indicative of the decreased antioxidant capacity of HSA in particular subtypes of ALL.

In the smaller set of AML cases and matched controls, the Cys34 adduct of homocysteine with loss of $H_2O$ (adduct 856.10) was consistently highly-ranked by the three variable selection methods used to test for associations between adducts and case-control status. The homocysteine adduct was substantially lower in AML cases compared to controls with a FC of 0.66. This finding offers potential insight into the biological mechanisms of AML etiology. Homocysteine is involved in folate-mediated one-carbon metabolism and is a key intermediate in the methionine recycling pathway.[39] Remethylation of homocysteine produces methionine, which is a precursor for S-adenosylmethionine, a universal methyl group donor that plays an important role in DNA methylation.[40] Since DNA methylation is one of the main epigenetic mechanisms, altered levels of homocysteine could reflect disruptions in epigenetic regulation leading to abnormal methylation patterns, which are often seen in cancer cells.[41–43] DNA methylation is regulated by DNA methyltransferases (DNMTs), and *DNMT3A* is one of the most frequently mutated genes in adult AML.[44] Mutations in epigenetic regulators have been shown to occur as early events in hematopoietic stem cells to produce preleukemic clones that can accrue additional proliferative mutations leading to AML.[45–47] However, Bolouri et al. recently showed that the genomic, transcriptomic, and epigenomic landscapes of adult and childhood AML differ substantially.[44] They further showed that mutations in epigenetic regulators, such as *DNMT3A*, were either absent or occurred less frequently in childhood AML compared to adults.[44] However, epigenetics of childhood AML is not well characterized compared to adult AML, and epigenetic aberrations involving hypermethylation of CpG island gene promoters and hypomethylation of other regions may contribute to childhood AML.[44,48,49]

Plasma homocysteine is also an inverse indicator of folate status,[50,51] and the observed lower levels of the homocysteine adduct in AML cases may reflect higher folate levels among cases. This is interesting because elevated plasma homocysteine levels have been associated with cardiovascular disease risk, and an increased folate intake has a protective effect on the development of neural tube defects.[39] However, folate has been shown to have a dual effect on cancer, where folate is protective against cancer initiation but promotes the progression of pre-neoplastic cells and subclinical cancers.[52] The adverse effects of folic acid supplementation and the resulting high folate status was first observed in the 1940s by Sidney Farber and colleagues who saw that the administration of folic acid accelerated the progression of acute leukemia in children.[40] Moreover, cancer proliferation is inhibited by limiting the supply of folic acid and its active tetrahydrofolate cofactor metabolites, which is the basis of the use of antifolate drugs in cancer therapy.[52]

Maternal prenatal folate intake through diet or supplements and risk of AML has been studied previously, but the findings have been inconclusive. Some studies have found null associations,[53,54] while a large, pooled analysis of 12 case-control studies from the Childhood Leukemia International Consortium[55] found a protective effect of maternal folic acid supplementation against AML (odds ratio [OR] = 0.68; 95% CI: 0.48, 0.96).[56] In a study of genetic variation in folate metabolism, no associations were observed for childhood AML and polymorphisms in the enzyme, methylenetetrahydrofolate reductase (*MTHFR*) that is central for DNA synthesis and regulation of 5-methyl tetrahydrofolate, which is a methyl donor for conversion of homocysteine to methionine.[57] However, in the same study, an increased risk of AML was observed for genetic variation in methionine synthase (*MTR*), the enzyme catalyzing remethylation of homocysteine to methionine, with an OR of 2.74 (95% CI: 1.07, 7.01).[57]

Dysfunction in MTR can disrupt the synthesis of purines and thymidine and thereby impact rapidly dividing blood cells in the bone marrow.[39] Our results, showing lower levels of the homocysteine adduct in AML cases, are consistent with these previous findings and further suggest development of AML may be mediated by alterations in methionine metabolism.

In addition to being remethylated to methionine, homocysteine can also be irreversibly degraded via the transsulfuration pathway to cysteine.[39] Homocysteine is metabolized by cystathionine β-synthase (CBS) to cystathioine, which is subsequently converted to cysteine by cystathionine γ-lyase.[39] This relationship between homocysteine and cysteine may explain why the Cys34 cysteinylated adduct (adduct 851.43) was also highly ranked for AML (Figure 4.4A ,B). Moreover, CBS contributes to the endogenous production of hydrogen sulfide, which has recently been shown to promote cancer progression.[58,59] Hydrogen sulfide is a signaling molecule that could potentially promote cancer by mechanisms such as disrupting redox homeostasis or by stimulating cell proliferation and survival.[58] Thus, while the relationship between CBS activity and leukemogenesis is unknown, it is possible that CBS dysfunction may promote AML through modulation of homocysteine and hydrogen sulfide.

To better understand how altered levels of homocysteine may be related to changes in epigenetic regulations or folate status, our adductomics findings should be integrated with other omics to obtain a cohesive view of the disease mechanisms involved in AML. Metabolomics may reveal global changes in small molecules related to the folate pathway, and genetic variation in other key enzymes in the homocysteine-related pathways, such as methionine synthase reductase and CBS, should also be examined. More extensive genome-wide epigenomic studies may refine our understanding of how epigenetic events are involved in childhood AML.

Because this was a hypothesis-generating study and some of the subgroups (particularly AML and T-cell ALL) were limited in sample size, our findings warrant replication in future studies. We cannot dismiss the possibility of false positive associations, particularly for the T-cell ALL subgroup which consisted of 38 subjects. Nonetheless, oxidative damage is consistently seen in cancer, and it is possible oxidative stress may also play a role in ALL.[36,37] Associations between the homocysteine adduct and AML were robust to the different variable selection methods and had a relatively high cvAUC. However, these results will also need to be replicated with a larger sample size with the inclusion of the various subtypes of AML. In this analysis, few discriminating adducts were detected and the effects sizes were relatively small. Although HSA adducts encompass exposures occurring during the last month of gestation, this single time point may not have captured all relevant exposures in the etiology of childhood leukemia. It is possible that important exposures may have occurred earlier in gestation and may not be reflected in HSA adducts. It is also possible that early markers of leukemogenesis and susceptibility may not be readily detectable at birth, given the long latency of the disease. However, it is important to emphasize that our analyses of newborn DBS detected differences in cases and controls in DBS collected years prior to diagnosis. Integration of different omics may reveal the biological mechanisms of the initial events in leukemogenesis and could provide new avenues for prevention and treatment of childhood leukemia.

## 4.6 Conclusion

In conclusion, we performed untargeted adductomics using 783 DBS from a population-based case-control study to identify HSA-Cys34 adducts associated with childhood leukemia. Although we found no differences in Cys34 adduct abundances between childhood leukemia

cases and controls overall, particular subtypes (*i.e.*, B-cell ALL with t(12;21) and T-cell ALL) had higher abundances of adducts of RCS, suggestive of oxidative stress and lipid peroxidation as potentially etiologic factors. Interestingly, there was a 0.66-fold difference in abundances of a Cys34 homocysteine adduct (with loss of $H_2O$) between AML cases and matched controls. Since homocysteine is an important intermediate in the folate-methionine metabolism, this may point to alterations in one-carbon metabolism and epigenetic changes as predictors of AML. Future studies should include larger sample sizes of ALL/AML cases and their subtypes to replicate these findings and further explore their implications.

## 4.7 Compliance with ethical standards

***Conflict of interest***
The authors declare they have no conflict of interest.

***Ethics approval***
The CCLS and Center for Integrative Research on Childhood Leukemia and the Environment were approved by the University of California Committee for the Protection of Human Subjects, the California Health and Human Services Agency Committee for the Protection of Human Subjects, and the institutional review boards of all participating hospitals, as appropriate.

***Informed consent***
Written informed consent was obtained from parents of all participating subjects in the CCLS.

## 4.8 Acknowledgements

This work was supported by the National Institute for Environmental Health Sciences of the U.S. National Institutes of Health (NIEHS grants P01ES018172, P50ES018172, R01ES009137 and P42ES0470518), by the U.S. Environmental Protection Agency (USEPA grants RD83451101 and RD83615901), and by a grant for a pilot project from Children with Cancer, a registered Charity in the U.K.

The biospecimens used in this study were obtained from the California Biobank Program, (SIS request number 26), in accordance with Section 6555(b), 17 CCR. The NIEHS, USEPA, and California Department of Public Health are not responsible for the results or conclusions drawn by the authors of this publication.

## 4.9 References


(1) Whitehead, T. P.; Metayer, C.; Wiemels, J. L.; Singer, A. W.; Miller, M. D. Childhood Leukemia and Primary Prevention. *Curr. Probl. Pediatr. Adolesc. Health Care* **2016**, *46* (10), 317–352.

(2) Greaves, M. Childhood Leukemia. *Bmj* **2002**, *324*, 283–287.

(3) Ward, E.; DeSantis, C.; Robbins, A.; Kohler, B.; Jemal, A. Childhood and Adolescent Cancer Statistics, 2014. *CA. Cancer J. Clin.* **2014**, *64* (2), 83–103.

(4) Greaves, M. A Causal Mechanism for Childhood Acute Lymphoblastic Leukaemia. *Nat. Rev. Cancer* **2018**, *18* (8), 471–484.

