UNIVERSITY OF CALIFORNIA,
IRVINE

DNA methylation markers for glyphosate exposure and mammographic density

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Epidemiology

by

Rachel McFarland Lucia

Dissertation Committee:
Associate Professor Hannah Lui Park, Chair
Assistant Professor Trina M. Norden-Krichmar
Associate Professor Luohua Jiang

2021

Dedication

To my daughters, who are worth every sleepless night and missed deadline.

Upon the hearth the fire is red,
Beneath the roof there is a bed;
But not yet weary are our feet,
Still round the corner we may meet
A sudden tree or standing stone
That none have seen but we alone.

Still round the corner there may wait
A new road or a secret gate,
And though we pass them by today,
Tomorrow we may come this way
And take the hidden paths that run
Towards the Moon or to the Sun.

J.R.R. Tolkien, *The Fellowship of the Ring*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACD, acid citrate dextrose
AMPA, aminomethylphosphonic acid
ANOVA, analysis of variance
ASA24, Automated Self-Administered 24-hour Dietary Assessment Tool
AUC, area under the receiver operating characteristic curve
BI-RADS, Breast Imaging Reporting and Data System
BMI, body mass index
BMIQ, beta mixture quantile normalization
CI, confidence interval
DHQII, Diet History Questionnaire II
DMP, differentially methylated probe
DMR, differentially methylated region
EDTA, ethylenediaminetetraacetic acid
EPA, Environmental Protection Agency
EPSPS, 5-enolpyruvylshikimate-3-phosphate synthase
FDA, Food and Drug Administration
FDR, false discovery rate
FFQ, food frequency questionnaire
FNDDS, Food and Nutrient Database for Dietary Studies
GO, gene ontology
GR, glyphosate-resistant
HEI, Healthy Eating Index
HRT, hormone replacement therapy
IARC, International Agency for Research on Cancer
ICC, intraclass correlation
IQR, interquartile range
Kcal, kilocalories
KEGG, Kyoto Encyclopedia of Genes and Genomes
LASSO, least absolute shrinkage and selection operator
LC-MS/MS, liquid chromatography with tandem mass spectrometry
LOD, limit of detection
LOQ, limit of quantification
MEE, Markers for Environmental Exposures
SD, standard deviation
SOC, Standard Occupational Classification
UCI, University of California, Irvine
USDA, United States Department of Agriculture
WHO, World Health Organization

Acknowledgements

My thesis chair, Hannah Lui Park, has my profound gratitude for her support and mentorship over the past five years. She has been my biggest academic advocate, my coach and my cheerleader, my mentor and my friend.

I would also like to thank my committee members, Trina Norden-Krichmar (always available for a laugh and a reality check) and Luohua Jiang, who have been an indispensable source of guidance and support as I worked to complete this dissertation.

I have heavily relied upon the expertise of our study team: the investigators (Argyrios Ziogas, Trina Norden-Krichmar, Deborah Goodman, Irene Masunaka and Andrea Alvarez), my fellow graduate student (Wei-Lin Huang: a bottomless well of commiseration, support, and good ideas), and our other students (Melodie Elyasian, Danielle Forman, Amanda Hidajat, Alexis Kim, Daphne Thampy, and Kailynn Yang), whose hard work was essential to our success.

I am profoundly grateful to the women who participated in these studies for their enthusiastic contributions to the advancement of scientific understanding.

Finally, my deepest gratitude to my family. To my parents, who have encouraged me in countless big and small ways, including to pursue this degree when I was sure that I was done with school forever, and my brother Ben, my first role model, who always made sure I did my reading. To my husband Peter, who has been steadfastly by my side and always believed in me when I doubted myself. Last but far from least, to my daughters, Ellie and Kate, for the joy.

Versions of Chapters 1 and 4 have been previously published as follows:

> Lucia RM, Huang W, Alvarez A, Thampy D, Elyasian M, Hidajat A, Yang K, Forman D, Pebdani A, Masunaka I, Brain S, Heditsian D, Lee V, Goodman D, Norden-Krichmar TM, Odegaard AO, Ziogas A, Park HL (2020). Rationale, study design, and cohort characteristics for the Markers for Environmental Exposures (MEE) Study. International Journal of Environmental Research and Public Health, 17(5):1774.
>
> Lucia RM, Huang W-L, Alvarez A, Masunaka I, Ziogas A, Goodman D, Odegaard AO, Norden-Krichmar TM, Park HL (2021). Association of mammographic density with blood DNA methylation. Epigenetics. 1-16.

I was the primary author on both publications, performed all data analyses, and drafted both manuscripts with support and feedback from my coauthors. Parts of the text, tables, and figures are reproduced here under the open-access license.

Financial support was provided by the California Breast Cancer Research Grants Program Office of the University of California, 22UB-2311, and the Chao Family Comprehensive Cancer Center. This reported research includes work performed in the mass spectrometry core supported by the National Cancer Institute of the National Institutes of Health under grant number P30CA033572. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Curriculum Vitae

Rachel McFarland Lucia

2013	B.S. in Biomedical Engineering Johns Hopkins University Honors Graduate
2014 – 2017	Clinical Research Assistant Ambry Genetics
2017 – 2018	Teaching Assistant, Department of Developmental and Cell Biology University of California, Irvine
2017 – 2021	Research Assistant, Department of Epidemiology and Biostatistics University of California, Irvine
2020	Teaching Assistant, Program in Public Health University of California, Irvine
2021	Research Fellow, Global Epidemiology AbbVie
2021	Ph.D. in Epidemiology University of California, Irvine

PUBLICATIONS

Lucia RM, Huang W-L, Alvarez A, Masunaka I, Ziogas A, Goodman D, Odegaard AO, Norden-Krichmar TM, Park HL (2021). Association of mammographic density with blood DNA methylation. Epigenetics. 1-16.

Matin J, **Lucia RM**, Lal K, Columbus A, Goodman D, Larsen KM, Ziogas A, Park HL (2021). Factors associated with women's unwillingness to decrease alcohol intake to decrease breast cancer risk. Journal of Primary Care and Community Health, 12.

Lucia RM, Huang W, Alvarez A, Thampy D, Elyasian M, Hidajat A, Yang K, Forman D, Pebdani A, Masunaka I, Brain S, Heditsian D, Lee V, Goodman D, Norden-Krichmar TM, Odegaard AO, Ziogas A, Park HL (2020). Rationale, study design, and cohort characteristics for the Markers for Environmental Exposures (MEE) Study. International Journal of Environmental Research and Public Health, 17(5):1774.

Yang X, Leslie G, Doroszuk A, Schneider S, Allen J, [… and 111 others, including **McFarland R**], Tischkowitz M (2020). Cancer risks associated with germline PALB2 pathogenic variants: an international study of 524 families. Journal of Clinical Oncology, 38(7):674-685.

Lu HM, Li S, Black MH, Lee S, Hoiness R, Wu S, Mu W, Huether R, Chen J, Sridhar S, Tian Y, **McFarland R**, Dolinsky J, Tippin Davis B, Mexal S, Dunlop C, Elliott A (2019). Association of breast and ovarian cancers with predisposition genes identified by large-scale sequencing. JAMA Oncology, 5(1):51-7.

Rana HQ, Gelman R, LaDuca H, **McFarland R**, Dalton E, Thompson J, Speare V, Dolinsky JS, Chao EC, Garber JE (2018). Differences in TP53 mutation carrier phenotypes emerge from panel-based testing. Journal of the National Cancer Institute, 110(8):863-870.

LaDuca H, **McFarland R** [co-first author], Gutierrez S, Yussuf A, Ho N, Pepper J, Reineke P, Cain T, Blanco K, Horton C, Dolinsky JS (2018). Quality of clinician-reported cancer history when ordering genetic testing. JCO Clinical Cancer Informatics, 2: 1-11.

Espenschied CR, LaDuca H, Li S, **McFarland R**, Gau CL, Hampel H (2017). Multigene panel testing provides a new perspective on Lynch syndrome. Journal of Clinical Oncology, 35(22):2568-75.

Couch FJ, Shimelis H, Hu C, Hart SN, Polley EC, Na J, Hallberg E, Moore R, Thomas A, Lilyquist J, Feng B, **McFarland R**, Pesaran T, Huether R, LaDuca H, Chao EC, Goldgar DE, Dolinsky JS (2017). Associations between cancer predisposition testing panel genes and breast cancer. JAMA Oncology, 3(9):1190-6.

Pritzlaff M, Summerour P, **McFarland R**, Li S, Reineke P, Dolinsky JS, Godgar DE, Shimelis H, Couch FJ, Chao EC, LaDuca H (2017). Male breast cancer in a multi-gene panel testing cohort: insights and unexpected results. Breast Cancer Research and Treatment, 161(3), 575-586.

Leedom TP, LaDuca H, **McFarland R**, Li S, Dolinsky JS, Chao EC (2016). Breast cancer risk is similar for CHEK2 founder and non-founder mutation carriers. Cancer Genetics, 209(9), 403-7.

Kapoor NS, Curcio LD, Blakemore CA, Bremner AK, **McFarland R**, West JG, Banks KC (2015). Multigene panel testing detects equal rates of pathogenic BRCA1/2 mutations and has a higher diagnostic yield compared to limited BRCA1/2 analysis alone in patients at risk for hereditary breast cancer. Annals of Surgical Oncology, 22(10), 3282-8.

CONFERENCE PRESENTATIONS

Lucia RM, Huang W, Alvarez A, Masunaka I, Goodman D, Norden-Krichmar TM, Odegaard AO, Ziogas A, Park HL. Epigenome-wide association study of mammographic density. Poster session presented at: American Society of Human Genetics Annual Meeting; 2020 Oct 27-30; Virtual meeting.

Lucia RM, Huang W, Alvarez A, Masunaka I, Goodman D, Norden-Krichmar TM, Odegaard AO, Ziogas A, Park HL. Association of mammographic density with blood DNA methylation. Poster session presented at: American Academy of Cancer Research Annual Meeting; 2020 Jun 22-24; Virtual meeting.

Lucia RM, Huang W, Pathak KV, McGilvrey M, Alvarez A, Ziogas A, Goodman D, Odegaard AO, Pirrotte P, Norden-Krichmar TM, Park HL. Epigenome-wide association study of the herbicide glyphosate and DNA methylation. Poster session presented at: American Society of Human Genetics Annual Meeting; 2019 Oct 15-19; Houston, TX.

Lucia RM, Alvarez A, Masunaka I, Ziogas A, Goodman D, Odegaard A, Lee V, Brain S, Heditsian D, Norden-Krichmar TM, Park HL. A pilot study of DNA methylation, epigenetic age, and breast cancer risk status. Poster session presented at: American Society of Human Genetics Annual Meeting; 2018 Oct 16-20; San Diego, CA.

McFarland R, LaDuca H, Li S, Dolinsky JS. Positive predictors of inherited cancer susceptibility among women with ovarian and endometrial cancer. Poster session presented at: 47th Annual Society of Gynecologic Oncology Annual Meeting on Women's Cancer; 2016 Mar 19-22; San Diego, CA.

Abstract of the Dissertation

DNA methylation markers for glyphosate exposure and mammographic density

by

Rachel McFarland Lucia

Doctor of Philosophy in Epidemiology

University of California, Irvine, 2021

Associate Professor Hannah Lui Park, Chair

Glyphosate is the most commonly used herbicide in the world. Animal studies and epidemiologic evidence suggest that there may be adverse health effects of exposure to glyphosate. Glyphosate and its primary metabolite aminomethylphosphonic acid (AMPA) are frequently detected in a wide variety of foods, but studies of dietary factors associated with these pesticide residues are limited. Many previous studies have identified epigenetic signatures of environmental exposures and other risk factors for disease. DNA methylation can provide a tool for understanding the molecular impact of exposures and may serve as a biomarker for past exposures. However, no studies to date have examined the relationship between blood DNA methylation and urinary glyphosate and AMPA. Mammographic density, an important risk factor for breast cancer, may also have epigenetic impacts, but previous research on the subject has been uninformative.

Three hundred ninety-two postmenopausal women responded to diet, health, and environmental questionnaires and provided biospecimens including blood and two urine samples. Diet was characterized using up to three 24-hour dietary recalls to capture intake of specific food groups and overall diet quality. Site-specific DNA methylation in blood was

measured using the Illumina Infinium MethylationEPIC BeadChip. Urinary glyphosate and AMPA were measured using LC-MS/MS. Mammographic density was obtained from the most recent mammogram report. Using linear mixed-effects models, I explored relationships between dietary characteristics, demographic/behavioral factors, and urinary glyphosate and AMPA concentrations. After standard methylation data filtering and preprocessing, CpG sites throughout the genome were examined for associations with urinary glyphosate and AMPA concentrations and with mammographic density.

Grain consumption was associated with higher urinary glyphosate concentrations, even among women who reported often or always eating organic grains. Soy and alcohol consumption and increased frequency of fast-food dining were associated with greater AMPA, while fruit and corn consumption were associated with lower AMPA. Methylation at 24 CpG sites was associated with urinary glyphosate concentration. These sites were combined into a methylation index which accurately predicted glyphosate concentration in an internal validation cohort. Glyphosate- and AMPA-associated methylation occurred near genes associated with cancer (*SF3B2, MSH4*, and *TRIM31*) and endocrine disruption (*ESR1*), and AMPA was associated with greater epigenetic age acceleration. I identified 250 CpG sites and 37 regions for which DNA methylation was significantly associated with mammographic density. The top sites were located within genes associated with cancer, and were more likely to be located in regulatory regions of the genome.

This is the largest study to date examining paired dietary data and urinary glyphosate and AMPA and identified possible sources of glyphosate and AMPA in the American diet, including grains, soy, and alcoholic beverages. Glyphosate and AMPA exposure are associated with DNA methylation differences that could promote the development of cancer,

informing the hypothesis that glyphosate and/or AMPA exposure could elevate the risk for disease. After further validation, these methylation differences may be developed into a biomarker of glyphosate exposure. This research paints a picture of epigenetic dysregulation associated with mammographic density and suggests the potential involvement of several genes in the biological mechanisms behind differences in breast density between women.

Introduction

DNA methylation and other epigenetic mechanisms are major regulators of gene expression and phenotype. Methylation usually occurs at CpG sites, where a cytosine nucleotide is adjacent to a guanine. The impact of DNA methylation depends on the genomic context; methylation at gene promoters is negatively correlated with gene expression, while methylation within gene bodies can be positively or negatively correlated with gene expression[1]. Dysregulation of DNA methylation is a key feature of cancer. A consistent pattern of changes in DNA methylation occur in cancer cells: global hypomethylation throughout the genome, paired with hypermethylation at the promoters of tumor-suppressor genes[2–6]. This pattern has also been observed in peripheral blood and may precede cancer development. Several studies have identified a pattern of global hypomethylation associated with later development of breast cancer[7–10]. In addition, some commercial tests or "liquid biopsies" utilize DNA methylation-based biomarkers for cancer detection, including for lung cancer[11] and for colorectal cancer[12,13].

Considerable attention has been given to identifying epigenetic signatures of environmental exposures and other risk factors for disease. Blood DNA methylation at several sites, most notably within the *AHRR* and *F2RL3* genes, has been consistently associated with smoking in numerous studies[14–16]. Biologic age can also be estimated using DNA methylation, and accelerated biologic age is associated with numerous diseases, including breast cancer[17–19]. Other studies have examined the epigenetic impact of body mass index[20,21], menopause[22,23], age at menarche[24], and environmental exposures such as endocrine-disrupting compounds[25,26], pesticides[27], and air pollution[28–30]. Most studies attempt to identify individual CpG sites associated with the exposure of interest. A promising

approach to develop exposure biomarkers by combining information from multiple sites into DNA methylation indices has been applied to smoking and alcohol consumption[31–33].

Identifying epigenetic signatures of environmental and lifestyle exposures has two important roles. First, epigenetic signatures may fulfill an important role as a biological "memory" of past exposures. For example, former smokers typically have DNA methylation levels somewhere between current and never smokers, with a gradual return to methylation patterns consistent with never smokers over a period of at least 10 years[31,34]. Other research suggests that the epigenetic impacts of transient exposures can persist for many years. For example, maternal smoking during pregnancy has been associated with differences in DNA methylation in the child that persist into later childhood[35].

Second, identifying the specific genes whose methylation is perturbed may provide a deeper understanding of the mechanism by which exposures influence risk for disease. In some previous research, epigenetic biomarkers of exposures have been extended to predict disease incidence, suggesting that DNA methylation serves as a biological mechanism of conferring risk. For example, methylation levels at 3 sites previously associated with smoking accurately predicted lung cancer incidence in a prospective nested case-control study[36]. In fact, the methylation-derived index outperformed self-reported smoking status at identifying incident lung cancer cases (AUC 0.80 for methylation index vs. 0.72 for smoking status), particularly for light smokers. This could be due to bias in self-reported smoking status; however, another possible explanation is that DNA methylation in peripheral blood serves as an intermediate biologic phenotype which precedes the development of disease. Other studies have supported this hypothesis by showing that methylation at smoking-associated loci is strongly associated with lung cancer incidence, even after adjustment for

smoking status[37,38], and that DNA methylation differences associated with breast cancer risk predate the development of cancer by many years[8]. Under this framework, exploring differentially methylated genes and pathways associated with an exposure could provide important insight into the molecular mechanisms by which the exposure affects risk for disease.

This work aims to explore the epigenetic impact of two distinct risk factors for disease: exposure to the herbicide glyphosate, and mammographic breast density. The long-term goal is to understand the epigenetic dysregulation that occurs in response to environmental insults and to develop DNA methylation signatures for known or potential breast cancer risk factors. Examining the functional significance of the DNA methylation patterns associated with these factors will clarify the molecular mechanisms by which they may influence the development of cancer and other diseases.

The broad-spectrum herbicide glyphosate is the most used pesticide in the world[39]. In agriculture, it is used for weed control, reduction of soil tillage, preparation of seed beds, and pre-harvest desiccation of grain crops and legumes[39,40] in addition to non-agricultural uses in residential areas, parks, forests, pastures, and lawns[41]. Use of glyphosate-based herbicides has increased dramatically since the introduction of glyphosate-resistant crop species in the late 1990s[40], although they are also commonly used on conventional crops[39,41,42] . From 1996 to 2014, the total volume of glyphosate applied in the U.S. increased over 9-fold[39]. This increase in glyphosate use is likely accompanied by increases in human exposure to glyphosate and its primary metabolite, aminomethylphosphonic acid (AMPA). Both compounds are frequently detected in urine samples from non-farming populations. In the U.S., recent studies detected glyphosate in 70-90% of participants[43–46].