(5) Erdmann, F.; Kielkowski, D.; Schonfeld, S. J.; Kellett, P.; Stanulla, M.; Dickens, C.; Kaatsch, P.; Singh, E.; Schüz, J. Childhood Cancer Incidence Patterns by Race, Sex and Age for 2000-2006: A Report from the South African National Cancer Registry. *Int. J. Cancer* **2015**, *136* (11), 2628–2639.

(6) Wiemels, J. Perspectives on the Causes of Childhood Leukemia. *Chem. Biol. Interact.* **2012**, *196* (3), 59–67.

(7) Greaves, M. Infection, Immune Responses and the Aetiology of Childhood Leukaemia. *Nat. Rev. Cancer* **2006**, *6* (3), 193–203.

(8) Greaves, M. In Utero Origins of Childhood Leukaemia. *Early Hum. Dev.* **2005**, *81* (1), 123–129.

(9) Wiemels, J. L.; Ford, A. M.; Van Wering, E. R.; Postma, A.; Greaves, M. Protracted and Variable Latency of Acute Lymphoblastic Leukemia after TEL-AML1 Gene Fusion in Utero. *Blood* **1999**, *94* (3), 1057–1062.

(10) Mori, H.; Colman, S. M.; Xiao, Z.; Ford, A. M.; Healy, L. E.; Donaldson, C.; Hows, J. M.; Navarrete, C.; Greaves, M. Chromosome Translocations and Covert Leukemic Clones Are Generated during Normal Fetal Development. *Proc. Natl. Acad. Sci.* **2002**, *99* (12), 8242–8247.

(11) Greaves, M. F.; Wiemels, J. Origins of Chromosome Translocations in Childhood Leukaemia. *Nat. Rev. Cancer* **2003**, *3* (9), 639–649.

(12) Aldini, G.; Vistoli, G.; Regazzoni, L.; Gamberoni, L.; Facino, R. M.; Yamaguchi, S.; Uchida, K.; Carini, M. Albumin Is the Main Nucleophilic Target of Human Plasma: A Protective Role Against Pro-Atherogenic Electrophilic Reactive Carbonyl Species? *Chem. Res. Toxicol.* **2008**, *21* (4), 824–835.

(13) Yano, Y.; Grigoryan, H.; Schiffman, C.; Edmands, W. M. B.; Petrick, L.; Hall, K.; Whitehead, T. P.; Metayer, C.; Dudoit, S.; Rappaport, S. M. Untargeted Adductomics of Cys34 Modifications to Human Serum Albumin in Newborn Dried Blood Spots. Manuscript Submitted for Publication. **2018**.

(14) Rappaport, S. M.; Li, H.; Grigoryan, H.; Funk, W. E.; Williams, E. R. Adductomics: Characterizing Exposures to Reactive Electrophiles. *Toxicol. Lett.* **2012**, *213* (1), 83–90.

(15) Grigoryan, H.; Li, H.; Iavarone, A. T.; Williams, E. R.; Rappaport, S. M. Cys34 Adducts of Reactive Oxygen Species in Human Serum Albumin. *Chem. Res. Toxicol.* **2012**, *25* (8), 1633–1642.

(16) Metayer, C.; Zhang, L.; Wiemels, J. L.; Bartley, K.; Schiffman, J.; Ma, X.; Aldrich, M. C.; Chang, J. S.; Selvin, S.; Fu, C. H.; et al. Tobacco Smoke Exposure and the Risk of Childhood Acute Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype. *Cancer Epidemiol. Biomarkers Prev.* **2013**, *22* (9), 1600–1611.

(17) California Department of Public Health, Genetic Disease Screening Program https://www.cdph.ca.gov/Programs/CFH/DGDS/Pages/default.aspx (accessed Oct 10, 2018).
(18) Grigoryan, H.; Edmands, W.; Lu, S. S.; Yano, Y.; Regazzoni, L.; Iavarone, A. T.; Williams, E. R.; Rappaport, S. M. Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin. *Anal. Chem.* **2016**, *88* (21), 10504–10512.
(19) R Development Core Team, (2016). R: A Language and Environment for Statistical Computing. https://www.r-project.org/ (accessed Mar 8, 2017).
(20) Schiffman, C.; Petrick, L.; Perttula, K.; Yano, Y.; Carlsson, H.; Whitehead, T.; Metayer, C.; Hayes, J.; Edmands, W. M. B.; Rappaport, S.; et al. Data-Adaptive Pipeline for Filtering and Normalizing Metabolomics Data. **2018**, 1–12.
(21) Cole, M.; Risso, D. Scone: Single Cell Overview of Normalized Expression Data. *R Packag. version 1.2.0* **2018**.
(22) Petrick, L.; Edmands, W.; Schiffman, C.; Grigoryan, H.; Perttula, K.; Yano, Y.; Dudoit, S.; Whitehead, T.; Metayer, C.; Rappaport, S. An Untargeted Metabolomics Method for Archived Newborn Dried Blood Spots in Epidemiologic Studies. *Metabolomics* **2017**, *13* (3), 27.
(23) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57* (1), 289–300.
(24) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58* (1), 267–288.
(25) Bach, F. R. Bolasso: Model Consistent Lasso Estimation through the Bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*; ACM Press: New York, New York, USA, 2008; pp 33–40.
(26) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2* (3), 18–22.
(27) Calle, M. L.; Urrea, V. Letter to the Editor: Stability of Random Forest Importance Measures. *Brief. Bioinform.* **2011**, *12* (1), 86–89.
(28) LeDell, E.; Petersen, M.; Laan, M. Van Der. Cross-Validated Area Under the ROC Curve Confidence Intervals. *R Packag. version 1.1.0* **2016**.
(29) De Livera, A. M.; Dias, D. A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P. Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **2012**, *84* (24), 10768–10776.
(30) Risso, D.; Ngai, J.; Speed, T. P.; Dudoit, S. Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples. *Nat. Biotechnol.* **2014**, *32* (9), 896–902.
(31) Liu, S.; Grigoryan, H.; Edmands, W. M. B.; Dagnino, S.; Sinharay, R.; Cullinan, P.; Collins, P.; Chung, K. F.; Barratt, B.; Kelly, F. J.; et al. Cys34 Adductomes Differ between Patients with Chronic Lung or Heart Disease and Healthy Controls in Central London. *Environ. Sci. Technol.* **2018**, *52* (4), 2307–2313.
(32) Grigoryan, H.; Edmands, W. M. B.; Lan, Q.; Carlsson, H.; Vermeulen, R.; Zhang, L.; Yin, S.-N.; Li, G.-L.; Smith, M. T.; Rothman, N.; et al. Adductomic Signatures of Benzene Exposure Provide Insights into Cancer Induction. *Carcinogenesis* **2018**, *39* (5), 661–668.
(33) Lu, S. S.; Grigoryan, H.; Edmands, W. M. B.; Hu, W.; Iavarone, A. T.; Hubbard, A.; Rothman, N.; Vermeulen, R.; Lan, Q.; Rappaport, S. M. Profiling the Serum Albumin Cys34 Adductome of Solid Fuel Users in Xuanwei and Fuyuan, China. *Environ. Sci.*

*Technol.* **2017**, *51* (1), 46–57.
(34) Pui, C.-H.; Carroll, W. L.; Meshinchi, S.; Arceci, R. J. Biology, Risk Stratification, and Therapy of Pediatric Acute Leukemias: An Update. *J. Clin. Oncol.* **2011**, *29* (5), 551–565.
(35) Lee, S. E.; Park, Y. S. Role of Lipid Peroxidation-Derived α , β -Unsaturated Aldehydes in Vascular Dysfunction. *Oxid. Med. Cell. Longev.* **2013**, *2013*, 1–7.
(36) Battisti, V.; Maders, L. D. K.; Bagatini, M. D.; Santos, K. F.; Spanevello, R. M.; Maldonado, P. a.; Brulé, A. O.; Araújo, M. D. C.; Schetinger, M. R. C.; Morsch, V. M. Measurement of Oxidative Stress and Antioxidant Status in Acute Lymphoblastic Leukemia Patients. *Clin. Biochem.* **2008**, *41*, 511–518.
(37) Colombo, G.; Clerici, M.; Giustarini, D.; Rossi, R.; Milzani, A.; Dalle-Donne, I. Redox Albuminomics: Oxidized Albumin in Human Diseases. *Antioxid. Redox Signal.* **2012**, *17* (11), 1515–1527.
(38) Dalle-Donne, I.; Aldini, G.; Carini, M.; Colombo, R.; Rossi, R.; Milzani, A. Protein Carbonylation, Cellular Dysfunction, and Disease Progression. *J. Cell. Mol. Med.* **2006**, *10* (2), 389–406.
(39) Blom, H. J.; Smulders, Y. Overview of Homocysteine and Folate Metabolism. With Special References to Cardiovascular Disease and Neural Tube Defects. *J. Inherit. Metab. Dis.* **2011**, *34* (1), 75–81.
(40) Strickland, K. C.; Krupenko, N. I.; Krupenko, S. A. Molecular Mechanisms Underlying the Potentially Adverse Effects of Folate. *Clin. Chem. Lab. Med.* **2013**, *51* (3), 607–616.
(41) Rush, L. J.; Plass, C. Alterations of DNA Methylation in Hematologic Malignancies. *Cancer Lett.* **2002**, *185* (1), 1–12.
(42) Yamashita, Y.; Yuan, J.; Suetake, I.; Suzuki, H.; Ishikawa, Y.; Choi, Y. L.; Ueno, T.; Soda, M.; Hamada, T.; Haruta, H.; et al. Array-Based Genomic Resequencing of Human Leukemia. *Oncogene* **2010**, *29* (25), 3723–3731.
(43) Ehrlich, M. DNA Methylation in Cancer: Too Much, but Also Too Little. *Oncogene* **2002**, *21* (35), 5400–5413.
(44) Bolouri, H.; Farrar, J. E.; Triche, T.; Ries, R. E.; Lim, E. L.; Alonzo, T. A.; Ma, Y.; Moore, R.; Mungall, A. J.; Marra, M. A.; et al. The Molecular Landscape of Pediatric Acute Myeloid Leukemia Reveals Recurrent Structural Alterations and Age-Specific Mutational Interactions. *Nat. Med.* **2018**, *24* (1), 103–112.
(45) Corces-Zimmerman, M. R.; Hong, W.-J.; Weissman, I. L.; Medeiros, B. C.; Majeti, R. Preleukemic Mutations in Human Acute Myeloid Leukemia Affect Epigenetic Regulators and Persist in Remission. *Proc. Natl. Acad. Sci.* **2014**, *111* (7), 2548–2553.
(46) Shlush, L. I.; Zandi, S.; Mitchell, A.; Chen, W. C.; Brandwein, J. M.; Gupta, V.; Kennedy, J. A.; Schimmer, A. D.; Schuh, A. C.; Yee, K. W.; et al. Identification of Pre-Leukaemic Haematopoietic Stem Cells in Acute Leukaemia. *Nature* **2014**, *506* (7488), 328–333.
(47) Grimwade, D.; Ivey, A.; Huntly, B. J. P. Molecular Landscape of Acute Myeloid Leukemia in Younger Adults and Its Clinical Relevance. *Blood* **2016**, *127* (1), 29–41.
(48) Newcombe, A. A.; Gibson, B. E. S.; Keeshan, K. Harnessing the Potential of Epigenetic Therapies for Childhood Acute Myeloid Leukemia. *Exp. Hematol.* **2018**, *63*, 1–11.
(49) Liang, D.; Liu, H.; Yang, C.; Jaing, T.; Hung, I.; Yeh, T.; Chen, S.; Hou, J.; Huang, Y.; Shih, Y.; et al. Cooperating Gene Mutations in Childhood Acute Myeloid Leukemia with and DNMT3A. *Blood* **2013**, *121* (15), 2988–2995.
(50) Ueland, P. M.; Hustad, S. Homocysteine and Folate Status in an Era of Folic Acid Fortification: Balancing Benefits, Risks, and B-Vitamins. *Clin. Chem.* **2008**, *54* (5), 779–

781.
(51) Pfeiffer, C. M.; Caudill, S. P.; Gunter, E. W.; Osterloh, J.; Sampson, E. J. Biochemical Indicators of B Vitamin Status in the US Population after Folic Acid Fortification: Results from the National Health and Nutrition Examination Survey 1999–2000. *Am. J. Clin. Nutr.* **2005**, *82* (2), 442–450.
(52) Smith, A. D.; Kim, Y.-I.; Refsum, H. Is Folic Acid Good for Everyone? *Am. J. Clin. Nutr.* **2008**, *87* (3), 517–533.
(53) Chokkalingam, A. P.; Chun, D. S.; Noonan, E. J.; Pfeiffer, C. M.; Zhang, M.; Month, S. R.; Taggart, D. R.; Wiemels, J. L.; Metayer, C.; Buffler, P. a. Blood Levels of Folate at Birth and Risk of Childhood Leukemia. *Cancer Epidemiol. Biomarkers Prev.* **2013**, *22* (6), 1088–1094.
(54) Singer, A. W.; Selvin, S.; Block, G.; Golden, C.; Carmichael, S. L.; Metayer, C. Maternal Prenatal Intake of One-Carbon Metabolism Nutrients and Risk of Childhood Leukemia. *Cancer Causes Control* **2016**, *27* (7), 929–940.
(55) Metayer, C.; Milne, E.; Clavel, J.; Infante-Rivard, C.; Petridou, E.; Taylor, M.; Sch??z, J.; Spector, L. G.; Dockerty, J. D.; Magnani, C.; et al. The Childhood Leukemia International Consortium. *Cancer Epidemiol.* **2013**, *37* (3), 336–347.
(56) Metayer, C.; Milne, E.; Dockerty, J. D.; Clavel, J.; Pombo-De-Oliveira, M. S.; Wesseling, C.; Spector, L. G.; Schüz, J.; Petridon, E.; Ezzat, S.; et al. Maternal Supplementation with Folic Acid and Other Vitamins and Risk of Leukemia in Offspring: A Childhood Leukemia Internotionol Consortium Study. *Epidemiology* **2014**, *25* (6), 811–822.
(57) Lightfoot, T. J.; Johnston, W. T.; Painter, D.; Simpson, J.; Roman, E.; Skibola, C. F.; Smith, M. T.; Allan, J. M.; Taylor, G. M. Genetic Variation in the Folate Metabolic Pathway and Risk of Childhood Leukemia. *Blood* **2010**, *115* (19), 3923–3929.
(58) Zhu, H.; Blake, S.; Chan, K. T.; Pearson, R. B.; Kang, J. Cystathionine β -Synthase in Physiology and Cancer. *Biomed Res. Int.* **2018**, *2018*, 1–11.
(59) Wu, D.; Si, W.; Wang, M.; Lv, S.; Ji, A.; Li, Y. Hydrogen Sulfide in Cancer: Friend or Foe? *Nitric Oxide* **2015**, *50*, 38–45.

## 4.10 Supporting information

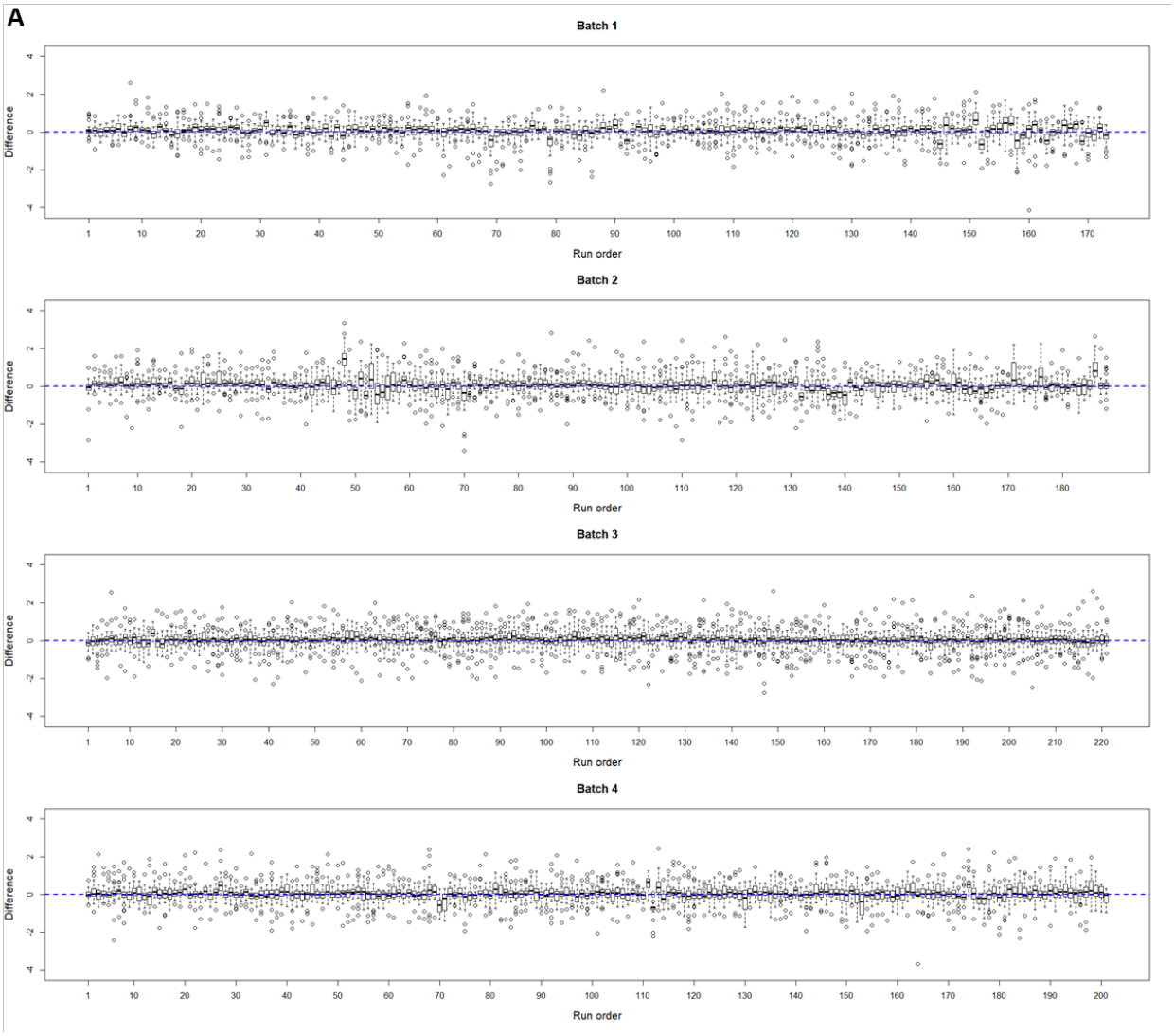
### 4.10.1 Supplemental Methods

**_k_-nearest neighbor imputation**

To find the *k*-nearest adduct neighbors, pairwise distances between adducts were calculated using the Euclidean distance based on all non-missing values. To optimize the value of the parameter *k*, 9 low abundance adducts missing in over 30% of the samples were randomly chosen to be made missing in 100 samples, their values were imputed using values of *k* ranging from 1 to 8, and the mean squared error was calculated for each *k*. The *k* with the smallest mean squared error, $k = 3$, was used to impute missing values in the dataset (see Figure S4.3).