The potential health effects of glyphosate on humans are widely debated. In 2015 the World Health Organization's (WHO) International Agency for Research on Cancer (IARC) determined glyphosate to be a probable human carcinogen[47], while other agencies including the U.S. Environmental Protection Agency (EPA) disagree with this assessment[48,49]. Recently, a pilot study on 250 women in Hawaii found that higher AMPA levels were associated with breast cancer risk[50]. In agricultural workers, glyphosate has been associated with kidney toxicity[51] and increased risk for non-Hodgkin's lymphoma[52]. Glyphosate has also been shown to induce fatty liver disease[53], endocrine disruption[54], and changes to the microbiome[55–57] in animal models, and the growth of breast cancer cells in vitro[58]. However, the potential health effects of glyphosate and AMPA are not well-characterized in human populations and thus remain controversial. A deeper understanding of the impacts of glyphosate and AMPA exposure on the epigenome could suggest pathways that may be impacted by exposure to these compounds and enhance our understanding of their health effects.

Furthermore, assessment of exposure to glyphosate and AMPA is challenging. First, glyphosate has a short half-life of 8-10 hours in humans[59,60], making it difficult to obtain an accurate estimate of chronic exposure through testing of biospecimens. Glyphosate and AMPA exposure are also difficult to assess by self-report in the general population, who are most likely to be exposed to these compounds through residues on food. A recent study found that urinary glyphosate and AMPA concentrations decreased by more than 70% after just 5 days of a completely organic diet[61]. Both compounds have been detected on a wide variety of foods, including soybeans[62], corn[62], wheat and cereals[63–65], and many others[63,64,66–68]. It is unclear which, if any, dietary factors are associated with glyphosate and AMPA levels,

and thus it is difficult for individuals who are not occupationally exposed to estimate their level of exposure. Identifying DNA methylation markers which are associated with glyphosate and/or AMPA levels could allow for more robust estimation of an individual's past exposure to glyphosate and AMPA.

One key step towards understanding risk factors for glyphosate and AMPA exposure is to identify characteristics associated with greater concentrations of these compounds. These may be demographic or behavioral characteristics; of particular interest are features of the diet, because the general population is typically exposed to glyphosate and AMPA through the diet. There is a paucity of research to identify dietary characteristics associated with elevated glyphosate and AMPA levels. Studies in Germany[69] and Slovenia[70] showed positive relationships between glyphosate levels and the consumption of pulses, nuts, and brown rice. However, both the typical diet and the presence of glyphosate and AMPA in the food supply are likely to differ between Europe and the United States, limiting the generalizability of these studies to American populations. The only studies to examine dietary factors associated with urinary glyphosate or AMPA in the U.S. found associations between urinary glyphosate and consumption of tea[68] and caffeinated beverages[46]. However, these studies were conducted in targeted populations of undergraduate students and pregnant women, again limiting the applicability of their findings to the population at large. More research on the subject is urgently needed given the increasing presence of glyphosate and AMPA in the food supply, leading to my first research aim.

Aim 1: Determine dietary and other characteristics associated with elevated glyphosate and AMPA levels

To achieve this aim, we recruited a cohort of nearly 400 postmenopausal women

residing in Southern California, who provided two urine samples paired with 24-hour dietary recalls recording their food and beverage intake for a 24-hour period. Because of the short half-lives of glyphosate and AMPA, I examined dietary intakes of major food groups and specific foods and beverages previously associated with or suspected to be associated with glyphosate levels over the day preceding a urine sample. Linear mixed models were used to determine which, if any, dietary, demographic, and behavioral factors were associated with self-reported organic eating and glyphosate and AMPA concentrations in urine. This study suggests possible sources of glyphosate and AMPA in the U.S. food supply and identifies other factors that are associated with elevated glyphosate and/or AMPA concentrations in urine.

After completing these initial studies exploring risk factors for glyphosate exposure, I conducted the first epigenome-wide association study of urinary glyphosate and AMPA in order to improve exposure assessment and our understanding of the molecular impact of glyphosate and AMPA exposure.

Aim 2: Identify a DNA methylation signature of urinary glyphosate and AMPA

In order to accomplish this aim, I used paired DNA methylation and urinary glyphosate and AMPA measurements in a cohort of nearly 400 postmenopausal women residing in Southern California. Average glyphosate and AMPA concentrations from two urine samples were examined for associations with DNA methylation at 850,000 CpG sites measured with the Illumina HumanMethylationEPIC BeadChip. After identifying individual CpG sites and regions whose methylation is associated with glyphosate or AMPA, I developed a methylation index to predict glyphosate and AMPA levels. This index was tested in a separate validation set. The functional and genomic context of the glyphosate-associated

sites was examined for enrichment of pathways which could promote the development of disease. This study is the first to examine the epigenetic impact of glyphosate and AMPA in humans and is the first step towards developing DNA methylation-based biomarkers of glyphosate and AMPA exposure, which could improve exposure assessment for these herbicides.

The framework applied to the development of epigenetic biomarkers of glyphosate and AMPA can be extended to other exposures of particular interest for breast cancer research, which leads to my next aim.

Aim 3: Identify a DNA methylation signature of mammographic breast density

Mammographic breast density, or the extent of radiologically dense fibroglandular tissue appearing on a mammogram, varies between women and is largely heritable[71,72], but also influenced by an individual's lifestyle[73–75] and exposure to exogenous hormones or other drugs[76,77]. Elevated mammographic density is one of the strongest risk factors for breast cancer, with women in the highest density category exhibiting at least a 4-fold increase in breast cancer risk compared to women in the lowest density category[78–80]. Some evidence suggests that mammographic density mediates the associations between breast cancer risk factors such as history of breast biopsy, nulliparity, age at first birth, and hormone therapy, and breast cancer risk[81–83]. However, our understanding of how mammographic density influences the molecular environment and the mechanism by which it affects cancer risk is incomplete. Identifying epigenetic differences which are associated with elevated mammographic density may provide an enhanced understanding of the mechanism by which mammographic density influences cancer risk. This could open the door for novel interventions and improve risk stratification by identifying women who are

at the greatest risk for mammographic density-mediated breast cancer.

The epigenetic impacts of mammographic density have not been extensively studied. One study in a cohort of Australian women did not identify any loci associated with mammographic density at genome-wide significance[84]. A more recent study utilized prediagnostic data from women who subsequently developed breast cancer and also did not identify any CpG sites which were significant at the genome-wide level[85]. However, they identified 140 regions which were differentially methylated in women with high- vs. low-density breasts. These are the only previous studies of mammographic density and DNA methylation, and are either uninformative or applicable only to a specific subpopulation. Thus, further studies of the relationship between mammographic density and DNA methylation are needed.

To achieve this aim, I leveraged paired mammographic density and DNA methylation measurements in a cohort of 385 postmenopausal women with no personal history of breast cancer. DNA methylation was quantified using the Illumina HumanMethylationEPIC BeadChip, which measures DNA methylation levels at over 850,000 sites throughout the genome, and the four-category BI-RADS mammographic density was obtained from the most recent mammogram report. Using a novel resampling-based approach to decrease the chance of type I errors, I identified individual CpG sites and genomic regions whose methylation is associated with elevated mammographic density. The methodology was validated using well-characterized relationships between DNA methylation and smoking status as a positive control. The genes and other genomic features located near the density-associated sites and regions were examined to determine their functional significance. This study advances our understanding of the molecular impact of mammographic breast density

and suggest pathways which may be influenced by the presence of dense breast tissue.

Taken together, these research aims will contribute to our knowledge of epigenetic responses to environmental exposures and aid in the development of DNA methylation-based biomarkers.

Chapter 1: Rationale, Design, and Cohort Characteristics of the Markers for Environmental Exposures (MEE) Study

1.1 Introduction

Studies suggest that human cancers are due to a combination of factors, including heredity, random mutations, and environmental factors[86–90]. In addition, environmental factors, including exposure to tobacco smoke, alcohol consumption, obesity, inadequate physical activity, radiation, and chemical exposures, have been shown to be associated with many other major chronic diseases, including cardiovascular disease[91], stroke[92], chronic obstructive pulmonary disease[93], and diabetes[94] as well as other conditions ranging from infertility to neurodegenerative diseases[95,96]. While some of these factors are well-established risk factors for disease, the evidence for many others is insufficient.

Reliable estimates of exposure are needed to analyze the potential relationships between exposures and health. Self-reporting is commonly used to assess environmental exposures in epidemiological research. However, compared to direct measurements, such as biomarker measurement, self-reported data are prone to recall bias, conscious or unconscious. For example, studies have found that about 20% of pregnant women with cotinine levels indicative of current smoking failed to report their smoking habits when asked[97,98]. For many other environmental exposures, self-report is difficult because individuals may be unaware of the dose, frequency, or time frame of exposure. Biomarkers can serve as objective markers for environmental exposures which may be used to improve our understanding of how exposures relate to health.

In recent years, interest has mounted in identifying epigenetic markers for exposures, or markers based on differences in non-sequence modifications of the DNA[99–101]. DNA

methylation is the best-studied epigenetic mechanism in this area and has several properties that make it ideal for marker development. First, it has a desirable balance between plasticity (allowing a biologic response to environmental insults) and long-term stability (allowing the assessment of past exposures). This has allowed the identification of DNA methylation-based biomarkers of numerous environmental exposures, most notably cigarette smoking, which has been consistently associated with methylation at sites within the *AHRR* and *F2RL3* genes in numerous studies[14,15,102]. In one study, *AHRR* methylation predicted current smoking status with an area under the curve (AUC) of 0.99[103]. Studies of smoking have also demonstrated the long-term stability of DNA methylation biomarkers, with one study showing that former smokers may return to a "nonsmoking" state over a period of at least 10 years[31]. Another useful quality is that DNA methylation can be relatively easily measured from blood, saliva, or tissue DNA, including banked DNA, which will allow leveraging of existing large cohort studies and biorepositories in order to identify associations of environmental exposures with health outcomes. Specific patterns of DNA methylation have been identified for some other environmental exposures, including air pollution[104,105] and polycyclic aromatic hydrocarbons[106,107], but the impact of many environmental exposures on the epigenome remains unknown.

This chapter describes the design, methods, and participants of the Markers for Environmental Exposures (MEE) Study, a cross-sectional study conducted from 2017–2019 comprised of 392 postmenopausal women aged 45 to 66 years in Southern California. Questionnaire data, medical record data, and biospecimens were collected from each participant. The repository of data and biospecimens (and molecular profiles obtained from them) can be linked to future cancer registry data and other publicly available data. This

conceptual framework will support the long-term vision of the MEE Study to (1) identify biomarkers, especially epigenetic markers, of environmental exposures, such as pesticides, air/food/water contaminants, and industrial chemicals that are commonly encountered in the general population; and (2) support the study of potential relationships between environmental exposures and health and health-related factors.

1.2 Methods

Study design

The MEE Study is a cross-sectional study with potential for follow-up data collection and linkage. The study was approved by the University of California, Irvine (UCI) Institutional Review Board (HS# 2016-3127), and the investigators worked closely with a panel of three patient and research advocates from the community (Susie Brain, Diane Heditsian, and Vivian Lee). While the study was originally designed with the intention of identifying epigenetic markers for pesticide exposure and factors related to breast cancer risk, the data and specimens collected can potentially be used in the context of other diseases, conditions, and environmental exposures. All participants provided written consent to participate in the study and HIPAA authorization.

The study workflow is outlined in Figure 1.1. Participants were asked to complete six online questionnaires, provide biospecimens, and provide authorization to access their breast health-related medical records (Table 1.1). The six questionnaires consisted of an environmental exposures questionnaire, a breast health questionnaire, three 24-hour dietary recalls completed on three separate days (two weekdays and one weekend day), and a food frequency questionnaire. The biospecimens consisted of two self-collected first-void urine samples, a venous peripheral blood sample, and an optional saliva sample. Study

participants received a nutrition report based on their 24-hour dietary recalls and a report of their urinary levels of certain pesticides (glyphosate and organophosphates); participants received no monetary compensation.

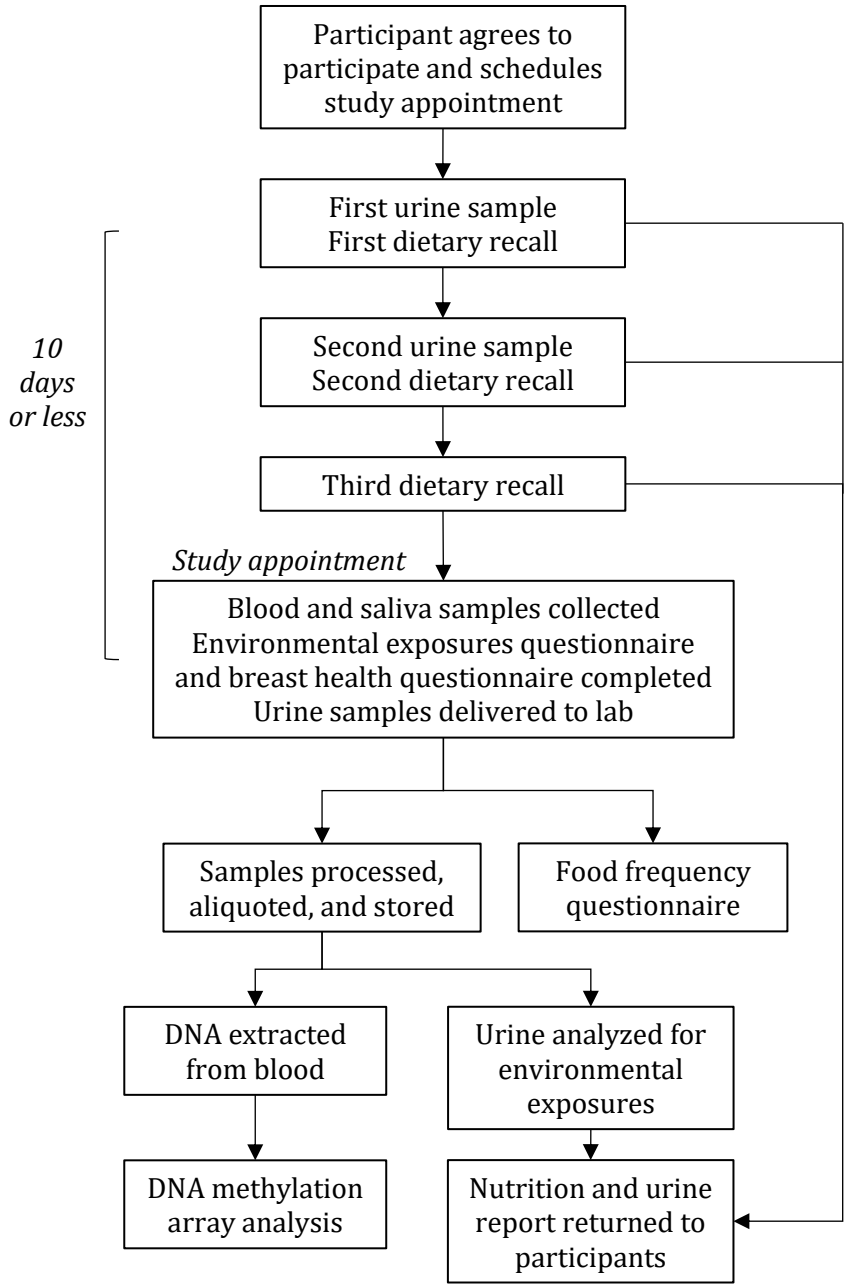

Figure 1.1. MEE Study workflow

Table 1.1. Data and specimens collected

Data Sources	Description
Environmental exposures questionnaire	Environmental exposures, source of drinking water, organic eating behaviors, residence history, occupation
Breast health questionnaire	Personal medical history, reproductive history, family history of cancer, demographic data
Dietary recalls (ASA24)	Complete report of all food, drink, and supplements consumed the previous day (3 recalls were requested)
Food frequency questionnaire (DHQII)	Summary of dietary intake frequencies over the previous year
Electronic medical record	Mammogram reports
Specimens Collected	**Description**
Blood	Peripheral blood, separated into serum, plasma, and buffy coat; DNA extracted from buffy coat
Urine	Two first-void urine samples
Saliva	Optional saliva sample

Population

Four hundred (400) participants completed the study protocol from October 2017 to May 2019 (Figure 1.2). Participants were female residents of primarily Orange County, CA. Women were eligible to participate in the study if they were aged 45–66 years, postmenopausal (had not had their period in over one year), and had never been diagnosed with breast cancer or had a mastectomy. Eligibility criteria were based on the study's main aim of identifying epigenetic markers for pesticide exposure and factors related to breast cancer risk. Age is known to be associated with epigenetic changes[17,18] as well as breast cancer risk[108], and menopause has also been shown to be associated with epigenetic changes[23]. Women with a personal history of breast cancer or mastectomy were excluded in order to capture the population at risk for incident breast cancer. After completion of study recruitment, the menopause status of eight participants could not be determined, leaving a final study cohort of 392 women.

There were two primary methods of study recruitment (Figure 1.2). The first ("targeted recruitment") involved personal mailings to patients who had received a screening mammogram at a University of California, Irvine Health facility and had consented to be contacted for future research studies as part of the Athena Breast Health Network[109]. For women who were eligible based on their most recent breast health questionnaire, three attempts were made to contact them using varied methods (telephone, mail, and email). Participants were also recruited from the broader community ("community recruitment") through email announcements to UCI employees; announcements/flyers posted on social media, local community boards, and at events; and via word-of-mouth. This method relied on potential participants initiating contact with the study team after seeing the announcement or flyer. Translated materials were available for patients whose primary language was Spanish. The study coordinator was fluent in both English and Spanish.