### 4.10.2 Supplemental Figures

**Figure S4. 1** Visual assessment of the reproducibility of duplicate injections. (A) Boxplot of the difference in duplicate measurements for each subject. Samples are shown by run order in each batch. (B) Relative log abundance each sample measurement. Duplicate measurements are shown side-by-side for each sample. Samples are colored by every other sample and shown in run order by each batch.



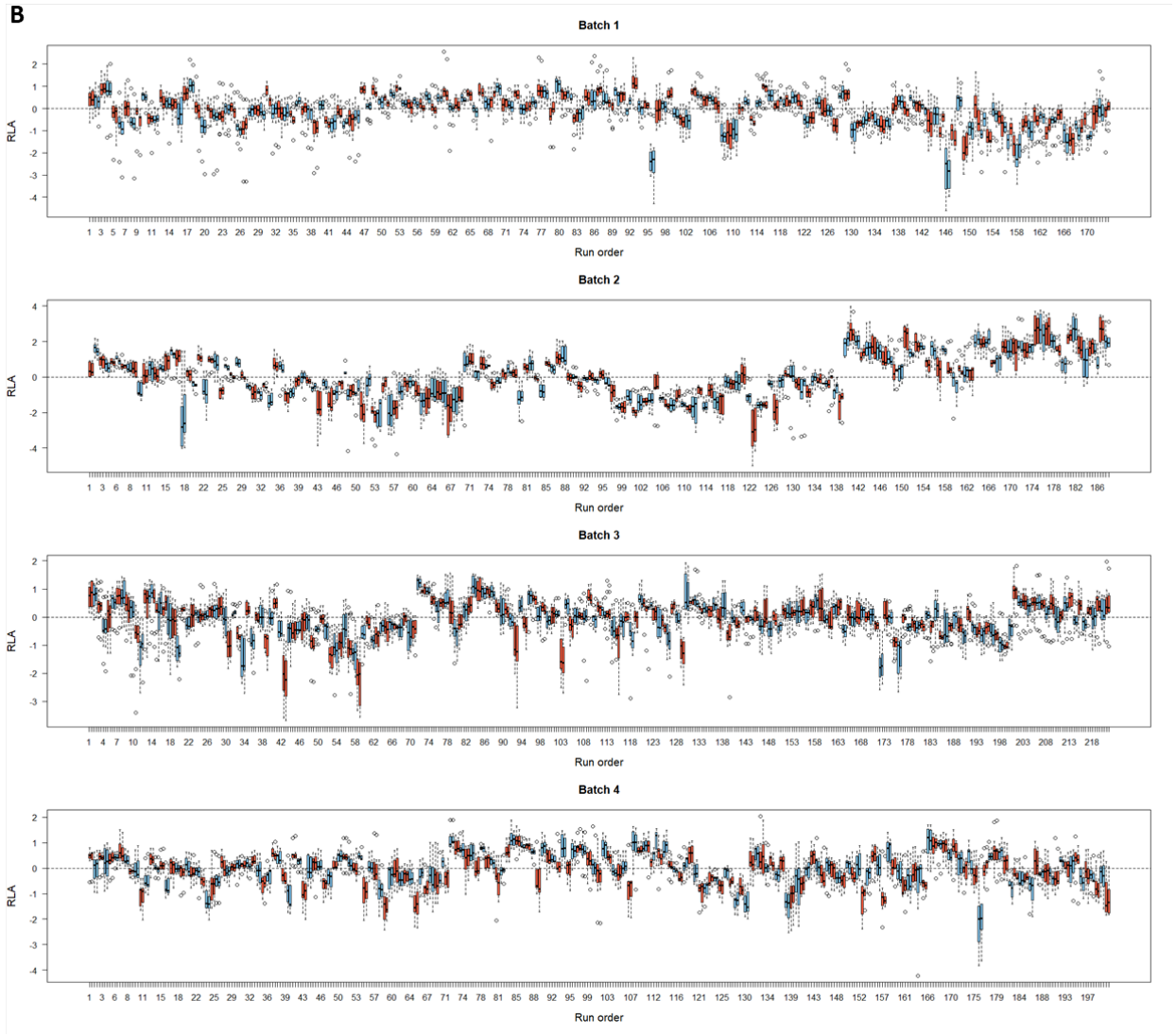

B
Batch 1
RLA
Run order
Batch 2
RLA
Run order
Batch 3
RLA
Run order
Batch 4
RLA
Run order

**Figure S4. 2** Percentage of samples with missing values for each adduct. Four adducts missing in over 50% of the samples (red dotted line) were excluded from the analysis.

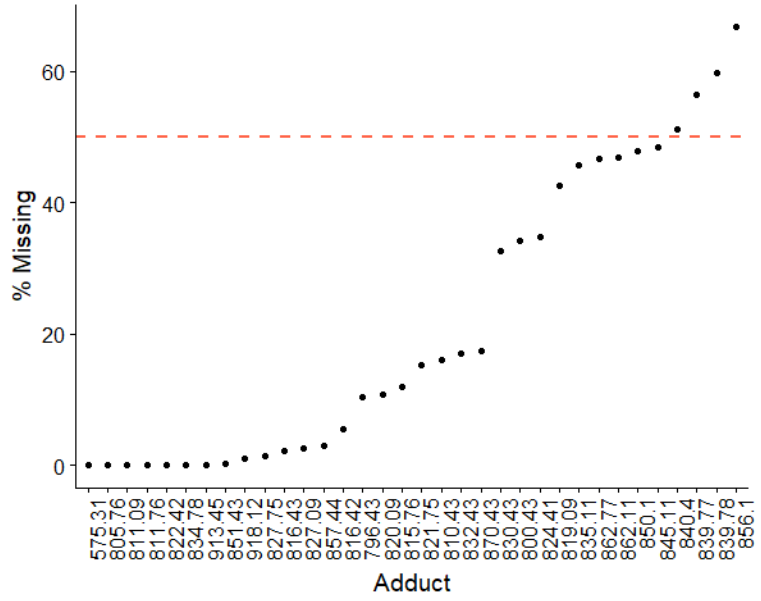



**Figure S4. 3** Optimization of the parameter *k* in *k*-nearest neighbor imputation. The mean squared error (MSE) calculated for each *k* is shown. The *k* with the smallest MSE, *k* = 3, was selected.

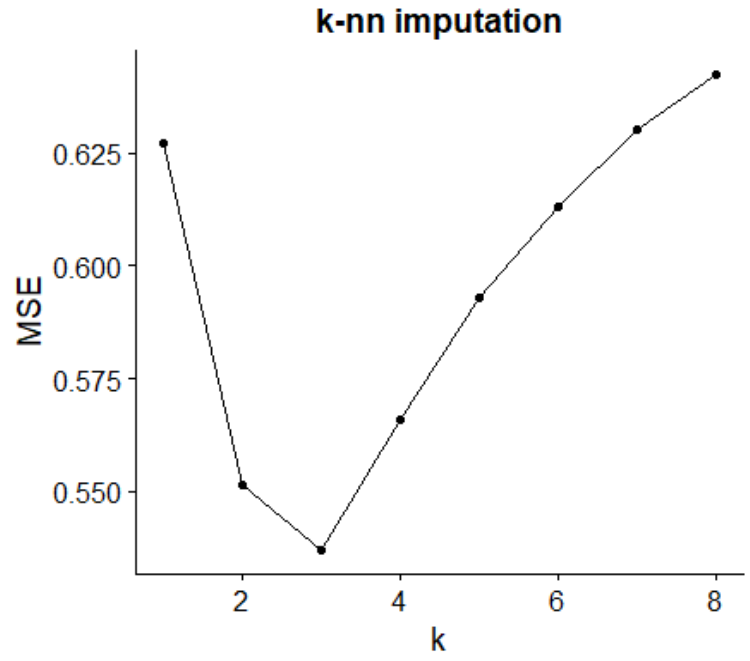


**Figure S4. 4** Correlation between the first estimated unwanted factor of variation from RUVg and (A) the internal standard (IS), (B) the housekeeping peptide (HK).

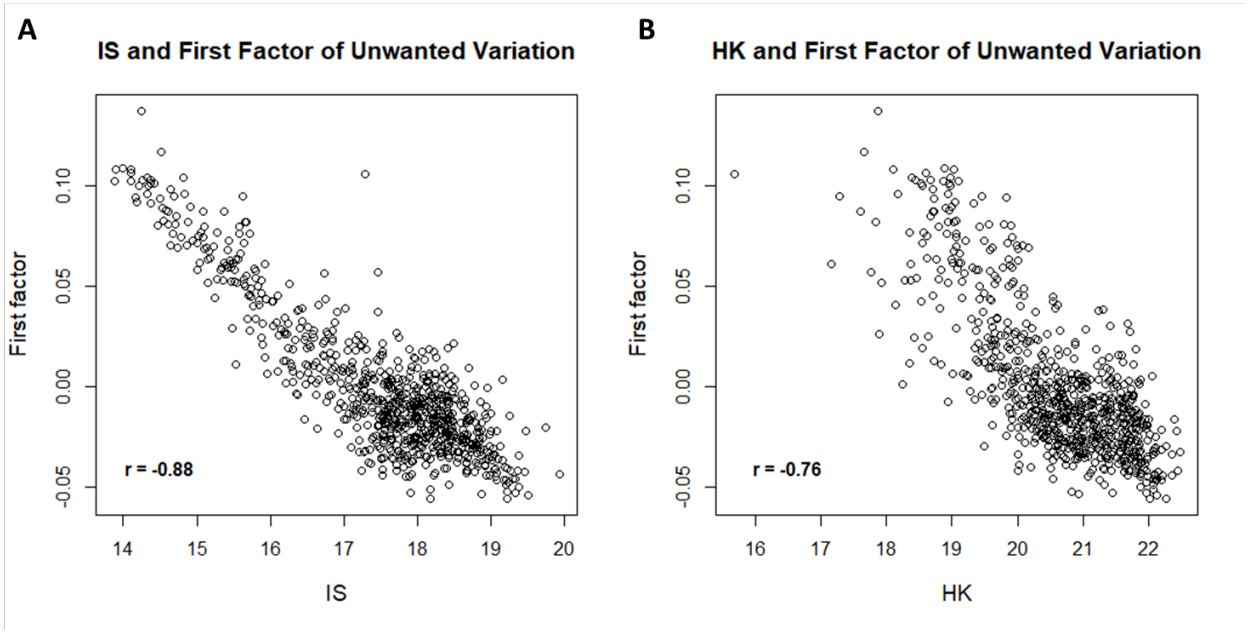
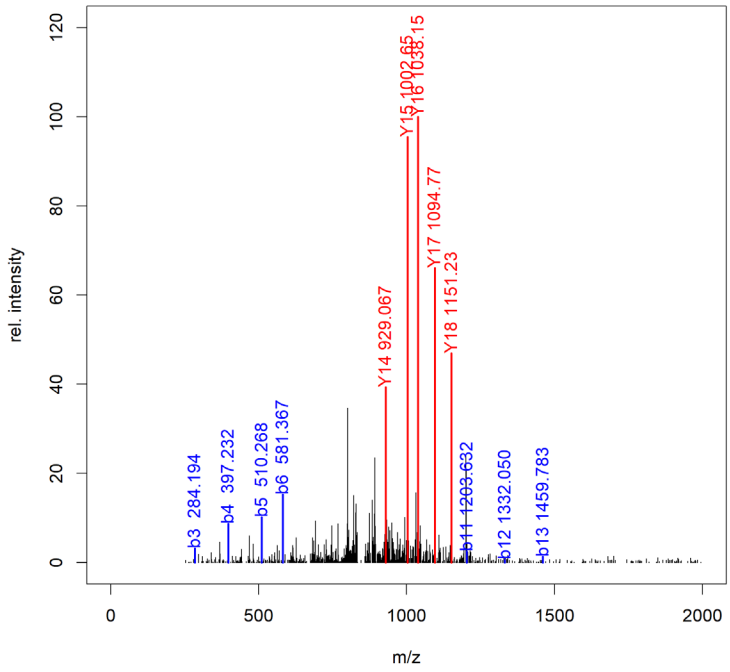


**Figure S4. 5** MS2 spectra of the two adducts that were new to this study: adducts 862.11 and 862.77.

***Adduct 862.11***

data02_138_238_PrecursorScan_3932_MZ_862.1133_RT_26.71_MS.MS.Scan_3933




***Adduct 862.77***

data02_59_279_PrecursorScan_4156_MZ_862.7728_RT_26.47_MS.MS.Scan_4159

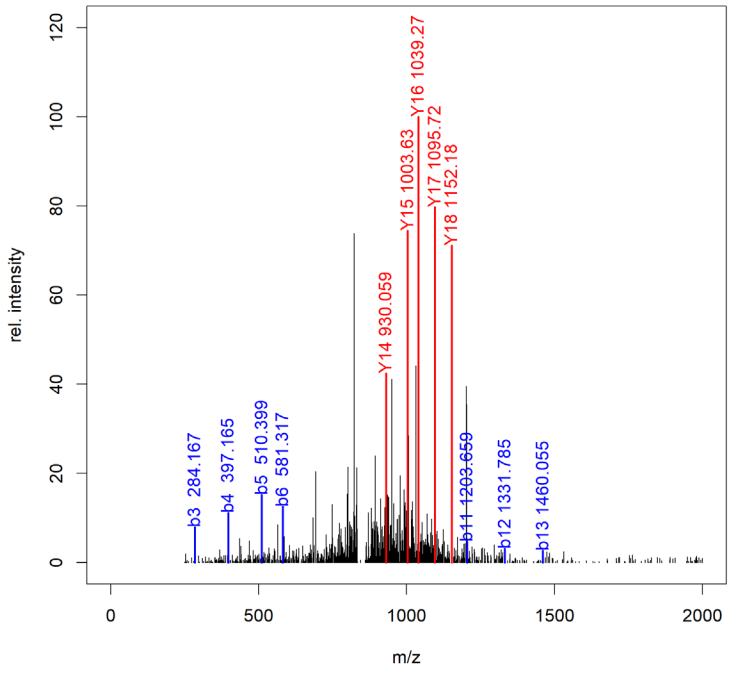


**Figure S4. 6** Variable selection for childhood leukemia cases and controls overall using (A) multivariable linear regression, (B) logistic lasso regression on bootstrapped data, and (C) random forest. (A) Volcano plot showing the case/control fold change for each adduct and the corresponding nominal *p*-value for each adduct. The dotted line represents a nominal *p*-value of 0.05. (B) Adducts ranked by the proportion of times each adduct was selected into the lasso model out of 500 bootstrap iterations. (C) The top 20 adducts ranked by random forest variable importance (*i.e.*, mean decrease in Gini index).

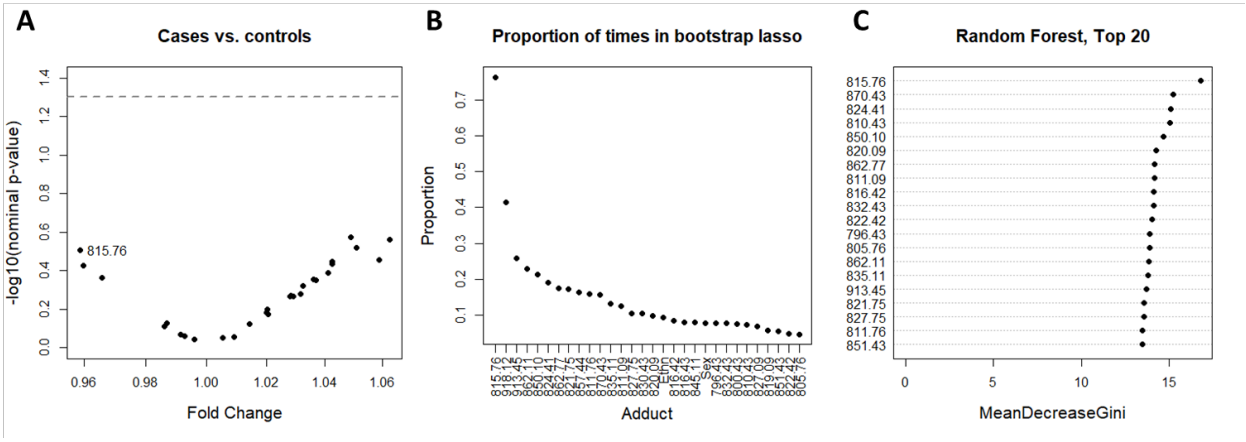



**Figure S4. 7** Variable selection for ALL cases and controls using (A) multivariable linear regression, (B) logistic lasso regression on bootstrapped data, and (C) random forest. (A) Volcano plot showing the case/control fold change for each adduct and the corresponding nominal *p*-value for each adduct. The dotted line represents a nominal *p*-value of 0.05. (B) Adducts ranked by the proportion of times each adduct was selected into the lasso model out of 500 bootstrap iterations. (C) The top 20 adducts ranked by random forest variable importance (*i.e.*, mean decrease in Gini index).