Specimen collection, processing, and storage

After enrolling in the study and scheduling a blood collection appointment, participants were mailed study materials including printed instructions in English or Spanish and two sterile urine collection cups labeled with their study ID. The 4 oz (118 mL) collection cups were individually sealed, made of high-density polyethylene plastic and had a spill-proof, leak-resistant screw-on sealing cap. A security seal on the cap prevented any tampering. Participants were advised that during the 10-day period prior to their scheduled blood collection appointment, they would be contacted on two separate days at 7:30 PM to ask them to self-collect their first-void urine samples the following morning and place the filled collection cups in a freezer (−20°C) until their appointment. Contact was via text

message or email based on participant preference. Participants were not told which days they would be contacted, ensuring capture of their typical behaviors and exposures.

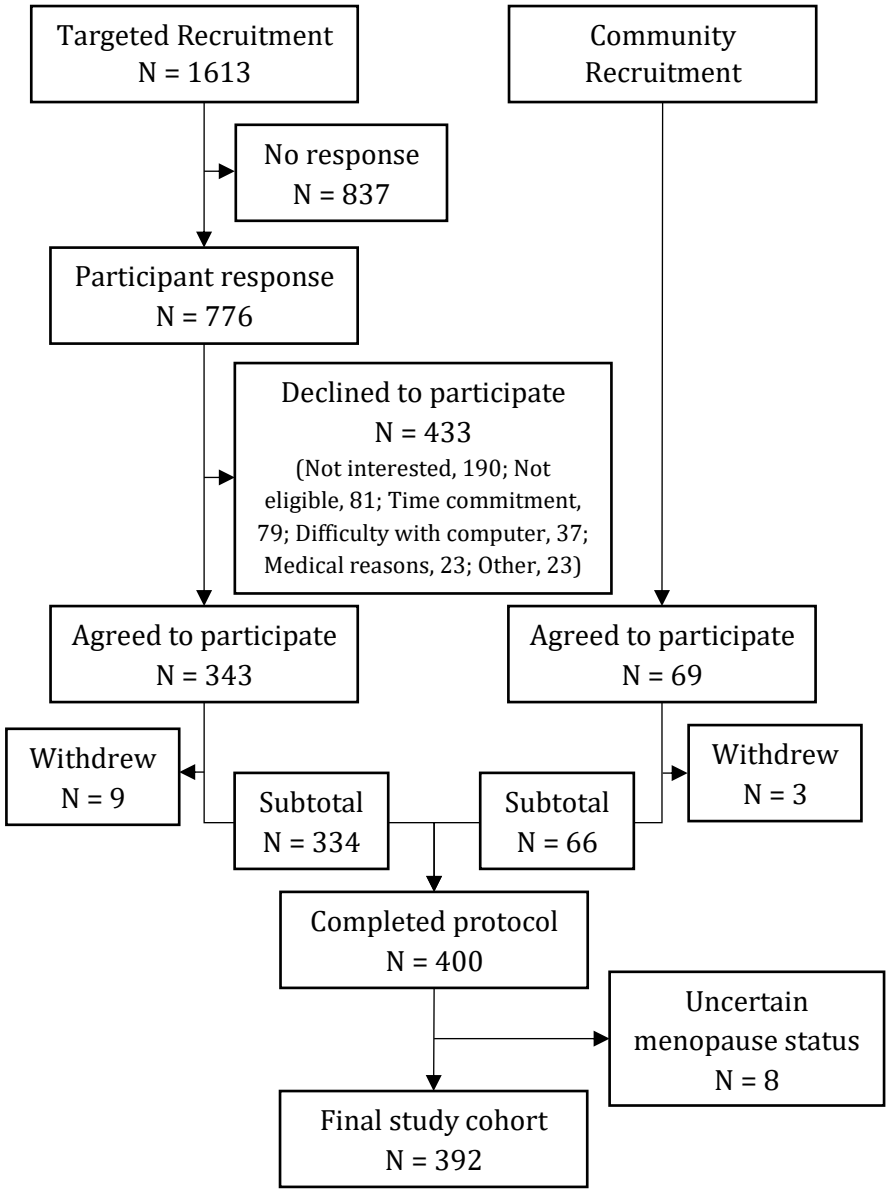

Figure 1.2. Flow chart of study recruitment
Recruitment for the MEE Study used two primary methods: personalized mailings to members of a breast cancer screening cohort ("Targeted Recruitment") and general announcements in the community ("Community Recruitment").

Peripheral blood was collected at the participant's home or other location of the participant's choice by the study coordinator, who was also a certified phlebotomist. Three

glass Vacutainer® blood collection tubes were used: one containing ethylenediaminetetraacetic acid (EDTA), one containing acid citrate dextrose (ACD), and one with no anticoagulants. Blood tubes were wrapped in an insulated bag and placed in a portable cooler with ice so the samples were cold but did not reach the point of freezing, while the coordinator was in the field. At the time of the blood collection appointment, the study coordinator also collected the previously frozen urine samples and placed them in the portable cooler. Upon arrival at the lab, the urine samples were stored at 4°C to thaw overnight; the next day, the samples were aliquoted and stored at −80°C for future analysis.

Within six hours of blood collection, blood samples were centrifuged at 2000 rpm for 10 min and isolated for serum (from tube with no anticoagulant), plasma (from EDTA and ACD tubes, separately) and buffy coat (from EDTA and ACD tubes, pooled). The serum and plasma were divided into 2 mL aliquots and stored at −80°C for future use. DNA was extracted from the buffy coat using the QIAamp DNA Blood Maxi Kit (Cat No. 51194, QIAGEN, Hilden, Germany). Extracted DNA was quantitated using the Qubit™ dsDNA HS Assay (Cat No. Q32854, ThermoFisher Scientific, Waltham, MA, USA) and the BioTek Synergy HT (BioTek, Winooski, VT, USA) microplate reader, and stored at −80°C. Optional saliva samples were collected from the participants in 15 mL Falcon tubes and stored at −80 °C for future analysis.

Questionnaire data collection and processing

Data on environmental exposures and breast health were collected using electronic questionnaires on the Research Electronic Data Capture (REDCap) and Salesforce platforms. The REDCap electronic data capture tool was hosted at the University of California, Irvine and is a secure, web-based application designed to support data capture for research

studies[110]. Salesforce is a secure, cloud-based customer relationship management system which is also used for research questionnaire data collection[111]. Environmental exposure questions included those regarding participants' residential history, frequency of organic food consumption, frequency of dining out, smoking history, and hormone use history. Organic eating behaviors were assessed with both a general question asking "Do you eat organic food?" as well as food-group specific questions for fruit, vegetables, grains, meat, and dairy, with response choices for each question being seldom or never, sometimes, often or always, and don't know or not sure. This question is consistent with previous studies of organic eating behaviors[112,113]. Food-group specific questions also included an option for "I do not eat the food" to account for different dietary patterns. Organic food was defined in the questionnaire as food that is either labeled United States Department of Agriculture (USDA) Organic, purchased locally from an organic farm, grown without pesticides in a home garden, or raised on organic feed without hormones and without antibiotics. Participants provided their current residential address, previous cities or ZIP codes, and length of residence at each location. Geocoded addresses can be used to model residential exposure to air pollution, as previously described in similar studies[104,114], or other neighborhood-specific environmental factors, such as green space. Never smokers were defined as those who had not ever smoked regularly for 6 months or more. Questions regarding occupation and use of pesticides were added to the questionnaire approximately midway through study enrollment, so just 248 (63.3%) participants were asked these questions. Breast health questions included those regarding the participant's history of mammography screening, personal and family history of breast and other cancers, personal gynecologic history, alcohol consumption, and physical activity. All questionnaires used branching logic, and the full text of each questionnaire is

provided in Appendix A (Table A1). Given the strong association between mammographic density and breast cancer risk[78,80] and its response to exposures such as exogenous hormones[76] and BMI[74], mammographic density (BI-RADS categories A, B, C, and D) was also collected from participants' most recent mammogram reports.

Questionnaire data were reviewed and curated, and some variables were regrouped to improve clarity and conciseness. For example, for race/ethnicity, only 5.2% of participants indicated a race/ethnicity other than non-Hispanic white, Hispanic, or Asian. These participants were grouped into the category "Other," while those who indicated Unknown or left the field blank were considered missing. Occupations were classified into six groups based on the Standard Occupational Classification (SOC) from the U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, and O*NET, the U.S. Occupational Information Network[115,116]. Occupation classification was performed by two separate individuals (Wei-Lin Huang and Danielle Forman), and any discrepancies were resolved by a third-party adjudicator (Hannah Lui Park). Cities within Orange County were classified into either North or South Orange County[117]. For physical activity, participants were classified based on whether their weekly physical activity met the Physical Activity Guidelines for Americans, issued by The Department of Health and Human Services, which recommend at least 150 minutes per week of moderate-intensity aerobic activity, at least 75 minutes per week of vigorous-intensity aerobic activity, or an equivalent combination of moderate- and vigorous- intensity aerobic activity for adults[118].

Dietary data collection and processing

Participants were advised that during the 10-day period prior to their scheduled blood collection appointment, they would be contacted by the study team on three days at

10:00 AM to ask them to complete a 24-hour dietary recall for the previous day. Contact was via text message or email according to participant preference. Participants were not told which days they would be contacted, and requests were made to include one weekend day, in order to capture their typical diet. Two recalls were requested for the days preceding first-void urine samples so the data would be paired.

Dietary intake data for 24-hour recalls were collected and analyzed using the Automated Self-Administered 24-hour (ASA24) Dietary Assessment Tool, version 2016, developed by the National Cancer Institute (Bethesda, MD, USA)[119]. ASA24 is modeled on the USDA's Automated Multiple-Pass Method (AMPM) and collects data using a meal-based approach to evaluate the consumption of specific food groups in addition to overall diet by calculating a diet quality score[120] and is available in Spanish.

Participants were also asked about their usual food consumption, including portion sizes, during the previous 12 months using the Diet History Questionnaire (DHQ II), a freely available food frequency questionnaire (FFQ) developed by the National Cancer Institute[121]. English speakers received the web version of DHQ II and Spanish speakers received a paper version of the DHQ I since a paper version of the DHQ II was not available in Spanish. DHQ I has been validated in three studies[122–124]. Updates to the second version included additions of foods and supplements but were minimal such that validation findings are unlikely to be significantly modified. Diet*Calc Software (V1.4.3, National Cancer Institute, Bethesda, MD, USA)[125] was used to analyze the DHQ data and generate nutrient and food group intake estimates. Dietary recalls which had total kilocalories (Kcal) lower than 500 Kcal/day or higher than 3500 Kcal/day were removed; these are cutoffs commonly used to exclude outliers in diet data for epidemiologic studies[126].

DNA methylation data collection and processing

Genomic DNA samples (1000 ng) were bisulfite converted using the Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's recommendations at the University of Southern California Molecular Genomics Core Facility. Bisulfite-converted DNA was then used as a substrate for the Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA, USA), which measures methylation levels at over 850,000 sites throughout the genome[127]. After assessing sample quality using the standard controls included in the array, methylation array data were filtered and normalized using the recommended pre-processing steps for methylation array data[128].

Statistical analysis

Basic cohort characteristics were compared for the two recruitment methods (targeted and community recruitment) using the student's t-test for age and Fisher's exact test for all other variables. Fisher's exact test was used rather than the chi-squared test due to small counts in some cells of the contingency table, which may affect the precision of the chi-squared approximation.

1.3 Results

Specimen and questionnaire collection

Four hundred (400) postmenopausal women aged 45 to 66 years completed the study protocol (Table 1.2). After recruitment was completed, 8 participants were excluded for having an ambiguous menopause status, leaving a final study cohort of 392 women. All participants provided a blood sample and two urine samples. Over half of the participants (*n* = 219, 55.9%) provided the optional saliva sample. Some urine samples and dietary recalls

were not completed on the days requested because participants forgot to complete them, were unexpectedly out of town or unavailable, or due to temporary problems accessing the ASA24 system. In these cases, participants were asked to complete their urine collection or dietary recall the next day that was feasible. Some study appointments had to be rescheduled due to illness or other reasons. The average time between collection of the two urine samples was 2.5 days (SD 1.8, range 1–17 days). The average time between collection of the first urine sample and the blood sample was 4.3 days (SD 2.9, range 0–26 days). DNA was extracted from buffy coat of all 392 participants' blood samples.

All participants completed the breast health questionnaire and environmental exposures questionnaire. After excluding ASA24 dietary recalls with out-of-range Kcal, 371 participants (94.6%) had at least one dietary recall, and 67.9% of participants had all three requested recalls (Table 1.2). Based on the study design, two dietary recalls for each participant were intended to be paired with urine samples. 338 participants (86.2%) had at least one paired dietary recall and 229 participants (58.4%) had both paired dietary recalls. 259 participants (66.1%) completed the food frequency questionnaire. There were no significant differences in questionnaire completion rates between participants from the two recruitment methods (data not shown).

Participant demographics

Since this study had a specific recruiting age range, the average age of participants was 56.9 years with small variation (SD 4.5 years). The majority of participants were non-Hispanic white (65.8%) followed by Hispanic (17.9%), Asian (11.1%), and other race/ethnicity (5.2%). Over 50% of participants graduated from college or had more education. Of the participants who indicated their occupation (N = 248), 81.9% were

currently employed either full-time or part-time. Among these participants, 69.0% currently had an occupation related to management, business, science or arts, followed by sales and office-related occupations (18.7%) and service-related occupations (9.9%). Over 90% of participants lived in either North or South Orange County. A small percentage of participants (9.4%) were from Los Angeles, San Bernardino, Riverside and San Diego Counties (Table 1.2). There were no significant differences between participants from the two recruitment methods for age, race/ethnicity, or education (data not shown).

Table 1.2. Cohort characteristics for the MEE Study

Characteristic	Mean (SD) or N (%) Total N = 392
Age, years, mean (SD)	56.9 (4.5)
Race/Ethnicity, N (%)	
Non-Hispanic white	254 (65.8%)
Hispanic	69 (17.9%)
Asian	43 (11.1%)
Other	20 (5.2%)
Missing	6
Education, N (%)	
High school graduate or less	33 (8.5%)
Some college or technical school	86 (22.1%)
College graduate or more	271 (69.5%)
Missing	2
Occupation, N (%)	
Unemployed or disabled	10 (4.0%)
Homemaker	12 (4.8%)
Retired	23 (9.3%)
Employed (Full-time or Part-time)	203 (81.9%)
Missing	144
Current residence, N (%)	
Los Angeles County	27 (6.9%)
North Orange County	179 (45.7%)
South Orange County	176 (44.9%)
Other	10 (2.6%)
Missing	0
Diet questionnaires completed, N (%)	
ASA24 dietary recall	
None	21 (5.4%)
1	32 (8.2%)
2	73 (18.6%)
≥3	266 (67.9%)
Paired urine and ASA dietary recall*	
None	54 (13.8%)
1	109 (27.8%)
2	229 (58.4%)
Food frequency questionnaire	259 (66.1%)

*A dietary recall was considered to be "paired" with a urine sample if the recall was provided for the day immediately prior to the first-void urine sample.

Lifestyle and health history

The majority of cohort participants were never-smokers (72.9%) (Table 1.3). Only 17 participants (4.3%) currently smoked, while 90 participants (22.8%) previously smoked regularly. The majority of cohort participants (68.4%) consumed less than two alcoholic drinks per week or never consume alcohol, while 31.6% reported consuming two or more alcoholic drinks per week. Only 40.7% of participants met physical activity guidelines. The average body mass index (BMI) in the cohort was 26.8 kg/m^2.

About 80% of participants had given birth at least once, with a mean age of 27.7 years old at first birth. The average ages of menarche and menopause were 12.8 years and 48.7 years, respectively. 122 participants (31.1%) had either a hysterectomy or oophorectomy, with 62 (15.8%) having had both surgeries. Just over 20% of participants had first-degree relatives who had been diagnosed with breast cancer, while 3.4% had a first-degree family history of ovarian cancer. 61.3% of participants had heterogeneously or extremely dense breasts according to their most recent mammogram. The most commonly reported personal health conditions were back pain (N = 84, 21.4%), thyroid disease (N = 70, 17.8%), hypertension (N = 69, 17.6%), and depression (N = 50, 12.8%). Just over 10% (N = 44) of participants reported a previous diagnosis of cancer (besides breast cancer). The most commonly reported cancers were non-melanoma skin cancer (N = 10), melanoma (N = 8), cervical cancer (N = 7), and thyroid cancer (N = 6). Five participants (1.3%) had previously been diagnosed with ovarian cancer.

Table 1.3. Lifestyle factors and health histories among study participants

Characteristic	Mean (SD) or N (%) Total N = 392
Smoking status, N (%)	
Current smoker	17 (4.3%)
Former smoker	89 (22.8%)
Never smoker	285 (72.9%)
Missing	1
Alcohol consumption, N (%)	
Never	105 (27.0%)
Less than 2 drinks per week	161 (41.4%)
2–6 drinks per week	72 (18.5%)
7 or more drinks per week	51 (13.1%)
Missing	3
Weekly physical activity meets the Physical Activity Guidelines for Americans, N (%)	
No	223 (59.3%)
Yes	153 (40.7%)
Missing	16
BMI, kg/m^2, mean (SD)	26.8 (6.5)
Missing	0
Age of menarche, mean (SD)	12.8 (1.5)
Missing	1
Number of live births, N (%)	
0	85 (21.7%)
1	70 (17.9%)
2	140 (35.7%)
3	73 (18.6%)
> 3	24 (6.1%)
Missing	0
Age at first birth, for parous women (N = 307), mean (SD)	27.7 (6.2)
Missing	0
Age of menopause, mean (SD)	48.7 (6.1)
Missing	3
History of gynecologic surgery, N (%)	
Oophorectomy	84 (21.4%)
Hysterectomy	100 (25.5%)
None	270 (68.9%)
Missing	0

Table continues

Table 1.1. Data and specimens collected (continued)

Characteristic	Mean (SD) or N (%) Total N = 392
Hormone replacement therapy use, N (%)	
Never	248 (63.4%)
Previous	61 (15.6%)
Current	82 (21.0%)
Missing	1
Mammographic breast density, N (%)	
Almost entirely fatty	42 (10.9%)
Scattered fibroglandular densities	107 (27.8%)
Heterogeneously dense	160 (41.6%)
Extremely dense	76 (19.7%)
Missing	7
Family history of cancer in first-degree relatives, N (%)	
Breast cancer (invasive or ductal carcinoma in situ [DCIS])	81 (21.0%)
Ovarian cancer	13 (3.4%)
Missing	6

Environmental exposures

Participants were asked questions regarding environmental exposures through their daily activities; responses are summarized in Table 1.4. Frequency of organic food consumption in the cohort was mostly evenly distributed, with 39.1% of participants reporting "often or always" eating organic food and 31.7% reporting "seldom or never" eating organic food. Over 90% of participants primarily drank bottled or filtered water, while 9.2% drank tap water without a filter. Only 13.5% of participants had ever lived on a farm; 5.6% had lived on a farm for more than 10 years. Of the participants who were asked about their use of pesticides, 19.8% (*N* = 49 out of 248) reported using pesticides at home or work within the past week.