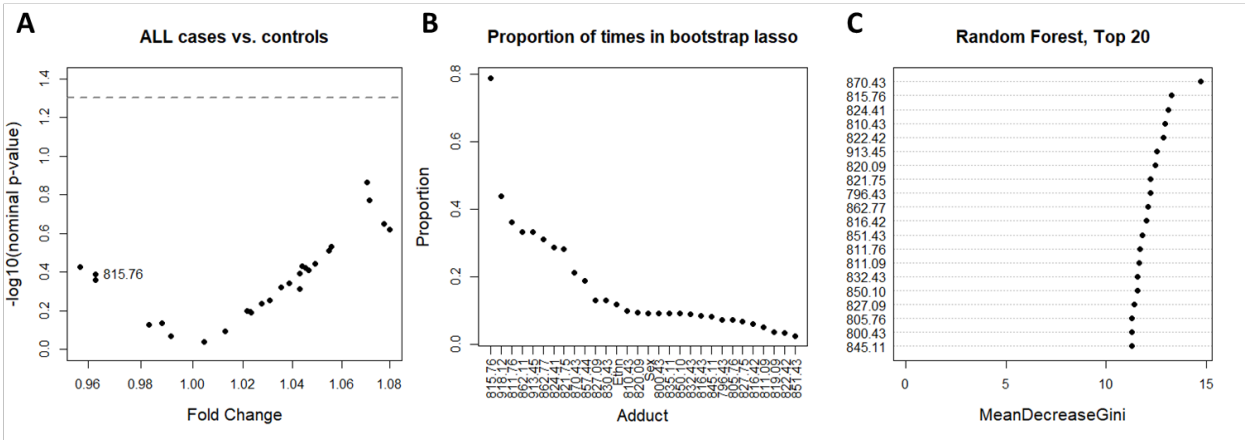


**Figure S4. 8** Variable selection for B-cell ALL cases and controls using (A) multivariable linear regression, (B) logistic lasso regression on bootstrapped data, and (C) random forest. (A) Volcano plot showing the case/control fold change for each adduct and the corresponding nominal *p*-value for each adduct. The dotted line represents a nominal *p*-value of 0.05. (B) Adducts ranked by the proportion of times each adduct was selected into the lasso model out of 500 bootstrap iterations. (C) The top 20 adducts ranked by random forest variable importance (*i.e.*, mean decrease in Gini index).

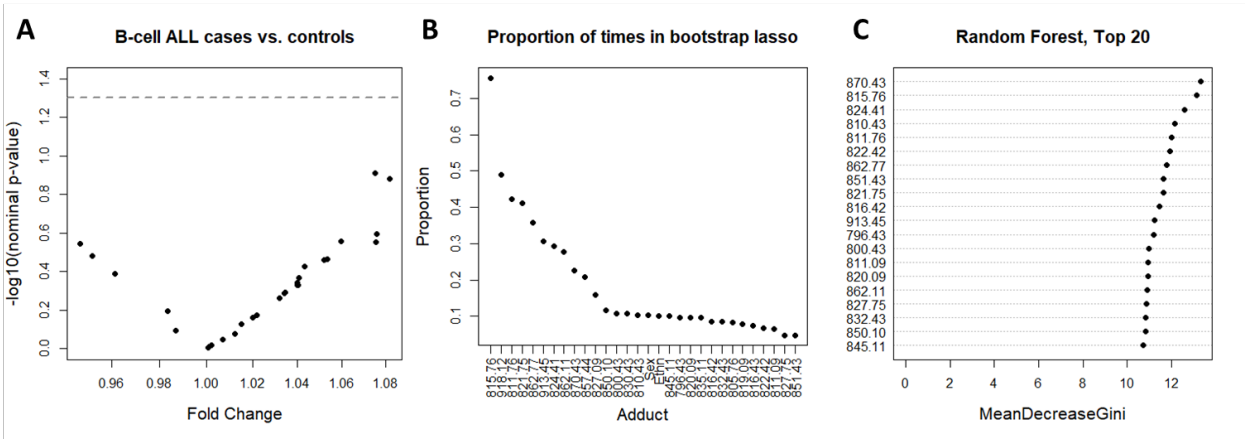



**Figure S4. 9** Variable selection for B-cell ALL cases with high-hyperdiploidy and controls using (A) multivariable linear regression, (B) logistic lasso regression on bootstrapped data, and (C) random forest. (A) Volcano plot showing the case/control fold change for each adduct and the corresponding nominal *p*-value for each adduct. The dotted line represents a nominal *p*-value of 0.05. (B) Adducts ranked by the proportion of times each adduct was selected into the lasso model out of 500 bootstrap iterations. (C) The top 20 adducts ranked by random forest variable importance (*i.e.*, mean decrease in Gini index).

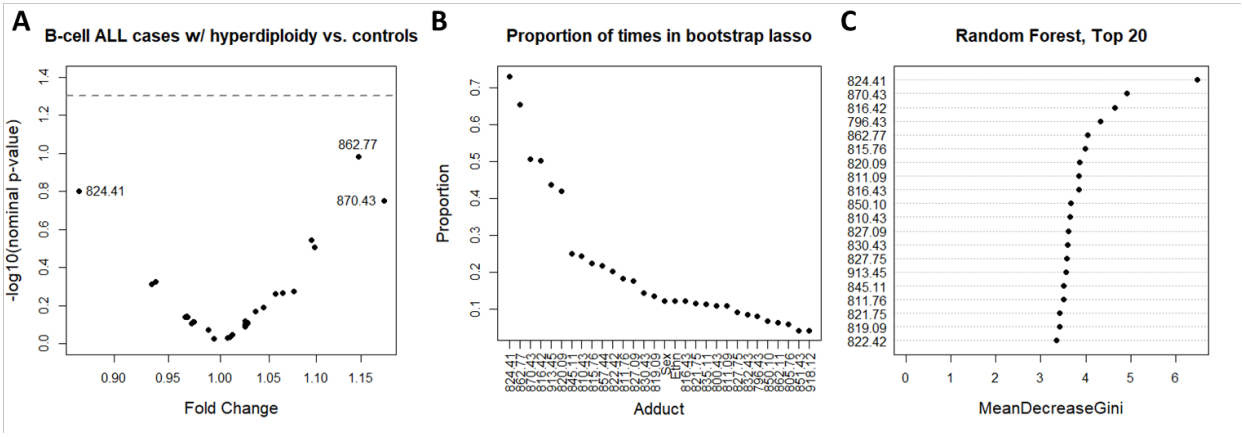


**Figure S4. 10** Variable selection for B-cell ALL cases with t(12;21) translocation and controls using (A) multivariable linear regression, (B) logistic lasso regression on bootstrapped data, and (C) random forest. (A) Volcano plot showing the case/control fold change for each adduct and the corresponding nominal *p*-value for each adduct. The dotted line represents a nominal *p*-value of 0.05. (B) Adducts ranked by the proportion of times each adduct was selected into the lasso model out of 500 bootstrap iterations. (C) The top 20 adducts ranked by random forest variable importance (*i.e.*, mean decrease in Gini index).

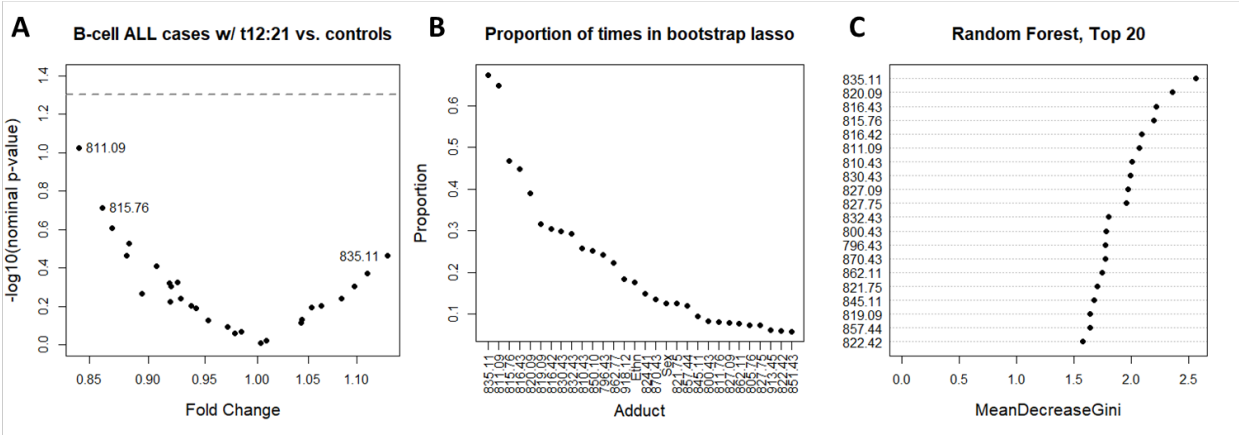
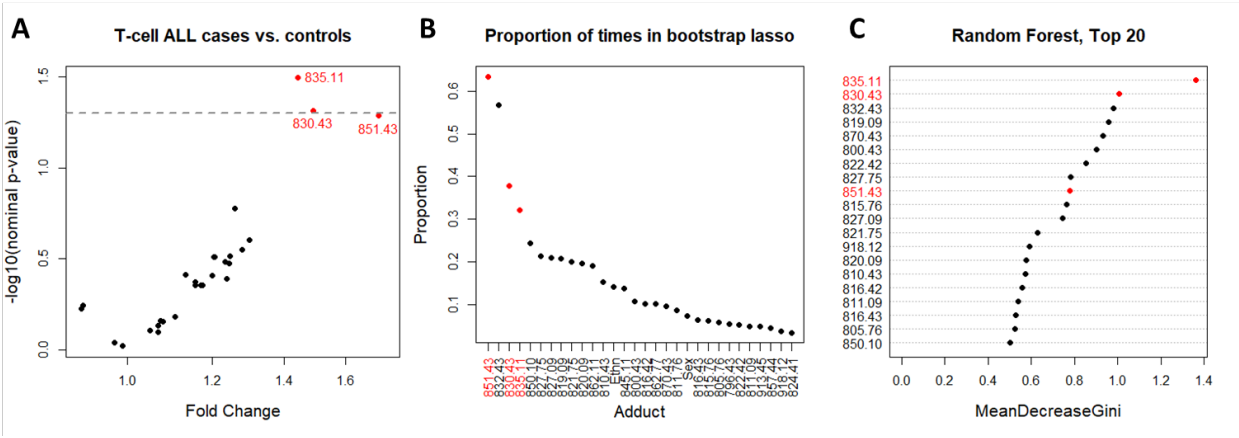



**Figure S4. 11** Variable selection for T-cell ALL cases and controls using (A) multivariable linear regression, (B) logistic lasso regression on bootstrapped data, and (C) random forest. (A) Volcano plot showing the case/control fold change for each adduct and the corresponding nominal *p*-value for each adduct. The dotted line represents a nominal *p*-value of 0.05. (B) Adducts ranked by the proportion of times each adduct was selected into the lasso model out of 500 bootstrap iterations. (C) The top 20 adducts ranked by random forest variable importance (*i.e.*, mean decrease in Gini index).