Table 1.4. Environmental factors among study participants

Characteristic	Mean (SD) or N (%) Total N = 392
Organic food consumption frequency, N (%)	
Seldom or never	124 (31.7%)
Sometimes	114 (29.2%)
Often or always	153 (39.1%)
Missing	1
Source of drinking water, N (%)	
Tap (unfiltered)	36 (9.2%)
Tap (filtered)	215 (55.0%)
Bottled water	139 (35.5%)
Don't know or not sure	1 (0.3%)
Missing	1
History of living on a farm, N (%)	
> 10 years	22 (5.6%)
≤ 10 years	31 (7.9%)
None	339 (86.5%)
Age when started living on a farm, among ever farm dwellers (N = 53), mean (SD)	8.2 (9.7)
Used pesticides at home or workplace within past 7 days, N (%)	
No	173 (69.8%)
Yes	49 (19.8%)
Don't know or not sure	26 (10.5%)
Missing	144

1.4 Discussion

Environmental exposures are ubiquitous in daily life and have been associated with a multitude of diseases. However, many environmental exposures are difficult to objectively assess due to inaccurate self-reported data or challenges with biomarker measurement[101]. The development of blood-based biomarkers for environmental exposures will allow us to more accurately study the long-term impact of these exposures by leveraging existing large cohort studies with banked blood and/or DNA samples. We established the MEE Study in order to bridge this gap in knowledge by developing DNA methylation-based biomarkers for environmental exposures and other risk factors for disease. Such biomarkers have been identified for some exposures[14,104,107] but many others have not been extensively studied.

Our targeted approach to recruitment had a high success rate: 20.7% of all potential participants who were contacted and 43.0% of those who responded to our initial contact attempts eventually completed the study. This is likely due to the targeted recruitment pool being members of a breast cancer screening program who consented to being contacted for future research studies. Of the women who were eligible for the study, the most common reasons given for declining to participate were that they were uninterested, had concerns about the time commitment involved, and had concerns about using a computer or mobile phone to complete study questionnaires. Because the community recruitment method relied on participants initiating contact with the study team, it is impossible to know the success rate or reasons for non-participation in this group. However, the general characteristics of participants recruited through each method were largely similar.

A primary strength of our study is the resulting biorepository of plasma, serum, urine, and saliva as well as extracted DNA from the buffy coat. Plasma and serum can be analyzed for an array of molecules including antibodies and other proteins, antigens, lipids, hormones, and exogenous substances. Multiple urine samples allow for the reasonably representative measurement of chemicals with short half-lives. However, the relatively short time period of biospecimen collection (up to 10 days between urine samples) remains a limitation of the study. Certain exposures of interest are likely to be relatively consistent over time, while others may exhibit strong temporal variability. We plan to leverage the MEE Study population in the future by recruiting a subset of cohort members to a longitudinal study, providing repeated biospecimens which will assess environmental exposures over a longer period of time. While analyses of urinary glyphosate, organophosphates, and bisphenols have been completed, remaining aliquots of stored urine samples are available for analysis

of other environmental chemicals as well as metabolomic and electrolyte analysis. Some of the extracted blood DNA has been used for DNA methylation array analysis, and remaining blood DNA can be used for other genetic and epigenetic analyses. Stored saliva samples can also be used for genetic and epigenetic analyses as well as microbiome or biochemical analyses in the future.

Another strength is our comprehensive data collection which includes detailed dietary information (multiple 24-hour dietary recalls and a food frequency questionnaire), self-reported environmental exposures, lifestyle factors, residential history, occupation, and personal and family health history, especially breast health-related data. While our sample size is much smaller than large cohort studies such as the Women's Health Initiative or Nurses' Health Study, this smaller sample size allowed deep, focused characterization of each participant with omics-scale data from diverse biospecimens (blood, saliva, and urine). This data was used to prepare a nutrition and environmental exposure report for return to study participants, and 99% of participants requested to receive the reports. Dietary recalls and first-void urine samples were unannounced, which should provide a more unbiased assessment of usual diet and exposures. Two dietary recalls were paired with urine samples; that is, participants received a request at 7:30 PM one evening to collect their first-void urine the next day, followed by a second message at 10:00 AM to request that they complete a 24-hour dietary recall reflecting the previous day's intake. Participants were not informed at the time of recruitment that urine samples and dietary recalls would be paired, so the request for a urine sample should not have affected the unannounced nature of the dietary recall. In addition, because the urine sample request was sent at 7:30 PM, most of that day's dietary intake would have already taken place, further reducing the chance of bias. Additionally,

participants provided information on their current and former residences, which may be used in the future to assess residential exposure to air pollution as well as other neighborhood factors. However, due to the questionnaire-based assessment of most environmental exposures, not all exposures were captured to minimize participant burden.

A major limitation of the MEE Study is the lack of follow-up and thus health outcome data. The study was originally conceived with a cross-sectional design to allow the development of exposure biomarkers, which does not require health outcome data. This limitation does not affect the primary goal to identify epigenetic markers of environmental exposures, which will support future studies of potential relationships between environmental exposures and health. However, the MEE Study cohort is well-characterized and its participants have shown a high willingness to participate in research. Thus, the data collected can be enhanced in the future through linkage to national and state-level cancer registries, Department of Motor Vehicles and voter registration records, hospital discharge records from the California Office of Statewide Health Planning and Development (OSHPD), mortality records, and other sources of data. Additional data may also be gathered through follow-up questionnaires.

The study cohort was drawn from the primarily suburban area of Orange County, California. Similar to the demographic characteristics of the region, study participants are fairly racially/ethnically diverse with a large proportion being of Asian and Hispanic heritage, and are largely college educated[129]. Generally, the study population exhibited higher rates of healthful behaviors than the general U.S. population. For example, just 4.3% of study participants were current smokers compared to 19.7% of women aged 45–64 nationwide; similarly, 22.4% of study participants had a BMI of 30 kg/m^2 or greater,

compared to 32.4% of women aged 45–64 nationwide[130]. An exception was the adherence to physical activity guidelines, which was similar to the national profile: 40.7% of participants met these guidelines compared to 40.9% of women aged 45-64 nationwide[130]. Since Orange County has a large Spanish-speaking population[129], study materials were translated to Spanish; seventeen study participants (4.3%) requested Spanish-language materials and communication. Additionally, the study population is entirely composed of postmenopausal women, which minimizes the biological variance in DNA methylation and other markers due to age, sex, or menopausal status, but may limit the generalizability of the results. Applicability of future study results will need to be validated in other populations.

1.5 Conclusions

The Markers for Environmental Exposures (MEE) Study is a well-characterized cohort of 392 postmenopausal women residing in Southern California. In addition to comprehensive exposure, dietary, and health history data, the study established a biorepository of blood, urine, and saliva samples. The long-term aims of the study are to (1) identify biomarkers, especially epigenetic markers, of environmental exposures, such as pesticides, air/food/water contaminants, and industrial chemicals that are commonly encountered in the general population; and (2) support the study of potential relationships between environmental exposures and health and health-related factors. While studies have been initiated to identify DNA methylation markers for some environmental exposures by the study investigators, the MEE Study is also intended to serve as a resource for researchers beyond the original investigators.

Chapter 2: Factors Associated with Organic Eating Behaviors and Urinary Glyphosate and AMPA in Postmenopausal Women

2.1 Introduction

The broad-spectrum herbicide glyphosate is the most used pesticide in the world[39]. Its agricultural applications include weed control, reduction of soil tillage, preparation of seed beds, and pre-harvest desiccation of grain crops and legumes[39,40] in addition to non-agricultural uses in residential areas, parks, forests, pastures, and lawns[41]. Use of glyphosate-based herbicides has increased dramatically since the introduction of glyphosate-resistant (GR) crop species in the late 1990s[40], although they are also commonly used on non-GR crops[39,41,42]. Glyphosate's herbicidal properties are due to inhibition of the shikimate pathway enzyme 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), which is found in plants and bacteria but not humans. Current commercially available GR crops are genetically engineered to withstand glyphosate through the introduction of a glyphosate-tolerant version of EPSPS (cp4) from bacteria, except for canola which has both the cp4 EPSPS and a bacterial glyphosate oxidoreductase[131,132].

Glyphosate is broken down by two main pathways: by C-P lyase to sarcosine and phosphate and more commonly by glyphosate oxidoreductase to aminomethylphosphonic acid (AMPA) and glyoxylate[133–135]. AMPA is further degraded by C-P lyase to produce methylamine and phosphate[135]. Most degradation of glyphosate in the environment is via soil microbes, although many plant species (both GR and non-GR) also appear to have at least some ability to metabolize glyphosate via the AMPA pathway[132,136–139].

From 1996 to 2014, the total volume of glyphosate applied in the U.S. increased 9.1-fold[39]. This increase in glyphosate use is likely accompanied by increases in human exposure

to glyphosate and AMPA. Both glyphosate and AMPA are frequently detected in urine samples from non-farming populations. In the U.S., recent studies detected glyphosate in 70-90% of participants[43–46]. The general population is most likely to be exposed to glyphosate and/or AMPA through residues on food; a recent study found that urinary glyphosate and AMPA concentrations decreased more than 70% after just 5 days of a completely organic diet[61]. Both compounds have been detected on a large variety of foods. In 2017, the U.S. Food and Drug Administration found that 40.0% of soybean samples and 17.9% of corn samples had glyphosate residues[62]. Glyphosate is also frequently detected in wheat and cereal crops[63–65] and glyphosate and/or AMPA has been detected in many other foods including lentils/beans[63,64,66], honey[66,67], various fruits and vegetables[63,64,66], tea[66,68], wine[64], infant foods[63], and prepared foods such as cookies/crackers and frozen meals[63].

The potential health effects of glyphosate on humans are widely debated. In 2015 the World Health Organization's (WHO) International Agency for Research on Cancer (IARC) determined glyphosate to be a probable human carcinogen[47], while other agencies including the U.S. Environmental Protection Agency (EPA) disagree with this assessment[48,49]. Recently, a pilot study on 250 women in Hawaii found that higher AMPA levels were associated with breast cancer risk[50]. In agricultural workers, glyphosate has been associated with kidney toxicity[51] and increased risk for non-Hodgkin's lymphoma[52]. Glyphosate has also been shown to induce fatty liver disease[53], endocrine disruption[54], and changes to the microbiome[55–57] in animal models, and the growth of breast cancer cells in vitro[58]. Thus, there is some interest in monitoring and limiting exposure to glyphosate and AMPA in humans[61,140–143].

Despite the evidence for widespread presence of glyphosate and AMPA in the food supply, few studies have examined dietary factors associated with personal glyphosate and AMPA exposure. This study evaluated urinary glyphosate and AMPA levels in a group of postmenopausal women in southern California and associations with their diet and other factors.

2.2 Methods

Population and data collection

Recruitment and data collection for the Markers for Environmental Exposures (MEE) Study have been previously described[144]. Briefly, postmenopausal women between ages 45 and 66 years without a personal history of breast cancer or mastectomy were eligible to enroll in the study. Study participants completed up to three 24-hour dietary recalls, questionnaires about their demographic and lifestyle information, including usual dietary habits, and provided two first-morning urine samples for analysis. All participants provided informed consent to join the study and the study protocol was approved by the University of California, Irvine Institutional Review Board, HS #2016-3127.

Glyphosate and AMPA measurements

Study participants provided two first-morning urine samples on two days within a 10-day period not known to them in advance. The urine samples were stored at -20°C until they could be transported to the laboratory, where they were thawed overnight at 4°C, then aliquoted and stored at -80°C for later analysis.

Urinary glyphosate and AMPA were measured using liquid chromatography with tandem mass spectrometry (LC-MS/MS) at the Collaborative Center for Translational Mass Spectrometry (CCTMS) (TGen, Phoenix, AZ) using a Vanquish UHPLC coupled to TSQ Altis

triple quadrupole mass spectrometer (Thermo Scientific, San Jose, CA) as previously described[44]. Both assays were linear (coefficient of determination $R^2 > 0.99$) over a range 0-5 ng/mL (Appendix B, Figure B1). The limits of detection (LOD) for glyphosate and AMPA were 0.014 and 0.013 ng/mL, and the limits of quantitation (LOQ) were 0.041 and 0.040 ng/mL, respectively (Appendix A, Table A2). Creatinine was measured using the DetectX urinary creatinine detection kit (Arbor Assays, Ann Arbor, MI, K002-H5) according to the manufacturer's instructions. The data were tested for batch effects using the Kruskal-Wallis test, and batch correction was performed for glyphosate values using the removeBatchEffects function in limma[145]. Glyphosate and AMPA values below the LOD were replaced with $LOD/\sqrt{2}$[146].

Dietary and demographic/behavioral measurements

Study participants were asked to complete three 24-hour dietary recalls using the Automated Self-Administered 24-hour (ASA24) Dietary Assessment Tool, version 2016 (National Cancer Institute, Bethesda, MD), which captures the type and amount of all food and drink consumed within a 24-hour period. Recalls were not announced in advance in order to capture typical dietary habits. A recall was requested for the day preceding each urine sample collection; if successfully completed, the recall and urine sample were considered "paired." Because the half-life of glyphosate in urine is approximately 8-10 hours[59,60], foods and drinks consumed in the 24 hours prior to specimen collection are likely to have the strongest influence on glyphosate levels.

Using the paired dietary recalls and urine samples, we examined associations between urinary glyphosate and AMPA and consumption of the 5 major food groups according to the Dietary Guidelines for Americans: fruit, vegetables, grains, protein foods,

and dairy, all derived from standard ASA24 output. Other subgroups of interest were foods and beverages previously reported or suspected to have elevated glyphosate levels or crops with glyphosate-resistant varieties. Among these subgroups, consumption of legumes, whole grains, refined grains, meat/poultry (combination of beef, poultry, pork, veal, lamb, and cured/organ meats), eggs, soy protein, and nuts and seeds were calculated from ASA24 output. For corn, oats, other soy products, coffee, tea, and alcoholic drinks (total and wine, beer, and spirits separately), the individual food and beverage items reported on each recall were reviewed to identify those that contained the food/beverage of interest (see Appendix A, Table A3 for a full list of all participant-reported items included in each category). Recalls reporting a total kilocalorie (Kcal) intake of <500 or >3500 Kcal were excluded from the analysis[126].

Other demographic and behavioral variables of interest, assessed by questionnaire, were examined for associations with urinary glyphosate and AMPA. These included age, self-reported race/ethnicity, body mass index (BMI) (calculated from self-reported height and weight), smoking status (current, former, or never), education (college graduate or higher, some college, high school graduate, some high school or less), ZIP code or city of residence, physical activity (≥150 minutes of moderate-intensity aerobic exercise per week or an equivalent combination of moderate- and vigorous-intensity exercise), frequency of dining at fast food restaurants, and usual drinking water source (bottled, filtered tap, or unfiltered tap). Household income was approximated using the median household income in the participant's ZIP code or city of residence (where ZIP code missing, N = 10) for householders aged 45-64 from the American Community Survey 2019 5-Year Estimates (data.census.gov). Overall dietary pattern was characterized according to the Healthy Eating Index 2015 (HEI),

a score of adherence to the 2015-2020 Dietary Guidelines for Americans[147]. The HEI was calculated 1) for individual recalls to characterize dietary pattern on a specific day and 2) for all available (paired and unpaired) recalls as a more comprehensive representation of usual eating habits. Study participants were asked how often they ate organic food, both in general and for six specific food groups (fruit, vegetables, grains, meat, eggs, and dairy), with the response choices "often or always," "sometimes," "seldom or never,"[148] and "I do not eat the food" (for specific food groups only). Participants were provided the following definition of organic food: food that is either labeled USDA Organic, purchased locally from an organic farm, grown without pesticides in a home garden, or raised on organic feed without hormones and without antibiotics[144,148].

Statistical analysis

A p-value of < 0.05 was considered statistically significant and all statistical tests were two-sided. All statistical analyses were conducted in R, version 3.5.0[149].

We examined relationships between dietary, demographic and behavioral factors and self-reported organic eating habits using Fisher's exact test (categorical variables) and the Kruskal-Wallis test (continuous variables).

Relationships between dietary, demographic, and behavioral factors and urinary glyphosate and AMPA were assessed using linear mixed-effects models. Major food groups (fruit, vegetables, grains, protein foods, and dairy) were evaluated as quartiles of consumption, while other foods and beverages were evaluated as binary (consumed/not consumed). Less than 5% of participants had any missing data (n = 5 participants for race/ethnicity, 1 for education, 12 for physical activity, 1 for smoking status). In the mixed-effects model, these missing data points were imputed with predictive mean matching using

the mice R package. The base model used the natural logarithm of urinary glyphosate or AMPA as the outcome and a fixed effect for the predictor of interest, with a fixed effect for creatinine to account for urine concentration and a random effect for individual. The adjusted model included additional fixed effects for total Kcal, dietary pattern (as HEI for the individual recall), and the non-diet-related variables that were nominally significant ($p < 0.10$) in the base model: race/ethnicity and physical activity level. For food groups with data about organic eating consumption (fruit, vegetables, grains, meat, eggs, and dairy), we repeated the adjusted model after stratifying by self-reported organic eating for that food group.