### 4.10.3 Supplemental Tables

**Table S4. 1** Top-10 ranking adducts from variable selection for childhood leukemia cases and controls overall. The top-10 adducts based on (1) nominal *p*-values from multivariable linear regression, with the corresponding fold change (FC); (2) proportion of times selected into the lasso model; and (3) the mean decrease in Gini index from random forest variable importance are shown.

| | **Linear regression** | | | **lasso** | | **Random forest** | |
|---|---|---|---|---|---|---|---|
| **Rank** | Adduct | p-value | FC | Adduct | Proportion | Adduct | Gini |
| 1 | 862.77 | 0.27 | 1.05 | 815.76 | 0.76 | 815.76 | 16.77 |
| 2 | 857.44 | 0.27 | 1.06 | 918.12 | 0.41 | 870.43 | 15.20 |
| 3 | 821.75 | 0.30 | 1.05 | 913.45 | 0.26 | 824.41 | 15.07 |
| 4 | 815.76 | 0.31 | 0.96 | 862.11 | 0.23 | 810.43 | 15.03 |
| 5 | 870.43 | 0.35 | 1.06 | 850.10 | 0.21 | 850.10 | 14.68 |
| 6 | 835.11 | 0.36 | 1.04 | 824.41 | 0.19 | 820.09 | 14.27 |
| 7 | 827.75 | 0.37 | 1.04 | 862.77 | 0.17 | 862.77 | 14.18 |
| 8 | 918.12 | 0.37 | 0.96 | 821.75 | 0.17 | 811.09 | 14.14 |
| 9 | 830.43 | 0.41 | 1.04 | 857.44 | 0.16 | 816.42 | 14.11 |
| 10 | 913.45 | 0.44 | 0.97 | 811.76 | 0.16 | 832.43 | 14.10 |

**Table S4. 2** Top-10 ranking adducts from variable selection for ALL cases and controls overall. The top-10 adducts based on (1) nominal *p*-values from multivariable linear regression, with the corresponding fold change (FC); (2) proportion of times selected into the lasso model; and (3) the mean decrease in Gini index from random forest variable importance are shown.

| | **Linear regression** | | | **lasso** | | **Random forest** | |
|---|---|---|---|---|---|---|---|
| **Rank** | Adduct | p-value | FC | Adduct | Proportion | Adduct | Gini |
| 1 | 862.77 | 0.14 | 1.07 | 815.76 | 0.79 | 870.43 | 14.70 |
| 2 | 821.75 | 0.17 | 1.07 | 918.12 | 0.44 | 815.76 | 13.27 |
| 3 | 857.44 | 0.22 | 1.08 | 811.76 | 0.36 | 824.41 | 13.10 |
| 4 | 870.43 | 0.24 | 1.08 | 862.11 | 0.33 | 810.43 | 12.92 |
| 5 | 796.43 | 0.29 | 1.06 | 913.45 | 0.33 | 822.42 | 12.84 |
| 6 | 830.43 | 0.31 | 1.05 | 862.77 | 0.31 | 913.45 | 12.53 |
| 7 | 850.10 | 0.36 | 1.05 | 824.41 | 0.29 | 820.09 | 12.46 |
| 8 | 845.11 | 0.37 | 1.04 | 821.75 | 0.28 | 821.75 | 12.22 |
| 9 | 918.12 | 0.38 | 0.96 | 870.43 | 0.21 | 796.43 | 12.19 |
| 10 | 835.11 | 0.38 | 1.05 | 857.44 | 0.19 | 862.77 | 12.08 |

**Table S4. 3** Top-10 ranking adducts from variable selection for B-cell ALL cases and controls. The top-10 adducts based on (1) nominal *p*-values from multivariable linear regression, with the corresponding fold change (FC); (2) proportion of times selected into the lasso model; and (3) the mean decrease in Gini index from random forest variable importance are shown.

| | **Linear regression** | | | **lasso** | | **Random forest** | |
|---|---|---|---|---|---|---|---|
| **Rank** | Adduct | p-value | FC | Adduct | Proportion | Adduct | Gini |
| 1 | 862.77 | 0.12 | 1.08 | 815.76 | 0.76 | 870.43 | 13.30 |
| 2 | 821.75 | 0.13 | 1.08 | 918.12 | 0.49 | 815.76 | 13.12 |
| 3 | 857.44 | 0.25 | 1.08 | 811.76 | 0.42 | 824.41 | 12.58 |
| 4 | 850.10 | 0.28 | 1.06 | 821.75 | 0.41 | 810.43 | 12.15 |
| 5 | 870.43 | 0.28 | 1.08 | 862.77 | 0.36 | 811.76 | 11.98 |
| 6 | 918.12 | 0.29 | 0.95 | 913.45 | 0.31 | 822.42 | 11.92 |
| 7 | 913.45 | 0.33 | 0.95 | 824.41 | 0.29 | 862.77 | 11.76 |
| 8 | 820.09 | 0.34 | 1.05 | 862.11 | 0.28 | 851.43 | 11.62 |
| 9 | 796.43 | 0.35 | 1.05 | 870.43 | 0.23 | 821.75 | 11.62 |
| 10 | 845.11 | 0.38 | 1.04 | 857.44 | 0.21 | 816.42 | 11.44 |

**Table S4. 4** Top-10 ranking adducts from variable selection for B-cell ALL cases with high-hyperdiploidy and controls. The top-10 adducts based on (1) nominal *p*-values from multivariable linear regression, with the corresponding fold change (FC); (2) proportion of times selected into the lasso model; and (3) the mean decrease in Gini index from random forest variable importance are shown.

| | **Linear regression** | | | **lasso** | | **Random forest** | |
|---|---|---|---|---|---|---|---|
| **Rank** | Adduct | p-value | FC | Adduct | Proportion | Adduct | Gini |
| 1 | 862.77 | 0.10 | 1.15 | 824.41 | 0.73 | 824.41 | 6.47 |
| 2 | 824.41 | 0.16 | 0.87 | 862.77 | 0.65 | 870.43 | 4.92 |
| 3 | 870.43 | 0.18 | 1.18 | 870.43 | 0.51 | 816.42 | 4.64 |
| 4 | 845.11 | 0.29 | 1.09 | 816.42 | 0.50 | 796.43 | 4.33 |
| 5 | 816.42 | 0.31 | 1.10 | 913.45 | 0.44 | 862.77 | 4.04 |
| 6 | 913.45 | 0.47 | 0.94 | 820.09 | 0.42 | 815.76 | 3.99 |
| 7 | 820.09 | 0.48 | 0.93 | 845.11 | 0.25 | 820.09 | 3.87 |
| 8 | 857.44 | 0.53 | 1.08 | 810.43 | 0.24 | 811.09 | 3.85 |
| 9 | 830.43 | 0.54 | 1.06 | 815.76 | 0.22 | 816.43 | 3.85 |
| 10 | 800.43 | 0.55 | 1.06 | 857.44 | 0.22 | 850.10 | 3.67 |

**Table S4. 5** Top-10 ranking adducts from variable selection for B-cell ALL cases with t(12;21) translocation and controls. The top-10 adducts based on (1) nominal *p*-values from multivariable linear regression, with the corresponding fold change (FC); (2) proportion of times selected into the lasso model; and (3) the mean decrease in Gini index from random forest variable importance are shown.

| | **Linear regression** | | | **lasso** | | **Random forest** | |
|---|---|---|---|---|---|---|---|
| **Rank** | Adduct | p-value | FC | Adduct | Proportion | Adduct | Gini |
| 1 | 811.09 | 0.09 | 0.84 | 835.11 | 0.67 | 835.11 | 2.56 |
| 2 | 815.76 | 0.19 | 0.86 | 811.09 | 0.65 | 820.09 | 2.36 |
| 3 | 816.42 | 0.25 | 0.87 | 815.76 | 0.47 | 816.43 | 2.22 |
| 4 | 862.77 | 0.30 | 0.88 | 816.43 | 0.45 | 815.76 | 2.20 |
| 5 | 796.43 | 0.34 | 0.88 | 820.09 | 0.39 | 816.42 | 2.09 |
| 6 | 835.11 | 0.34 | 1.13 | 819.09 | 0.32 | 811.09 | 2.07 |
| 7 | 822.42 | 0.39 | 0.91 | 816.42 | 0.30 | 810.43 | 2.01 |
| 8 | 819.09 | 0.43 | 1.11 | 830.43 | 0.30 | 830.43 | 1.99 |
| 9 | 805.76 | 0.47 | 0.93 | 832.43 | 0.29 | 827.09 | 1.98 |
| 10 | 845.11 | 0.48 | 0.92 | 810.43 | 0.26 | 827.75 | 1.96 |