A sensitivity analysis was conducted after excluding potentially problematic recalls. These included potentially incomplete recalls (e.g., study participant closed the browser window rather than clicking "finish") and recalls which included one or more poorly-mapped foods. ASA24 maps foods and beverages to the USDA Food and Nutrient Database for Dietary Studies (FNDDS) to obtain the nutritional content of the item. If a respondent-provided item was not recognized by the FNDDS, a best-guess attempt to map it to a known item is made. The list of these "unfound foods" was reviewed by two authors (RML and HLP) and classified as poorly-mapped (having a markedly different macronutrient profile and/or affecting the measurement of one of the foods of interest) or acceptable[120]. For example, a user's entry of "organic shade grown coffee" which was mapped to "soft drink" was considered poorly-mapped, while a user's entry of "sun-dried tomato" being mapped to "tomatoes, raw" was considered incorrect but acceptable.

2.3 Results

After exclusions, 338 individuals completed 567 paired recalls/urine samples and formed the main study cohort (Figure 2.1). 229 participants (67.8%) completed two paired dietary recalls, while the remaining 109 participants (32.2%) completed only one.

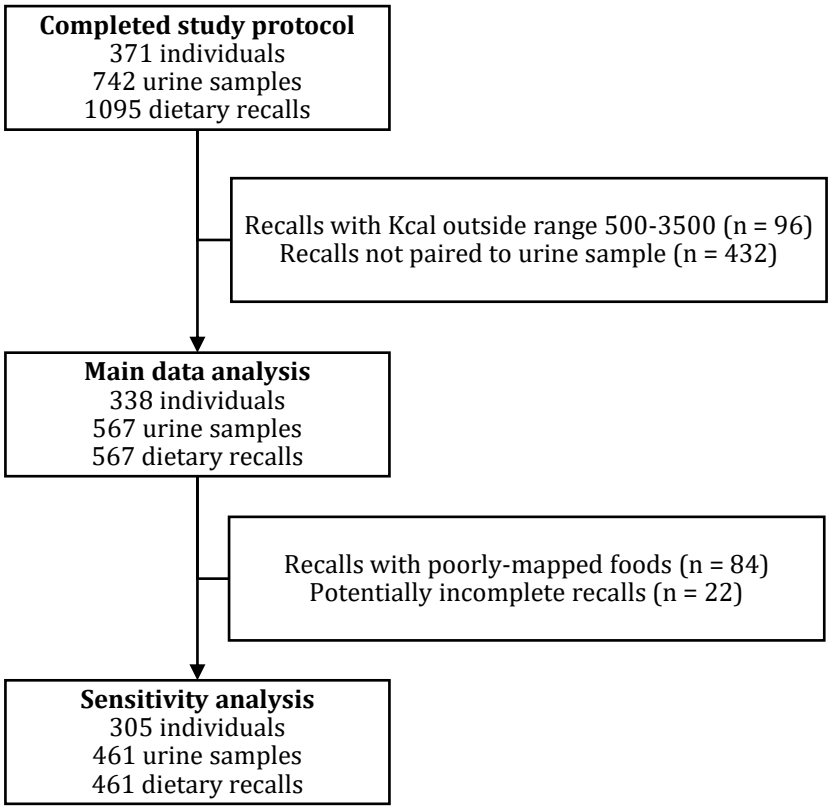

Figure 2.1. Diet study flow chart

Organic eating behaviors

Most study participants were non-Hispanic whites who were college graduates and reported never smoking (Table 2.1). The three categories of self-reported overall organic eating habits were approximately evenly distributed, with 37.9% reporting often or always eating organic food, 30.2% sometimes, and 32.0% seldom or never. About a third of women reported often or always eating organic fruit, vegetables, and eggs, while less than 20% often or always ate organic grains, meat, and dairy (Table 2.2).

Table 2.1. Subject characteristics, overall and stratified by self-reported organic eating

	Overall N = 338	Self-Reported Organic Eating			
		Seldom/Never N = 108 (32.0%)	Sometimes N = 102 (30.2%)	Often/Always N = 128 (37.9%)	
	Median (Q1, Q3) or N (%)	Median (Q1, Q3) or N (%)	Median (Q1, Q3) or N (%)	Median (Q1, Q3) or N (%)	p
Age	57 (54, 60)	57 (53, 60)	57 (54, 60)	58 (55, 61)	0.20
Race/Ethnicity					0.19
Non-Hispanic White	225 (67.6%)	67 (63.8%)	62 (62.0%)	96 (75.0%)	
Asian	34 (10.2%)	13 (12.4%)	13 (13.0%)	8 (6.3%)	
Hispanic	58 (17.4%)	22 (21.0%)	19 (19.0%)	17 (13.3%)	
Other Race	16 (4.8%)	3 (2.9%)	6 (6.0%)	7 (5.5%)	
Education					**0.002**
College Graduate	241 (71.5%)	71 (65.7%)	72 (70.6%)	98 (77.2%)	
Some College	68 (20.2%)	18 (16.7%)	26 (25.5%)	24 (18.9%)	
High School Graduate	21 (6.2%)	14 (13.0%)	2 (2.0%)	5 (3.9%)	
Some High School	7 (2.1%)	5 (4.6%)	2 (2.0%)	0 (0.0%)	
Median Annual Household Income					0.11
Less than $100,000	99 (29.3%)	40 (37.0%)	27 (26.4%)	32 (25.0%)	
$100,000 - $149,999	164 (48.5%)	48 (44.4%)	56 (54.9%)	60 (46.9%)	
$150,000 or more	75 (22.2%)	20 (18.5%)	19 (18.6%)	36 (28.1%)	
BMI (kg/m^2)	25.0 (22.1, 29.3)	26.5 (23.3, 31.4)	24.4 (22.1, 29.5)	24.0 (21.4, 28.1)	**0.002**
Physical Activity					0.24
<150 minutes/week	188 (57.7%)	65 (61.9%)	60 (60.6%)	63 (51.6%)	
≥150 minutes/week	138 (42.3%)	40 (38.1%)	39 (39.4%)	59 (48.4%)	
Smoking Status					0.74
Never	243 (72.1%)	81 (75.0%)	72 (71.3%)	90 (70.3%)	
Former	80 (23.7%)	23 (21.3%)	23 (22.8%)	34 (26.6%)	
Current	14 (4.2%)	4 (3.7%)	6 (5.9%)	4 (3.1%)	
Average HEI (all available recalls)	64.0 (55.1, 73.0)	58.9 (51.7, 69.3)	64.3 (55.2, 73.3)	67.0 (59.0, 74.5)	**<0.001**

Table continues

Table 2.1. Subject characteristics, overall and stratified by self-reported organic eating (continued)

	Overall N = 338	Self-Reported Organic Eating			
		Seldom/Never N = 108 (32.0%)	Sometimes N = 102 (30.2%)	Often/Always N = 128 (37.9%)	
	Median (Q1, Q3) or N (%)	Median (Q1, Q3) or N (%)	Median (Q1, Q3) or N (%)	Median (Q1, Q3) or N (%)	p
Fast Food Meals					**0.007**
Less than 1 time/month	173 (51.2%)	43 (39.8%)	50 (49.0%)	80 (62.5%)	
1-3 times/month	93 (27.5%)	30 (27.8%)	33 (32.4%)	30 (23.4%)	
1-3 times/week	65 (19.2%)	30 (27.8%)	18 (17.6%)	17 (13.3%)	
4 or more times/week	7 (2.1%)	5 (4.6%)	1 (1.0%)	1 (0.8%)	
Primary Drinking Water Source					0.09
Bottled	128 (37.9%)	48 (44.4%)	33 (32.4%)	47 (36.7%)	
Tap (Filtered)	179 (53.0%)	47 (43.5%)	58 (56.9%)	74 (57.8%)	
Tap (Unfiltered)	31 (9.2%)	13 (12.0%)	11 (10.8%)	7 (5.5%)	
Number of Paired Dietary Recalls					0.30
1	109 (32.2%)	36 (33.3%)	27 (26.5%)	46 (35.9%)	
2	229 (67.8%)	72 (66.7%)	75 (73.5%)	82 (64.1%)	

P-values for differences between organic eating categories from Fisher's exact test (categorical variables) or Kruskal-Wallis test (continuous variables). Missing data: 5 for race/ethnicity, 1 for education, 12 for physical activity, 1 for smoking status. BMI: body mass index, HEI: Healthy Eating Index

Table 2.2. Self-reported organic eating habits for six food groups

Organic Eating	Fruit N (%)	Vegetables N (%)	Grains N (%)	Meat N (%)	Eggs N (%)	Dairy N (%)
Often/Always	112 (33.1)	110 (32.5)	56 (16.6)	61 (18.0)	114 (33.7)	64 (18.9)
Sometimes	79 (23.4)	81 (24.0)	48 (14.2)	67 (19.8)	33 (9.8)	44 (13.0)
Seldom/Never	143 (42.3)	146 (43.2)	220 (65.1)	188 (55.6)	183 (54.1)	205 (60.7)
Do not consume	4 (1.2)	1 (0.3)	14 (4.1)	22 (6.5)	8 (2.4)	25 (7.4)

Education was significantly associated with overall organic eating habits ($p = 0.002$), with 77.2% of women who often or always ate organic food having graduated from college compared to 65.7% of seldom or never organic eaters (Table 2.1). Often/always organic eaters had a lower BMI compared to seldom/never (24.0 vs. 26.5 kg/m^2, $p = 0.002$). Often/always organic eaters also had a higher median HEI (calculated for all available recalls) of 67.0 compared to 58.9 in those who seldom or never ate organic food ($p < 0.001$), and ate at fast-food restaurants less often (37.5% vs. 60.2% had fast food once or more per month, $p = 0.007$). Other characteristics (age, race/ethnicity, household income, physical activity level, smoking status, and drinking water source) were not significantly associated with organic eating habits.

Of the major food groups, seldom/never organic eaters consumed less fruit, more grains (particularly refined grains), and more meat and poultry compared to often/always organic eaters (Table 2.3). Most of the specific foods and beverages examined were consumed relatively infrequently, with the exception of eggs, nuts and seeds, and coffee, which were each reported on more than half of the dietary recalls (Table 2.3). Legumes were consumed on 29.0% of recalls completed by those who often or always ate organic food compared to 16.7% completed by those who seldom or never did ($p = 0.01$). However, seldom/never organic eaters reported consuming significantly less alcohol, particularly

wine, compared to the other two categories of organic eating. Overall, 31% of recalls reported drinking alcoholic beverages, with nearly 25% consuming wine, but only 24.4% of recalls completed by seldom/never organic eaters reported drinking any alcohol and 15.0% drank wine.

Glyphosate and AMPA concentrations

Nearly all urine samples (89.9%) had detectable glyphosate and 67.2% had detectable AMPA (Table 2.4). The majority of samples (76.4% for glyphosate and 50.4% for AMPA) had values above the LOQ for both analytes. The median concentrations were 0.10 and 0.04 ng/mL for glyphosate and AMPA, respectively, with a maximum of 3.01 ng/mL for glyphosate and 1.51 ng/mL for AMPA. For both glyphosate and AMPA, the median concentration in ng/mL was slightly but significantly greater in seldom/never organic eaters (Table 2.4, Figure 2.2).

Table 2.3. Dietary recall characteristics, overall and stratified by self-reported organic eating

	Overall N = 567	Self-Reported Organic Eating			
		Seldom/Never N = 180 (31.7%)	Sometimes N = 177 (31.2%)	Often/Always N = 210 (37.0%)	
	Median (Q1, Q3)	Median (Q1, Q3)	Median (Q1, Q3)	Median (Q1, Q3)	p
Kcal	1653 (1301, 2100)	1699 (1360, 2133)	1644 (1252, 2152)	1631 (1293, 2010)	0.18
Major Food Groups					
Fruit (cup equivalents)	0.7 (0.0, 1.7)	0.5 (0.0, 1.2)	0.8 (0.1, 1.7)	0.8 (0.1, 1.7)	**0.005**
Vegetables (cup equivalents)	1.6 (0.8, 2.8)	1.4 (0.7, 2.8)	1.6 (0.7, 2.7)	1.8 (1.0, 2.8)	0.08
Grains (ounce equivalents)	4.1 (2.4, 6.0)	4.6 (2.9, 6.9)	4.0 (2.3, 5.9)	3.8 (2.0, 5.5)	**0.003**
Whole Grains (ounce equivalents)	0.5 (0.0, 1.6)	0.5 (0.0, 1.7)	0.4 (0.0, 1.4)	0.5 (0.0, 1.6)	0.26
Refined Grains (ounce equivalents)	3.0 (1.5, 5.0)	3.6 (1.9, 5.7)	3.1 (1.6, 5.0)	2.5 (1.1, 4.3)	**0.002**
Protein Foods (ounce equivalents)	5.3 (2.9, 8.0)	5.8 (3.0, 8.9)	5.2 (3.1, 7.6)	5.0 (2.5, 7.8)	0.10
Meat and Poultry (ounce equivalents)	2.3 (0.1, 4.7)	2.8 (1.0, 5.5)	2.0 (0.0, 4.3)	1.9 (0.0, 4.1)	**0.002**
Dairy (cup equivalents)	1.0 (0.4, 1.8)	1.1 (0.4, 2.0)	1.0 (0.4, 1.6)	1.1 (0.4, 1.8)	0.27
Other Foods and Beverages (Yes/No)	N (%)	N (%)	N (%)	N (%)	
Legumes	136 (24.0%)	30 (16.7%)	45 (25.4%)	61 (29.0%)	**0.01**
Corn	165 (29.1%)	49 (27.2%)	52 (29.4%)	64 (30.5%)	0.78
Oats	119 (21.0%)	44 (24.4%)	35 (19.8%)	40 (19.0%)	0.39
Eggs	410 (72.3%)	133 (73.9%)	127 (71.8%)	150 (71.4%)	0.85
Soy Protein	100 (17.6%)	36 (20.0%)	27 (15.3%)	37 (17.6%)	0.51
Soy Protein, Milk, and Other Soy Foods	171 (30.2%)	54 (30.0%)	56 (31.6%)	61 (29.0%)	0.86
Nuts and Seeds	331 (58.4%)	100 (55.6%)	101 (57.1%)	130 (61.9%)	0.40
Coffee	372 (65.6%)	112 (62.2%)	116 (65.5%)	144 (68.6%)	0.43
Tea	179 (31.6%)	52 (28.9%)	60 (33.9%)	67 (31.9%)	0.60
Alcohol	176 (31.0%)	44 (24.4%)	66 (37.3%)	66 (31.4%)	**0.03**
Wine	127 (22.4%)	27 (15.0%)	48 (27.1%)	52 (24.8%)	**0.006**
Spirits	37 (6.5%)	13 (7.2%)	14 (7.9%)	10 (4.8%)	0.39
Beer	32 (5.6%)	8 (4.4%)	12 (6.8%)	12 (5.7%)	0.62

P-values for differences between organic eating categories from Fisher's exact test (categorical variables) or Kruskal-Wallis test (continuous variables). Kcal: kilocalories

Table 2.4. Urinary glyphosate and AMPA measurements, overall and stratified by self-reported organic eating

	Overall N = 567			Self-Reported Organic Eating			
				Seldom/Never N = 180 (31.7%)	Sometimes N = 177 (31.2%)	Often/Always N = 210 (37.0%)	
	Value (ng/mL) Median (Q1, Q3)	≥LOD N (%)	≥LOQ N (%)	Value (ng/mL) Median (Q1, Q3)	Value (ng/mL) Median (Q1, Q3)	Value (ng/mL) Median (Q1, Q3)	p
Glyphosate	0.10 (0.04, 0.25)	510 (89.9%)	433 (76.4%)	0.13 (0.06, 0.27)	0.10 (0.04, 0.27)	0.09 (0.04, 0.22)	**0.04**
AMPA	0.04 (<LOD, 0.11)	381 (67.2%)	286 (50.4%)	0.05 (<LOD, 0.14)	0.04 (<LOD, 0.12)	0.03 (<LOD, 0.08)	**0.01**

P from Kruskal-Wallis test between organic eating categories. AMPA: aminomethylphosphonic acid

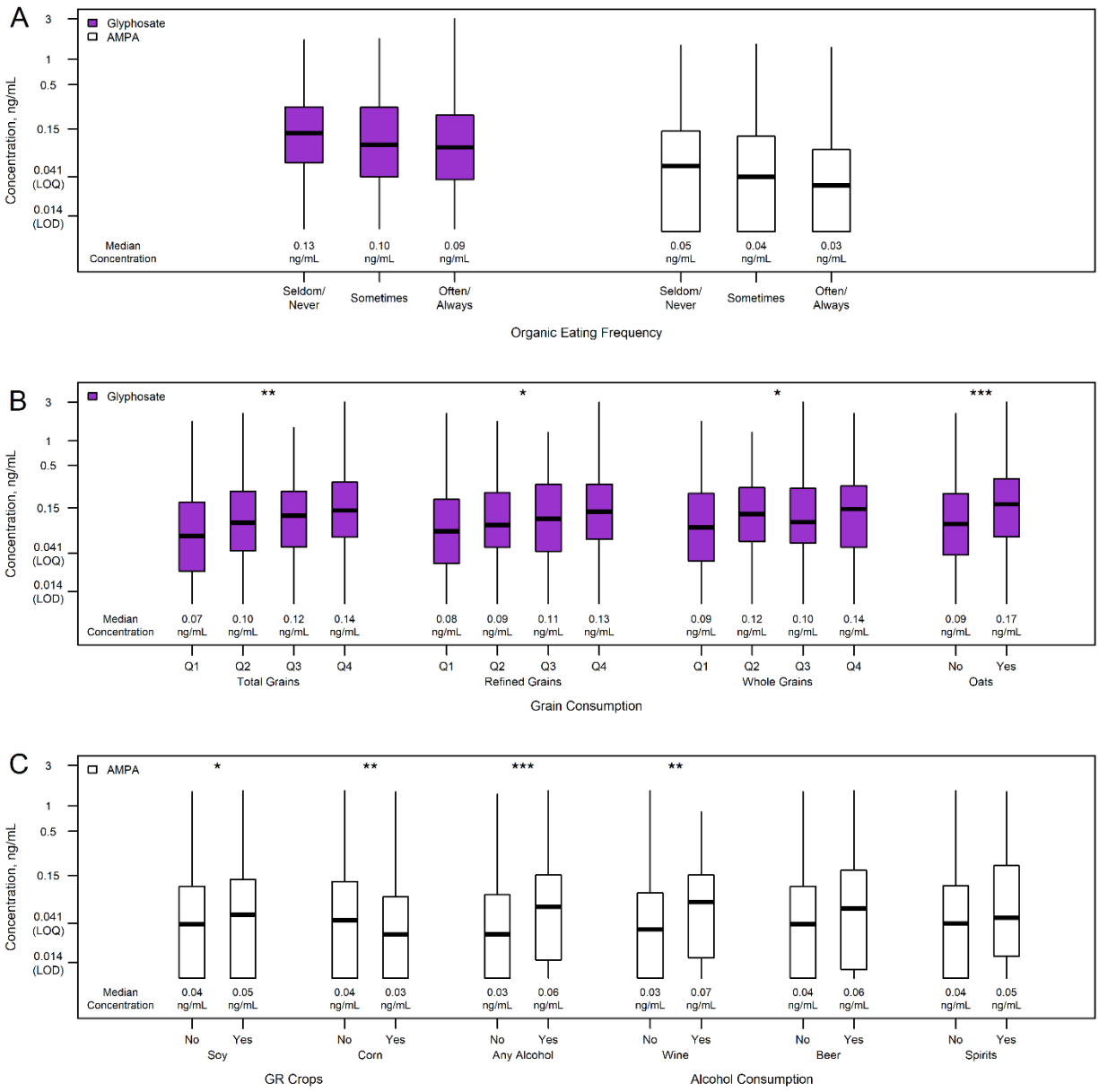

Figure 2.2. Urinary glyphosate and/or AMPA concentrations for selected categories A) Self-reported organic eating frequency, B) grain consumption, C) soy, corn, and alcohol consumption. Asterisks indicate statistical significance when adjusted for urinary creatinine, total Kcal, recall-specific Healthy Eating Index (HEI), race/ethnicity, and physical activity level: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. AMPA: aminomethylphosphonic acid; GR: Glyphosate-resistant