**Table S4. 6** Top-10 ranking adducts from variable selection for T-cell ALL cases and controls. The top-10 adducts based on (1) nominal *p*-values from multivariable linear regression, with the corresponding fold change (FC); (2) proportion of times selected into the lasso model; and (3) the mean decrease in Gini index from random forest variable importance are shown.

| | **Linear regression** | | | **lasso** | | **Random forest** | |
|---|---|---|---|---|---|---|---|
| **Rank** | Adduct | p-value | FC | Adduct | Proportion | Adduct | Gini |
| 1 | 835.11 | 0.03 | 1.44 | 851.43 | 0.63 | 835.11 | 1.36 |
| 2 | 830.43 | 0.05 | 1.49 | 832.43 | 0.57 | 830.43 | 1.00 |
| 3 | 851.43 | 0.05 | 1.72 | 830.43 | 0.38 | 832.43 | 0.98 |
| 4 | 800.43 | 0.17 | 1.26 | 835.11 | 0.32 | 819.09 | 0.96 |
| 5 | 827.75 | 0.25 | 1.30 | 850.10 | 0.24 | 870.43 | 0.93 |
| 6 | 918.12 | 0.28 | 1.28 | 827.75 | 0.21 | 800.43 | 0.90 |
| 7 | 913.45 | 0.31 | 1.25 | 827.09 | 0.21 | 822.42 | 0.85 |
| 8 | 805.76 | 0.31 | 1.20 | 819.09 | 0.21 | 827.75 | 0.78 |
| 9 | 827.09 | 0.31 | 1.21 | 821.75 | 0.20 | 851.43 | 0.78 |
| 10 | 811.09 | 0.33 | 1.23 | 820.09 | 0.20 | 815.76 | 0.76 |

**Table S4. 7** Top-10 ranking adducts from variable selection for AML cases and controls. The top-10 adducts based on (1) nominal *p*-values from multivariable linear regression, with the corresponding fold change (FC); (2) proportion of times selected into the lasso model; and (3) the mean decrease in Gini index from random forest variable importance are shown.

| | **Linear regression** | | | **lasso** | | **Random forest** | |
|---|---|---|---|---|---|---|---|
| **Rank** | Adduct | p-value | FC | Adduct | Proportion | Adduct | Gini |
| 1 | 850.10 | 0.01 | 0.66 | 850.10 | 0.69 | 835.11 | 2.56 |
| 2 | 851.43 | 0.05 | 0.65 | 827.09 | 0.59 | 850.10 | 2.40 |
| 3 | 845.11 | 0.26 | 0.85 | 800.43 | 0.37 | 827.09 | 2.01 |
| 4 | 870.43 | 0.35 | 0.83 | 851.43 | 0.36 | 811.09 | 1.99 |
| 5 | 816.42 | 0.36 | 0.89 | 862.77 | 0.30 | 815.76 | 1.93 |
| 6 | 918.12 | 0.37 | 0.90 | 811.76 | 0.29 | 800.43 | 1.87 |
| 7 | 862.77 | 0.39 | 0.88 | 811.09 | 0.28 | 862.77 | 1.83 |
| 8 | 811.09 | 0.43 | 0.92 | 832.43 | 0.27 | 845.11 | 1.76 |
| 9 | 913.45 | 0.45 | 0.92 | 835.11 | 0.27 | 816.42 | 1.71 |
| 10 | 821.75 | 0.56 | 0.91 | 816.42 | 0.26 | 832.43 | 1.70 |

# Chapter 5.

# Conclusions and future directions

As the global disease burden continues to shift from communicable to non-communicable diseases, particularly cancer, we need to better understand the etiologic factors and underlying disease mechanisms that may lead to disease prevention. In the last few decades, advances in sequencing technologies have allowed for in-depth explorations of genetic factors involved in cancer etiology. Similar tools were not readily available for the exploration of exposures, and epidemiological studies often relied on questionnaires and interviews, which are prone to measurement errors and reporting biases. However, technological advances in analytical instruments such as liquid chromatography-high resolution mass spectrometry along with high-dimensional bioinformatics/statistical tools have opened new avenues for comprehensive omics investigations of exposures impacting cancer risks. Since many of the etiologic factors of cancer remain unknown, it is anticipated that omics investigations will reveal unrecognized risk factors and enhance our understanding of the causes of cancer. Our untargeted HSA-Cys34 adductomics method is one such omics approach that allows for the exploration of both known and unknown exposures to reactive electrophiles contributing to the etiology of cancer.

The overarching goal of the work described here was to perform HSA-Cys34 adductomics using archived newborn DBS to discover *in utero* exposures associated with childhood leukemia. We first developed an adductomics method tailored for the analysis of newborn DBS to overcome the unique challenges associated with DBS, which included sample complexity and variation in blood volume across DBS. Normalization of blood volume using hemoglobin measurements of the DBS extracts via UV-Vis spectroscopy was very effective and could easily be integrated into other future omics analyses utilizing DBS (Chapter 2). Sample processing for adductomics measurements was designed to allow for high-throughput analyses and uses a simple solvent extraction to isolate HSA in the DBS extracts (Chapter 3). We validated this methodology with 49 archived DBS collected from newborns with mothers who actively smoked or were nonsmokers during pregnancy. In this pilot study, we saw that the Cys34 adduct of cyanide showed consistent discrimination between newborns of smoking and nonsmoking mothers. Hydrogen cyanide is in fact a constituent of cigarette smoke, and these results indicated that the DBS-adductomics method was suitable for exploring *in utero* exposures to reactive electrophiles. While our DBS-adductomics method was developed for the analysis of 4.7-mm newborn DBS punches preserved at -20℃, this method is also applicable for other DBS analyses of HSA-Cys34 adducts.

We then applied the DBS-adductomics method to analyze 783 archived newborn DBS collected from childhood leukemia cases and matched population-based controls participating in the California Childhood Leukemia Study (Chapter 4). This was the first large-scale adductomics analysis of archived newborn DBS. Our novel normalization workflow was optimized for the dataset and adjusted for factors of unwanted variation that were unique to DBS-adductomics, including batch effects, amount of digested HSA, instrument performance, DBS age, and blood volume. Since large numbers of samples are analyzed over long periods of time in many omics studies, our results indicate that it is important to minimize sources of unwanted variation prior

to data collection, and to adjust for this technical variation and normalize data before performing statistical analyses.

In comparing HSA-Cys34 adducts between childhood leukemia cases and controls, we identified several adducts that warrant further investigation. Interestingly, the Cys34 homocysteine adduct with loss of $H_2O$ was shown to consistently discriminate AML cases and controls. We hypothesized that the altered levels of homocysteine may be related to changes in epigenetic regulations or folate status among AML cases. No consistent associations were found for childhood leukemia cases and controls overall, as well as the total ALL and total B-cell ALL subgroups. However, when the B-cell ALL subgroup was further stratified into B-cell ALL with high-hyperdiploidy and B-cell ALL with t(12;21) subgroups, the Cys34 adducts of potassium, sulfenamide, and crotonaldehyde showed modest discrimination between cases and controls. Likewise, for the T-cell ALL subgroup, Cys34 adducts of crotonaldehyde, acrolein, and cysteine showed moderate discrimination. Because this was a hypothesis-generating study, and some of the subgroups were limited in sample size, our findings will need to be replicated in future studies.

The etiology of childhood leukemia, and cancer in general, is very complex. In order to obtain a more cohesive view of the etiology of childhood leukemia, our adductomics results should be complemented by other omics investigations. Our findings for B-cell ALL with t(12;21) and T-cell ALL point to the involvement of reactive oxygen and carbonyl species that have long been implicated in the progression of ALL.[1,2] Regarding AML, our results point to the possible involvement of the folate-mediated one-carbon metabolic pathway. Since the adductome is but one component of the biological system that is affected by exposures, our results would benefit from knowledge of effects to other omics layers in the same DBS specimens, including the epigenome, transcriptome, proteome, and the metabolome.

It is important to emphasize that our study used archived newborn DBS that were collected within 48 h of birth. While HSA encompasses exposures occurring one month prior to birth, this single time point may not have captured all relevant exposures in the etiology of childhood leukemia. It is possible that the first hit may have occurred at other critical time periods such as the earlier weeks of gestation. Furthermore, given the long latency spanning up to 14 years, it is also possible that early markers of disease manifestation or susceptibility may not be easily detectable at birth. Despite this, we were still able to detect differences in levels of particular adducts between ALL/AML cases and controls in archived newborn DBS collected years before diagnosis. Future integration of our results with multi-omic investigations of childhood leukemia may reveal biological changes related to leukemogenesis, and thereby present new opportunities for preventing this tragic disease.

## References


(1) Battisti, V.; Maders, L. D. K.; Bagatini, M. D.; Santos, K. F.; Spanevello, R. M.; Maldonado, P. a.; Brulé, A. O.; Araújo, M. D. C.; Schetinger, M. R. C.; Morsch, V. M. Measurement of Oxidative Stress and Antioxidant Status in Acute Lymphoblastic Leukemia Patients. *Clin. Biochem.* **2008**, *41*, 511–518.

(2) Hole, P. S.; Darley, R. L.; Tonks, A. Do Reactive Oxygen Species Play a Role in Myeloid Leukemias? *Blood* **2011**, *117* (22), 5816–5826.