Dietary and demographic factors associated with glyphosate and AMPA

Results were generally consistent between the base and adjusted models (Table 2.5 for recall-level characteristics, Table 2.6 for individual-level characteristics). A model

adjusted for just creatinine and dietary pattern also showed similar results (data not shown). Grain consumption, particularly refined grains, was associated with elevated urinary glyphosate in both models (Table 2.5**,** Figure 2.2). In the adjusted model, meat/poultry consumption was negatively associated with glyphosate levels ($p = 0.01$). Consuming oats was strongly associated with greater glyphosate levels (adjusted $p < 0.001$, Figure 2.2). Women who completed 150 minutes or more of moderate-intensity exercise per week had lower urinary glyphosate (Table 2.6, adjusted $p = 0.04$). No other dietary, demographic, or behavioral factors were significantly associated with urinary glyphosate. When stratified by organic eating habits, total grain consumption remained associated with elevated glyphosate in both often/always and seldom/never organic grain eaters (Table 2.7). The inverse relationship between meat consumption and glyphosate levels was present only in those who seldom or never ate organic meat ($p = 0.04$).

Table 2.5. Association of dietary characteristics with glyphosate and AMPA levels

	Glyphosate				AMPA			
	Base model		Adjusted model		Base model		Adjusted model	
	Coefficient (95% CI)	p	Coefficient (95% CI)	p	Coefficient (95% CI)	p	Coefficient (95% CI)	p
Major Food Groups (Quartile)								
Fruit	-0.06 (-0.15, 0.02)	0.14	-0.07 (-0.16, 0.02)	0.14	-0.11 (-0.21, -0.02)	**0.02**	-0.12 (-0.23, -0.01)	**0.03**
Vegetables	-0.05 (-0.14, 0.03)	0.23	-0.06 (-0.15, 0.03)	0.21	-0.02 (-0.11, 0.08)	0.71	-0.03 (-0.13, 0.07)	0.54
Grains	0.13 (0.05, 0.21)	**0.002**	0.14 (0.05, 0.23)	**0.004**	0.12 (0.03, 0.21)	**0.01**	0.06 (-0.04, 0.17)	0.24
Whole Grains	0.08 (-0.002, 0.15)	0.06	0.10 (0.01, 0.19)	**0.02**	-0.02 (-0.11, 0.06)	0.62	-0.01 (-0.11, 0.08)	0.78
Refined Grains	0.11 (0.03, 0.19)	**0.01**	0.11 (0.02, 0.21)	**0.02**	0.15 (0.06, 0.25)	**0.001**	0.11 (-0.005, 0.22)	0.06
Protein Foods	-0.06 (-0.14, 0.02)	0.18	-0.09 (-0.18, -0.003)	**0.04**	0.08 (-0.01, 0.17)	0.10	0.05 (-0.05, 0.15)	0.36
Meat and Poultry	-0.08 (-0.16, 0.005)	0.07	-0.12 (-0.20, -0.03)	**0.01**	0.03 (-0.06, 0.13)	0.51	-0.02 (-0.12, 0.08)	0.65
Dairy	-0.02 (-0.10, 0.07)	0.70	-0.04 (-0.13, 0.05)	0.37	-0.04 (-0.13, 0.06)	0.42	-0.08 (-0.18, 0.02)	0.11
Other Foods and Beverages (Yes/No)								
Legumes	0.18 (-0.03, 0.39)	0.09	0.21 (-0.01, 0.42)	0.06	-0.06 (-0.30, 0.18)	0.63	-0.06 (-0.30, 0.19)	0.65
Corn	-0.01 (-0.21, 0.18)	0.88	-0.05 (-0.25, 0.15)	0.63	-0.26 (-0.49, -0.04)	**0.02**	-0.31 (-0.54, -0.09)	**0.008**
Oats	0.37 (0.13, 0.60)	**0.003**	0.44 (0.19, 0.68)	**<0.001**	0.05 (-0.21, 0.31)	0.70	0.11 (-0.16, 0.38)	0.44
Eggs	0.004 (-0.19, 0.20)	0.97	-0.03 (-0.22, 0.17)	0.81	0.04 (-0.19, 0.27)	0.72	-0.02 (-0.25, 0.21)	0.86
Soy Protein	0.02 (-0.22, 0.26)	0.88	0.02 (-0.22, 0.26)	0.85	0.29 (0.02, 0.57)	**0.04**	0.32 (0.05, 0.59)	**0.02**
Soy Protein, Milk, Other Soy Foods	-0.06 (-0.26, 0.14)	0.56	-0.04 (-0.24, 0.16)	0.67	0.11 (-0.12, 0.33)	0.36	0.11 (-0.11, 0.34)	0.32
Nuts and Seeds	0.15 (-0.03, 0.34)	0.10	0.17 (-0.02, 0.37)	0.08	0.02 (-0.19, 0.23)	0.85	0.04 (-0.19, 0.26)	0.75
Coffee	0.04 (-0.18, 0.26)	0.73	-0.001 (-0.22, 0.22)	0.99	-0.01 (-0.24, 0.23)	0.95	-0.03 (-0.27, 0.21)	0.80
Tea	-0.06 (-0.27, 0.15)	0.60	-0.05 (-0.26, 0.16)	0.65	0.06 (-0.17, 0.29)	0.61	0.07 (-0.16, 0.30)	0.56
Alcohol	-0.01 (-0.23, 0.21)	0.92	-0.04 (-0.26, 0.19)	0.76	0.44 (0.21, 0.68)	**<0.001**	0.46 (0.22, 0.70)	**<0.001**
Wine	0.004 (-0.23, 0.24)	0.97	-0.01 (-0.25, 0.23)	0.95	0.41 (0.15, 0.67)	**0.002**	0.44 (0.18, 0.70)	**0.001**
Beer	-0.14 (-0.54, 0.25)	0.48	-0.19 (-0.59, 0.20)	0.34	0.30 (-0.15, 0.75)	0.19	0.26 (-0.19, 0.71)	0.26
Spirits	-0.003 (-0.38, 0.37)	0.99	-0.02 (-0.40, 0.36)	0.91	0.19 (-0.23, 0.62)	0.38	0.17 (-0.26, 0.59)	0.45
Dietary Pattern								
HEI (for individual dietary recall)	-0.002 (-0.01, 0.005)	0.58	-0.001 (-0.01, 0.01)	0.75	-0.01 (-0.01, 0.002)	0.15	-0.01 (-0.01, 0.002)	0.15

Base model adjusted for urinary creatinine with a random effect for individual; adjusted model additionally adjusted for total kilocalories, recall-specific Healthy Eating Index (HEI), race/ethnicity, and physical activity level. AMPA: aminomethylphosphonic acid

Table 2.6. Association of demographic/behavioral variables with glyphosate and AMPA levels

	Glyphosate				AMPA			
	Base model		Adjusted model		Base model		Adjusted model	
	Coefficient (95% CI)	p	Coefficient (95% CI)	p	Coefficient (95% CI)	p	Coefficient (95% CI)	p
Age	0.02 (-0.01, 0.05)	0.16	0.02 (-0.01, 0.05)	0.13	-0.002 (-0.03, 0.02)	0.89	0.001 (-0.03, 0.03)	0.92
Race/Ethnicity								
Non-Hispanic White	Ref		Ref		Ref		Ref	
Asian	-0.21 (-0.61, 0.19)	0.30	-0.21 (-0.61, 0.19)	0.29	0.19 (-0.21, 0.59)	0.36	0.23 (-0.17, 0.63)	0.26
Hispanic	0.06 (-0.26, 0.38)	0.70	0.04 (-0.28, 0.36)	0.80	0.03 (-0.29, 0.35)	0.85	0.01 (-0.31, 0.33)	0.97
Other Race	-0.52 (-1.09, 0.06)	0.08	-0.53 (-1.10, 0.05)	0.07	-0.64 (-1.23, -0.06)	**0.03**	-0.69 (-1.27, -0.11)	**0.02**
Education								
College Graduate	Ref		Ref		Ref		Ref	
Some College	0.05 (-0.25, 0.35)	0.74	-0.03 (-0.34, 0.28)	0.85	0.15 (-0.14, 0.45)	0.31	0.09 (-0.22, 0.40)	0.58
High School Graduate	0.19 (-0.31, 0.69)	0.46	0.10 (-0.43, 0.64)	0.71	0.24 (-0.26, 0.74)	0.35	0.13 (-0.41, 0.66)	0.64
Some High School	0.60 (-0.21, 1.41)	0.15	0.54 (-0.32, 1.39)	0.22	-0.68 (-1.48, 0.13)	0.10	-0.71 (-1.55, 0.14)	0.10
Median Annual Household Income								
Less than $100,000	Ref		Ref		Ref		Ref	
$100,000 - $149,999	-0.05 (-0.32, 0.23)	0.73	-0.05 (-0.33, 0.23)	0.71	0.17 (-0.11, 0.44)	0.24	0.15 (-0.13, 0.43)	0.30
$150,000 or more	-0.13 (-0.47, 0.20)	0.43	-0.11 (-0.46, 0.24)	0.53	-0.003 (-0.34, 0.33)	0.99	0.04 (-0.31, 0.39)	0.82
BMI	0.01 (-0.01, 0.03)	0.25	0.01 (-0.01, 0.03)	0.44	0.01 (-0.01, 0.03)	0.17	0.01 (-0.01, 0.03)	0.25
Physical Activity								
<150 minutes/week	Ref		Ref		Ref		Ref	
≥150 minutes/week	-0.26 (-0.49, -0.02)	**0.03**	-0.26 (-0.49, -0.02)	**0.04**	-0.18 (-0.42, 0.06)	0.14	-0.14 (-0.38, 0.10)	0.25
Smoking Status								
Never	Ref		Ref		Ref		Ref	
Former	0.01 (-0.27, 0.29)	0.94	-0.01 (-0.29, 0.27)	0.96	-0.23 (-0.51, 0.05)	0.11	-0.21 (-0.49, 0.06)	0.13
Current	-0.24 (-0.84, 0.35)	0.42	-0.31 (-0.92, 0.29)	0.31	-0.14 (-0.73, 0.46)	0.65	-0.11 (-0.71, 0.49)	0.71
Fast Food Meals								
Less than 1 time/month	Ref		Ref		Ref		Ref	
1-3 times/month	-0.05 (-0.33, 0.23)	0.73	-0.12 (-0.41, 0.17)	0.42	0.07 (-0.21, 0.35)	0.62	-0.01 (-0.30, 0.27)	0.92
1-3 times/week	0.18 (-0.14, 0.50)	0.27	0.05 (-0.29, 0.39)	0.76	0.22 (-0.10, 0.54)	0.18	0.10 (-0.24, 0.44)	0.55
4 or more times/week	0.76 (-0.07, 1.59)	0.07	0.64 (-0.20, 1.47)	0.14	1.37 (0.55, 2.20)	**0.001**	1.19 (0.35, 2.02)	**0.006**

Table continues

Table 2.6. Association of demographic/behavioral variables with glyphosate and AMPA levels (continued)

	Glyphosate				AMPA			
	Base model		Adjusted model		Base model		Adjusted model	
	Coefficient (95% CI)	p	Coefficient (95% CI)	p	Coefficient (95% CI)	p	Coefficient (95% CI)	p
Organic Eating								
Seldom/Never	0.21 (-0.08, 0.49)	0.15	0.17 (-0.12, 0.46)	0.24	0.27 (-0.01, 0.56)	0.06	0.18 (-0.11, 0.47)	0.22
Sometimes	0.12 (-0.17, 0.41)	0.42	0.11 (-0.18, 0.40)	0.46	0.22 (-0.06, 0.51)	0.13	0.19 (-0.10, 0.47)	0.21
Often/Always	Ref		Ref		Ref		Ref	
Primary Drinking Water Source								
Bottled	Ref		Ref		Ref		Ref	
Tap (Filtered)	-0.06 (-0.31, 0.19)	0.65	-0.04 (-0.30, 0.22)	0.74	-0.07 (-0.33, 0.18)	0.58	-0.09 (-0.34, 0.17)	0.52
Tap (Unfiltered)	-0.05 (-0.48, 0.38)	0.82	-0.09 (-0.52, 0.35)	0.70	-0.31 (-0.74, 0.12)	0.16	-0.34 (-0.77, 0.09)	0.12

Base model adjusted for urinary creatinine with a random effect for individual; adjusted model additionally adjusted for total Kcal, recall-specific Healthy Eating Index (HEI), race/ethnicity, and physical activity level. BMI: body mass index, AMPA: aminomethylphosphonic acid

Table 2.7. Association of quartile of consumption of major food groups with urinary glyphosate and AMPA, stratified by self-reported food-group-specific organic eating habits

Food Group	Organic Eating	Glyphosate		AMPA	
		Adjusted Coefficient (95% CI)	Adjusted p	Adjusted Coefficient (95% CI)	Adjusted p
Fruit	Often/Always	-0.09 (-0.24, 0.05)	0.20	-0.05 (-0.22, 0.11)	0.52
	Sometimes	0.09 (-0.10, 0.28)	0.37	-0.18 (-0.38, 0.03)	0.10
	Seldom/Never	-0.02 (-0.16, 0.13)	0.80	-0.03 (-0.20, 0.14)	0.73
Vegetables	Often/Always	-0.13 (-0.29, 0.03)	0.11	-0.02 (-0.18, 0.15)	0.84
	Sometimes	0.06 (-0.10, 0.22)	0.47	-0.06 (-0.24, 0.13)	0.56
	Seldom/Never	-0.04 (-0.19, 0.11)	0.59	0.08 (-0.09, 0.26)	0.36
Grains	Often/Always	0.25 (0.04, 0.45)	**0.02**	0.11 (-0.13, 0.36)	0.36
	Sometimes	0.13 (-0.14, 0.40)	0.36	0.09 (-0.18, 0.36)	0.53
	Seldom/Never	0.12 (0.006, 0.24)	**0.04**	0.05 (-0.08, 0.19)	0.44
Meat/Poultry	Often/Always	-0.09 (-0.28, 0.09)	0.34	-0.17 (-0.36, 0.02)	0.09
	Sometimes	-0.13 (-0.32, 0.06)	0.18	0.19 (-0.02, 0.40)	0.08
	Seldom/Never	-0.13 (-0.26, -0.01)	**0.04**	-0.001 (-0.14, 0.14)	0.99
Eggs	Often/Always	0.03 (-0.09, 0.15)	0.65	0.003 (-0.14, 0.14)	0.97
	Sometimes	0.12 (-0.17, 0.41)	0.42	-0.04 (-0.38, 0.31)	0.83
	Seldom/Never	-0.01 (-0.13, 0.10)	0.83	0.04 (-0.09, 0.18)	0.54
Dairy	Often/Always	0.01 (-0.17, 0.20)	0.87	-0.04 (-0.24, 0.16)	0.70
	Sometimes	-0.01 (-0.27, 0.25)	0.94	0.01 (-0.27, 0.28)	0.94
	Seldom/Never	-0.04 (-0.15, 0.07)	0.48	-0.10 (-0.24, 0.03)	0.12

Adjusted for urinary creatinine, total Kcal, recall-specific Healthy Eating Index (HEI), race/ethnicity, and physical activity level. AMPA: aminomethylphosphonic acid

Fruit (adjusted p = 0.03) and corn (adjusted p = 0.008) consumption were associated with lower AMPA levels (Table 2.5, Figure 2.2). Total grains and refined grains were positively associated with AMPA, but not after adjustment for race/ethnicity, physical activity, and dietary pattern. Consumption of soy protein was significantly associated with elevated AMPA (adjusted p = 0.02, Figure 2.2), but this relationship was attenuated when other forms of soy (such as soy milk and soy sauce) were included (adjusted p = 0.32). Women who reported consuming alcohol, particularly wine, on the day of the dietary recall had higher AMPA levels even after adjusting for dietary pattern, race/ethnicity, and physical activity (adjusted p <0.001, Figure 2.2). Women who were not of white, Asian, or Hispanic

race/ethnicity had significantly lower AMPA levels (Table 2.6, adjusted p = 0.02). Dining at fast food restaurants 4 or more times per week was associated with elevated AMPA (Table 2.6, adjusted p = 0.006). No food groups were significantly associated with AMPA levels in the organic-eating-stratified analysis (Table 2.7).

After adjusting for creatinine, dietary pattern, race/ethnicity, and physical activity, self-reported organic eating was no longer significantly associated with either glyphosate or AMPA, although concentrations of both analytes were non-significantly higher in those who reported seldom or never eating organic food (Table 2.5, Figure 2.2).

The sensitivity analysis excluding potentially problematic recalls was generally consistent with the main results (Appendix A, Table A4). The relationships between glyphosate concentration and physical activity and between AMPA concentration and fruit and soy protein consumption were suggestive but no longer statistically significant ($p > 0.05$ but <0.1 for all 3 relationships). All other variables that were significantly associated with glyphosate or AMPA in the main results remained statistically significant in the sensitivity analysis. Additionally, adjusting for self-reported organic food consumption frequency did not materially change the results (data not shown).

2.4 Discussion

This was the largest study to date pairing data from dietary recalls with measurements of urinary glyphosate and its metabolite AMPA. We observed elevated glyphosate in women who consumed more grains, even those who usually ate organic grains. We also observed that consuming alcohol (particularly wine) and soy protein were associated with higher AMPA concentrations, while fruit and corn consumption were

associated with lower AMPA. These findings were robust to adjustment for dietary pattern and other factors associated with glyphosate and AMPA concentrations.

These results suggest probable sources of glyphosate and AMPA in the American diet and broadly agree with the few previous studies done in other countries. Studies in Germany[69] and Slovenia[70] showed positive relationships between glyphosate levels and the consumption of pulses, nuts, and brown rice. In our study, glyphosate levels were higher among those who had consumed legumes and nuts/seeds on the day before providing their urine sample, although this association was just over the statistical significance threshold (adjusted p = 0.06 for legumes and 0.08 for nuts/seeds). We did not examine rice separate from other grains due to limitations of the study instrument, but our finding of elevated glyphosate in those who consumed more grains is consistent with these results. Compared to both other studies, a larger proportion of our study cohort had quantifiable levels of both analytes, likely due to several factors. First, our assay was more sensitive (LOD of 0.014 ng/mL for glyphosate and 0.013 ng/mL for AMPA compared to >0.05 in both other studies). Also, dietary habits and the presence of glyphosate and/or AMPA in the food supply likely differ between their European and our American populations. The only studies to examine dietary factors associated with urinary glyphosate or AMPA in the U.S. found associations between urinary glyphosate and consumption of tea[68] and caffeinated beverages[46], findings not replicated in our study. The sample sizes for these other studies were small (N = 36 and N = 77, respectively) and the study populations (undergraduate students and pregnant women) were markedly different from our cohort of postmenopausal women, in addition to their use of questionnaires instead of validated 24-hour dietary recalls.

Those who reported themselves to be frequent organic food consumers had a higher level of educational attainment, lower BMI, and consumed less fast food compared to those who infrequently ate organic food. Frequent organic eaters also adhered more closely to U.S. nutritional guidelines, consuming more fruit and legumes and less grains and meat, although they were also more likely to drink alcohol. These results complement previous research describing characteristics associated with more frequent organic eating[112,150–153].

Previous data are mixed regarding the relationship between organic food consumption and glyphosate concentrations, with some observational studies showing no statistically significant relationship between self-reported organic eating behaviors and glyphosate levels[46,154]. However, a small dietary intervention study showed a decrease in median urinary glyphosate from 0.51 to 0.12 ng/mL after 5 days of a completely organic diet[61]. In our study, we saw a more modest, but statistically significant, difference between those who reported often/always eating organic food compared to seldom/never (median glyphosate of 0.09 ng/mL vs. 0.13 ng/mL, respectively, $p = 0.04$). However, this relationship was no longer statistically significant after adjusting for dietary pattern and demographic/behavioral variables, suggesting that characteristics of individuals who frequently consume organic food may be a source of the inconsistency between observational and experimental data. Given the strong relationship between grain consumption and glyphosate levels that we observed, at least some of the elevated glyphosate levels in seldom/never organic consumers may be explained by increased consumption of glyphosate-rich grains. Additionally, typical dietary habits and purchasing behaviors will likely differ in the real world compared to a controlled experimental setting, likely attenuating relationships between organic eating and glyphosate levels.

Different dietary factors were associated with glyphosate and AMPA in our study, suggesting distinct sources of exposure. Humans are not known to possess the necessary enzymes to metabolize glyphosate, so AMPA excreted by humans is thought to derive primarily from direct consumption of food contaminated with AMPA. There is also evidence that species of the gut microbiome could convert glyphosate to AMPA[155,156], which has been observed in rats[157,158]. One study in humans measured low levels of AMPA in urine after oral exposure to glyphosate, but it was unclear whether this was from glyphosate metabolism in the gut or background exposure levels[59]. Most glyphosate used in the U.S. is applied to glyphosate-resistant (GR) crops[39], which can withstand treatment with glyphosate long enough for breakdown to AMPA to occur[138], so these crops are expected to have higher concentrations of AMPA. Current commercially available GR food crops include soybean, corn, and canola. We did not evaluate canola in our study because the dietary recall instrument could not accurately distinguish the primary source of canola in the American diet (canola oil) from other oils. In 2016, over 80% of soybeans and corn grown in the U.S. were GR varieties[40]. The widespread adoption of glyphosate-resistant soybeans is consistent with our finding that soy protein consumption was associated with higher AMPA concentrations. Previous studies[132,137,138,159] as well as monitoring by the USDA and FDA[62,160] have identified glyphosate and AMPA residues in soybean crops. However, the opposite of the expected effect was seen for corn: those who consumed corn had significantly lower AMPA levels. The cause of this inverse relationship is unclear. Some evidence suggests[132] that GR corn does not accumulate AMPA as readily as other GR crops, or that AMPA produced by GR corn is further broken down into other products. One study found no traces of AMPA in seeds of GR corn[136] and another did not detect AMPA in either standard or GR corn treated

with glyphosate[159]. The USDA and FDA did not test corn for AMPA. Therefore, a possible explanation for our finding is that consumption of corn displaced consumption of more AMPA-rich foods, such as wheat and other cereals.

Glyphosate is widely used as a pre-harvest desiccant on non-corn grain crops, including wheat and oats[39], which likely explains the higher glyphosate concentrations in women who ate more grains. AMPA was also elevated in frequent grain consumers in our study, but this relationship was attenuated after adjustment for dietary, demographic, and behavioral factors. In addition to the frequent detection of glyphosate in samples of cereal crops, AMPA has also been detected, although levels of both were almost always below maximum residue levels (MRLs) permitted by regulatory agencies[63–65]. A Canadian study detected glyphosate, but not AMPA, in wheat samples and determined that it survives the milling and baking process; concentrations were higher in whole-grain flour[161]. Although the use of glyphosate is prohibited on organic crops, participants who consumed more grains and reported often or always eating organic grains also had elevated glyphosate concentrations, suggesting either incomplete avoidance of conventional grains or the presence of glyphosate in organic grains. One study detected glyphosate in 25% of organic food products compared to 49% of conventional products in Canada, mostly at low concentrations consistent with drift from nearby conventional fields[63].

Despite consumer concerns over pesticide contamination in produce, we did not observe an elevated level of glyphosate or AMPA associated with fruit or vegetable consumption. In fact, there was an inverse relationship between total fruit consumption and AMPA levels. Given that less than 2% of the glyphosate applied to U.S. crops is used on fruit and vegetable acreage[39], this relationship may be due to fruit replacing foods higher in AMPA

in the diet. A similar cause could explain the inverse relationship between glyphosate levels and meat consumption; glyphosate is infrequently detected in meat[63,64]. Another possible explanation may involve the gut microbiome. One study found that some species of gut bacteria possess C-P lyase, which is responsible for the degradation of AMPA[162], and fruit consumption can influence the makeup of the gut microbiome[163]. Additionally, those who reported often or always eating organic food consumed more fruit in our cohort, although analyses adjusted for organic eating habits showed the same results.

There was a strong relationship between alcohol consumption, particularly wine, and AMPA levels, which is especially interesting given the observation that those who reported seldom or never eating organic food also consumed less alcohol. Glyphosate is used in grape cultivation for weed management[164]; in 2012, 43% of grape acres in the U.S. were treated with glyphosate[39]. Few studies have examined wine for glyphosate or AMPA residues. One Swiss study detected glyphosate in 100% and AMPA in 19% of wine samples tested, albeit at very low concentrations[64]. Testing by the U.S. Public Interest Research Group (PIRG) detected glyphosate in all 5 samples of wines tested[165]. If glyphosate is present in wine, the microbial activity which drives the fermentation process could metabolize some into AMPA, but evidence for this is limited. One study observed degradation of glyphosate by brewer's yeast (*S. cerevisiae*) but did not confirm that AMPA was the product[166]. Further research is needed to determine the presence of AMPA in wine and other alcoholic beverages in the U.S. market.

This study has several notable strengths. First, the 24-hour dietary recalls were paired to urine samples provided the following morning, which allowed accurate assessment of dietary exposures most likely to influence glyphosate/AMPA levels given their short half-

lives. Most participants provided two paired urine samples/dietary recalls, which provides a more accurate representation of typical diet habits compared to a single sample. We also asked detailed questions about usual organic eating behaviors, including for specific food groups.

However, participants did not report the organic status of the specific food items that they reported on the dietary recalls, so it is impossible to know whether the two 24-hour periods are reflective of their self-reported organic eating behaviors. Additionally, our study cohort was relatively homogeneous with respect to age, sex, and place of residence, which may limit generalizability of the results to other populations.

In the largest study to date of dietary habits and urinary glyphosate and AMPA, there was elevated glyphosate in those who ate more grains and elevated AMPA associated with soy and alcohol consumption. Fruit and vegetable (including corn) consumption is not likely to be a major source of glyphosate or AMPA in the American diet. Although more frequent organic eating was associated with lower glyphosate and AMPA levels in the urine, this relationship was not statistically significant after adjusting for other factors associated with organic eating habits. Reducing grain consumption may help consumers avoid exposure to glyphosate through the diet, but it remains unclear whether there are health effects associated with chronic exposure to glyphosate and AMPA.

Chapter 3: Association of Glyphosate Exposure with Blood DNA Methylation

3.1 Introduction

Glyphosate is the most used pesticide in the world[39], used in both agricultural (for weed control and pre-harvest desiccation) and non-agricultural settings[39,40,167]. Studies have detected glyphosate in the air, soil, drinking water, and food[140]. Use of glyphosate-based herbicides has increased dramatically since their introduction[39], largely due to the growing use of genetically modified glyphosate-resistant crops starting in the late 1990s[168]. Glyphosate and its primary metabolite, aminomethylphosphonic acid (AMPA), are frequently detected in the food supply[62–64,66,160], and recent studies in the U.S. have detected glyphosate in the urine of 70-90% of participants[43–46].

Concerns have been raised about the safety of glyphosate exposure in humans, and the topic remains controversial. In 2015 the International Agency for Research on Cancer (IARC) classified glyphosate as a probable human carcinogen[169]. This classification is supported by a meta-analysis which suggests an elevated risk of non-Hodgkin's lymphoma associated with glyphosate exposure[52], although another large cohort study did not find a relationship between glyphosate and cancer risk[170]. Epidemiologic studies have also found associations with other health problems including shortened gestational length[46], birth defects[171], endocrine disruption[44], and thyroid dysfunction[172]. Animal and in vitro studies have shown associations with endocrine disruption[54,173], fatty liver disease[53], breast cancer cell proliferation[58], changes to the microbiome[55–57,174], and increased risk of antibiotic resistance in bacteria[175]. The potential effects of AMPA exposure in humans are even less

well understood, although a recent study found elevated breast cancer risk in women with high urinary AMPA[50].

Epigenetic markers, such as DNA methylation, may be a powerful tool for understanding the potential effects of glyphosate exposure in humans. White blood cell DNA methylation has been associated with environmental exposures[176,177], including endocrine-disrupting compounds[25,26], other pesticides[27], and air pollution[28–30]. In addition, DNA methylation indices combining information from multiple sites have been developed as biomarkers of some exposures, including smoking and alcohol consumption[31–33]. Some DNA methylation markers have also been shown to be associated with disease risk[36,178].

We conducted an epigenome-wide association study to identify DNA methylation loci associated with urinary glyphosate and AMPA levels. The secondary goal was to develop blood DNA methylation indices to predict urinary glyphosate and AMPA levels. Discovery and validation of such loci may advance our understanding of the potential biological mechanisms linking glyphosate and AMPA exposure to human health outcomes and provide tools for measuring their longer-term exposures.

3.2 Methods

Study population

The study population consisted of postmenopausal women residing in southern California between the ages of 45-66 years with no personal history of breast cancer or mastectomy (N = 392). All participants provided informed consent to participate and the study was approved by the University of California, Irvine Institutional Review Board, HS #2016-3127.

Biospecimen collection and processing

Recruitment, questionnaires, and specimen collection were described in detail previously[144]. During the 10 days prior to blood sample collection, participants collected first-morning urine samples on two days not known to them in advance in order to capture their typical behaviors and exposures. Urine samples were stored in a freezer (-20°C) until they could be transported to the laboratory. Within 6 hours of peripheral blood collection, DNA was extracted from the buffy coat using the QIAamp DNA Blood Maxi Kit (Cat No. 51194, QIAGEN, Hilden, Germany) and stored at -80°C for later analysis.

Urinary glyphosate and AMPA were measured using liquid chromatography with tandem mass spectrometry (LC-MS/MS) at the Collaborative Center for Translational Mass Spectrometry (CCTMS) (TGen, Phoenix, AZ) using a Vanquish UHPLC coupled to TSQ Altis triple quadrupole mass spectrometer (Thermo Scientific, San Jose, CA) as previously described[44]. Briefly, glyphosate and AMPA assay validation was performed in a commercially available urine pool from LeeBio Solutions (Maryland Heights, MO) prior to analysis. The calibration curves for glyphosate and AMPA were prepared over a linear range of 0-5 ng/mL (coefficient of determination $R^2 > 0.99$) by spiking variable concentrations of glyphosate and AMPA and their respective isotopically labeled internal standards, $^{13}C_2{}^{15}N$-Glyphosate and $D_2{}^{13}C^{15}N$-AMPA (Sigma-Aldrich) at a fixed concentration of 50 ng/mL. Data acquisition and processing were performed using Xcalibur 4.1.50 and Quanbrowser, 4.1.50 (Thermo Scientific). Both assays were linear (coefficient of determination $R^2 > 0.99$) over a range 0-5 ng/mL (Appendix B, Figure B1). The limits of detection (LOD) for glyphosate and AMPA were 0.014 and 0.013 ng/mL, and the limits of quantitation (LOQ) were 0.041 and 0.040 ng/mL, respectively (Appendix A, Table A2). Creatinine was measured using the DetectX urinary

creatinine detection kit (Arbor Assays, Ann Arbor, MI, K002-H5) according to the manufacturer's instructions. The data were tested for batch effects using the Kruskal-Wallis test, and batch correction was performed for glyphosate values using the removeBatchEffects function in limma[145]. Four samples with implausibly low creatinine (<10 mg/dL) were excluded from subsequent analysis. Measurements were averaged for the two urine samples for each analyte after replacing values <LOD with LOD/$\sqrt{2}$[146].

Genomic DNA was bisulfite converted using the Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA, USA) and then DNA methylation at over 850,000 CpG sites was measured using the Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA, USA) at the University of Southern California Molecular Genomics Core. Laboratory staff were blinded to glyphosate and AMPA measurements. All samples passed internal controls and were included in the final data analysis. Methylation data pre-processing was performed according to recommended steps[128], including probe filtering, normalization, and batch correction. All data analysis was performed in R, version 3.6.2[149]. Probes with a detection p-value >0.05 were considered missing, and other low-quality probes were removed as follows: missing in at least 20% of samples (n = 648); had SNPs with global minor allele frequency >1% within 5 base pairs of the target sequence or mapping problems with the probe sequence (n = 99,109)[179]; hybridize to multiple locations (n = 15)[180]; or located on the X or Y chromosome (n = 16,927). Values were normalized with noob normalization as implemented in the minfi package (version 1.32.0)[181] for background correction and dye bias adjustment followed by beta mixture quantile normalization (BMIQ) to correct for type II probe bias[182]. Except where noted, filtering and normalization was completed using the ChAMP package, version 2.12.4[183]. Finally, DNA methylation measurements were corrected

for batch and position on chip using ComBat[184] as implemented in sva version 3.30.1[185]. We used a reference-based method[186] to estimate the proportions of six white blood cell types in our samples, which did not significantly differ with glyphosate or AMPA tertile (Appendix A, Table A5). We used the methylation M value (the base 2 logarithm of the ratio of methylated to unmethylated intensities) for all analyses; in some cases, we also report β values (percent methylation) for ease of interpretation[187].

Statistical analysis

Variability and correlation of urinary glyphosate and AMPA levels between and within samples were characterized with the intraclass correlation (ICC) and Pearson's correlation, respectively. Relevant covariates, including age, self-reported race/ethnicity, body mass index (BMI), smoking status (current, former, or never smoker), alcohol consumption, organic eating habits, physical activity, and recent herbicide use were collected via questionnaires. Organic eating habits were self-reported as "seldom or never," "sometimes," or "often or always"[148]. Physical activity was reported as typical frequency and duration of moderate and vigorous physical activity. These responses were converted to minutes per week of moderate exercise, with each minute of vigorous exercise equivalent to two minutes of moderate exercise[118]. Diet quality was estimated by using up to three unannounced ASA24 24-hour dietary recalls[119], completed during the 10 days prior to blood collection, to calculate the Healthy Eating Index-2015 (HEI) averaged across all recalls. The HEI is scored on a 0-100 scale with higher values indicating greater adherence to the Dietary Guidelines for Americans[147]. Some participants were asked whether they had used herbicides at home or work within the past seven days. This question was added to the questionnaire partway through study recruitment, so data on recent herbicide use is missing

for 37% (n = 144) of participants. The association between each of these variables and the natural logarithm of glyphosate and AMPA concentration (adjusted for urinary creatinine) was assessed using linear regression. In all linear regression analyses, two-sided p-values <0.05 were considered statistically significant.

DNA methylation analysis

Epigenetic age acceleration (residuals from a model regressing chronologic age on epigenetic age according to Horvath's clock[17]) was examined for association with glyphosate and AMPA, first in univariate models and then adjusted for age, race/ethnicity, BMI, smoking status, alcohol consumption, organic eating, HEI, and urinary creatinine levels, in addition to batch, position on chip, and white blood cell types. These covariates were selected based on known relationships with either DNA methylation, glyphosate/AMPA, or both; creatinine was included to account for differences in urine concentration.

For probe- and region-specific analyses, the study samples were divided into training (N = 332) and validation (N = 60) sets stratified by glyphosate tertile (Figure 3.1). Demographic and dietary variables were compared for the training and validation sets (Appendix A, Table A6). All variables except smoking were not significantly associated with the randomly assigned set; there was a larger proportion of former smokers in the validation set (p = 0.03). Because all of the former smokers in the validation set reported cessation >20 years prior to blood sample collection, we proceeded with the planned analysis. Using the training set, we identified differentially methylated probes (DMPs) and differentially methylated regions (DMRs) according to a resampling-based method[188]. The development of this method is described in detail in Appendix C.

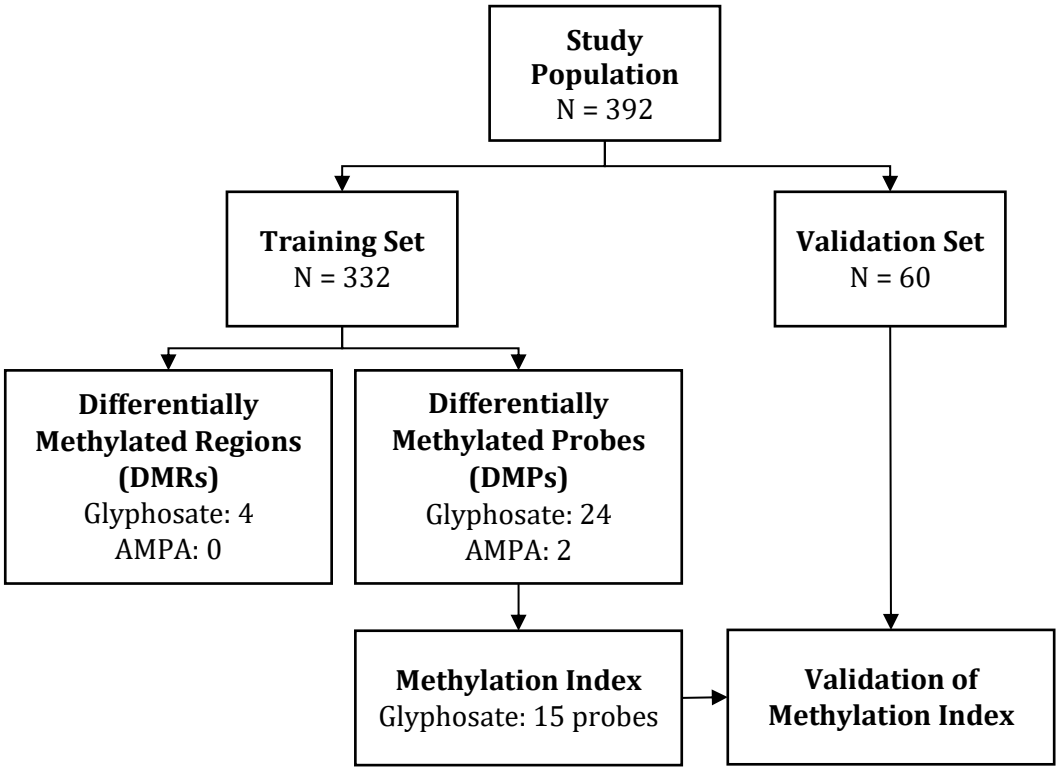

Figure 3.1. Flow chart for epigenome-wide association study of glyphosate and AMPA

We selected a random subsample consisting of 90% of the training set (N = 299). Using this subsample, we identified candidate DMPs using limma, version 3.44.3[145], to fit linear models and candidate DMRs using DMRcate, version 2.0.7[189], with a kernel bandwidth of 1000 base pairs and scaling factor of 2. Each model was adjusted for the same variables as in the epigenetic aging analysis. Models without adjustment for dietary variables (organic eating and HEI) did not show substantially different results (Pearson's R >0.99 for test statistics from the two versions of the model for both analytes). A false discovery rate (FDR) q-value of < 0.1 was considered statistically significant. This process was repeated in a series of 1,000 subsamples of 90% of the training set and sampling distributions for all summary statistics constructed. Probes and regions that were selected as candidate DMPs or DMRs in ≥90% of subsamples were considered differentially methylated and carried forward for

further analysis. This approach results in a more stable list of DMPs and DMRs, since DNA methylation microarray results are known to be sensitive to small differences in the study cohort[190–193]. The median FDR q-value was used to rank the relative significance of DMPs. For DMRs, overlapping regions were combined and then significance was ranked after combining median q-values for the probes within the region using Stouffer's method.

DMPs and DMRs were annotated using the Illumina manifest (version B4) to identify associated genes and genomic context. Probes were also mapped to ChromHMM data from ENCODE[194] to determine predicted chromatin state; data from the GM12878 lymphoblastic cell line was used as it is the tissue most similar to leukocytes. Chromatin states were pooled into six categories as follows: promoter (active, weak, or poised promoters), enhancer (strong or weak enhancers), transcribed (transcriptional transition, transcriptional elongation, or weak transcribed), polycomb-repressed, inactive (heterochromatin and repetitive regions), and insulators. DMPs were examined for enrichment of certain locations relative to CpG islands or chromatin states, compared to the background of all probes included on the array, using Fisher's exact test. We used the "gometh" function from the missMethyl package, version 1.16.0[195], to examine enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and gene ontology terms in genes associated with DMPs, and assessed enrichment of pathways using Ingenuity Pathway Analysis software, version 62089861 (Release Date: 2021-02-17). Terms or pathways with $p<0.01$ and at least 3 genes associated with DMPs were considered enriched.

We developed methylation indices to predict urinary glyphosate and AMPA levels using blood DNA methylation at the glyphosate- and AMPA-associated DMPs. Methylation (β) values for each DMP were passed to a least absolute shrinkage and selection operator

(LASSO) regression model with 10-fold cross validation repeated 100 times to estimate lambda (the penalty coefficient) and regression coefficients for each site. This was implemented using the glmnet package, version 2.0-18[196]. The performance of the methylation indices were assessed by calculating the Pearson correlation between predicted and actual glyphosate and AMPA concentration in the 60 validation samples. The indices' relationship with glyphosate and AMPA tertile was also tested using ANOVA and the area under the receiver operating characteristic curve (AUC) computed to characterize the discriminatory accuracy for the highest vs. lowest tertile.

3.3 Results

Glyphosate and AMPA concentrations

96% of study participants had detectable (>LOD) glyphosate and 82% had detectable AMPA in at least one urine sample. Table 3.1 describes glyphosate and AMPA concentrations stratified by various demographic, dietary, and lifestyle factors. The median glyphosate concentration (averaged between two urine samples for each participant) was 0.12 ng/mL (range <LOD-1.65 ng/mL) and the median AMPA concentration was 0.06 ng/mL (range <LOD-1.36 ng/mL). The maximum concentration in a single urine sample was 3.01 ng/mL for glyphosate and 1.51 ng/mL for AMPA. Both glyphosate and AMPA concentrations were strongly right-skewed and thus transformed with the natural logarithm for all further analyses (Figure 3.2). Between-sample agreement was moderate for both glyphosate (ICC = 0.53) and AMPA (ICC = 0.34), as was the within-sample correlation between glyphosate and AMPA measurements (Pearson's R = 0.48, $p < 0.001$) (Figure 3.3).

Table 3.1. Median glyphosate and AMPA, stratified by cohort characteristics

		Glyphosate			AMPA		
	N (%)	ng/mL Median (IQR)	µg/g creatinine Median (IQR)	p	ng/mL Median (IQR)	µg/g creatinine Median (IQR)	p
Overall	392	0.12 (0.06, 0.22)	0.20 (0.11, 0.38)		0.06 (0.02, 0.12)	0.10 (0.04, 0.22)	
Race/Ethnicity							
Asian	43 (11.1)	0.11 (0.07, 0.23)	0.19 (0.09, 0.34)	0.30	0.08 (0.02, 0.14)	0.13 (0.03, 0.24)	0.50
Hispanic	69 (17.9)	0.12 (0.07, 0.21)	0.20 (0.11, 0.37)	0.90	0.06 (0.02, 0.11)	0.08 (0.04, 0.17)	0.41
Other	20 (5.2)	0.07 (0.04, 0.13)	0.11 (0.06, 0.24)	**0.004**	0.02 (0.01, 0.05)	0.04 (0.02, 0.07)	**0.005**
White	254 (65.8)	0.12 (0.06, 0.24)	0.21 (0.12, 0.42)	Ref	0.05 (0.02, 0.12)	0.10 (0.04, 0.22)	Ref
Age (years)							
45-49	27 (6.9)	0.18 (0.09, 0.24)	0.23 (0.11, 0.33)	0.72	0.11 (0.05, 0.19)	0.12 (0.05, 0.29)	0.12
50-54	85 (21.7)	0.14 (0.06, 0.19)	0.18 (0.12, 0.33)		0.07 (0.02, 0.11)	0.09 (0.04, 0.17)	
55-59	155 (39.5)	0.11 (0.06, 0.21)	0.18 (0.10, 0.33)		0.05 (0.02, 0.12)	0.08 (0.03, 0.22)	
60-66	125 (31.9)	0.11 (0.06, 0.29)	0.23 (0.11, 0.56)		0.05 (0.01, 0.12)	0.10 (0.04, 0.22)	
BMI (kg/m^2)							
<25	188 (48.0)	0.12 (0.06, 0.24)	0.21 (0.10, 0.44)	0.78	0.06 (0.02, 0.12)	0.10 (0.04, 0.21)	0.63
25-29.9	116 (29.6)	0.12 (0.06, 0.22)	0.20 (0.12, 0.33)		0.05 (0.02, 0.12)	0.09 (0.04, 0.20)	
≥30	88 (22.4)	0.13 (0.07, 0.19)	0.19 (0.12, 0.31)		0.06 (0.01, 0.13)	0.09 (0.04, 0.22)	
Smoking Status							
Never	285 (72.9)	0.12 (0.06, 0.22)	0.22 (0.11, 0.38)	Ref	0.06 (0.02, 0.13)	0.11 (0.04, 0.22)	Ref
Former	89 (22.8)	0.10 (0.05, 0.22)	0.17 (0.10, 0.38)	0.52	0.04 (0.01, 0.09)	0.07 (0.03, 0.15)	**0.04**
Current	17 (4.3)	0.11 (0.05, 0.21)	0.17 (0.09, 0.33)	0.60	0.05 (0.02, 0.14)	0.09 (0.03, 0.26)	0.59
Alcohol (drinks/week)							
None	107 (27.4)	0.12 (0.06, 0.22)	0.22 (0.11, 0.48)	Ref	0.05 (0.02, 0.11)	0.08 (0.04, 0.22)	Ref
1 or fewer	161 (41.2)	0.12 (0.05, 0.24)	0.19 (0.09, 0.37)	0.14	0.06 (0.01, 0.12)	0.08 (0.03, 0.18)	0.49
2-6	72 (18.4)	0.14 (0.07, 0.19)	0.23 (0.14, 0.41)	0.92	0.06 (0.03, 0.12)	0.13 (0.04, 0.24)	0.64
7 or more	51 (13.0)	0.10 (0.05, 0.21)	0.16 (0.12, 0.25)	0.17	0.07 (0.03, 0.15)	0.12 (0.06, 0.21)	0.16
Organic Eating							
Seldom/Never	124 (31.7)	0.14 (0.07, 0.23)	0.21 (0.13, 0.38)	Ref	0.07 (0.03, 0.13)	0.12 (0.05, 0.23)	Ref
Sometimes	114 (29.2)	0.12 (0.05, 0.24)	0.19 (0.11, 0.38)	0.47	0.06 (0.02, 0.13)	0.10 (0.04, 0.25)	0.44
Often/Always	153 (39.1)	0.10 (0.06, 0.20)	0.19 (0.09, 0.38)	0.05	0.04 (0.01, 0.11)	0.07 (0.03, 0.18)	**0.01**

Table continues

Table 3.1. Median glyphosate and AMPA, stratified by cohort characteristics (continued)

		Glyphosate			AMPA		
	N (%)	ng/mL Median (IQR)	µg/g creatinine Median (IQR)	p	ng/mL Median (IQR)	µg/g creatinine Median (IQR)	p
HEI							
Quartile 1	93 (25.1)	0.13 (0.07, 0.22)	0.22 (0.13, 0.37)	0.99	0.07 (0.02, 0.12)	0.12 (0.05, 0.22)	**0.03**
Quartile 2	93 (25.1)	0.10 (0.05, 0.19)	0.19 (0.09, 0.35)		0.07 (0.02, 0.13)	0.13 (0.04, 0.26)	
Quartile 3	92 (24.8)	0.11 (0.06, 0.22)	0.17 (0.10, 0.41)		0.05 (0.02, 0.12)	0.08 (0.04, 0.16)	
Quartile 4	93 (25.1)	0.12 (0.06, 0.25)	0.24 (0.11, 0.54)		0.04 (<LOD, 0.12)	0.06 (0.03, 0.21)	
Physical Activity							
≥150 mins/week	153 (40.7)	0.10 (0.05, 0.24)	0.19 (0.09, 0.38)	**0.03**	0.05 (0.01, 0.11)	0.08 (0.04, 0.19)	0.08
<150 mins/week	223 (59.3)	0.13 (0.07, 0.22)	0.22 (0.12, 0.41)		0.06 (0.02, 0.13)	0.11 (0.04, 0.23)	
Herbicide Use (past week)							
Yes	21 (8.5)	0.14 (0.08, 0.20)	0.22 (0.11, 0.44)	0.61	0.04 (<LOD, 0.11)	0.11 (0.04, 0.16)	0.39
No	227 (91.5)	0.12 (0.06, 0.24)	0.21 (0.11, 0.38)		0.06 (0.02, 0.12)	0.10 (0.03, 0.20)	

P-values are from linear regression with the natural logarithm of glyphosate or AMPA as outcome, adjusted for urinary creatinine. Age, body mass index (BMI), and Healthy Eating Index (HEI) were evaluated as continuous variables; results were comparable when evaluated as categorical variables. Missing data: 6 for race/ethnicity, 1 each for smoking, alcohol, and organic eating, 21 for HEI, 16 for physical activity, and 144 for herbicide use. AMPA: aminomethylphosphonic acid

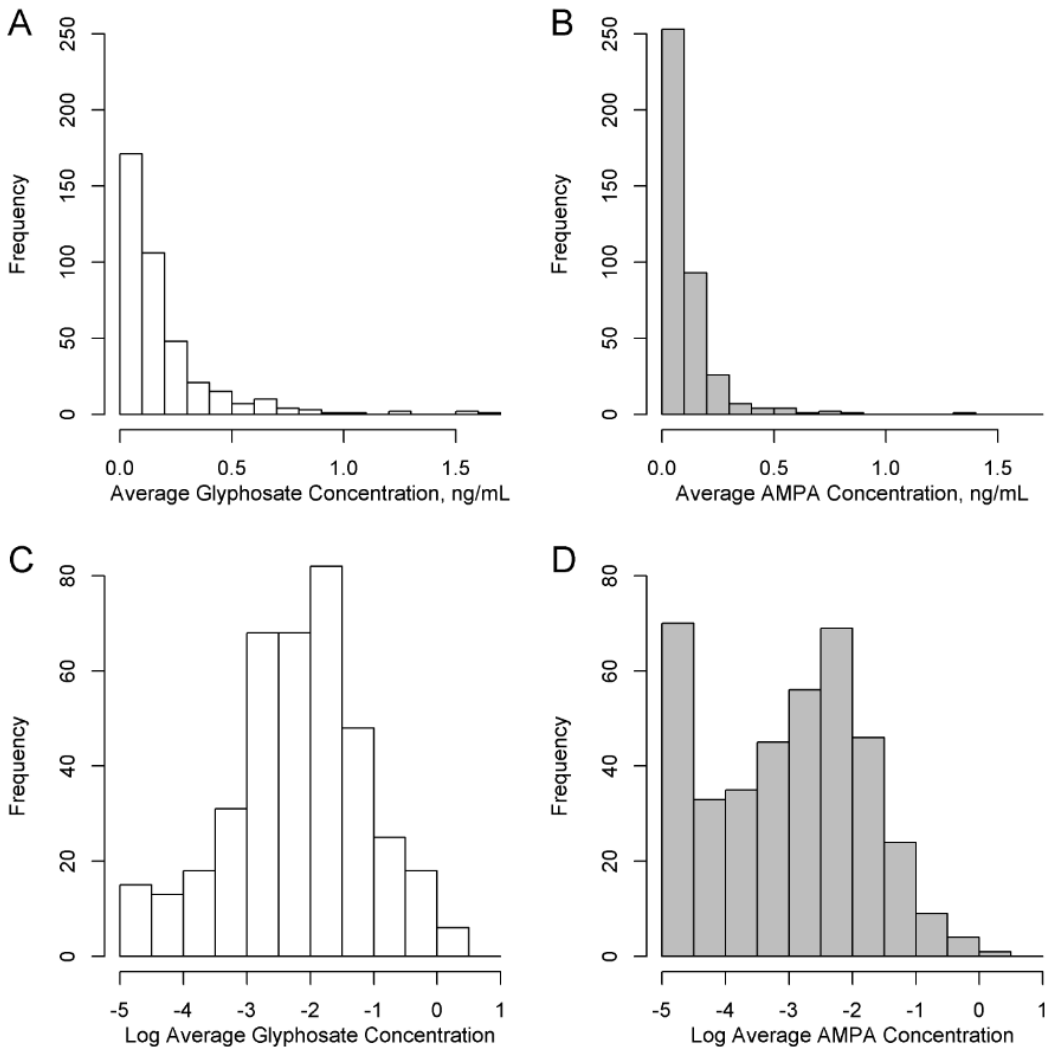

Figure 3.2. Histograms of urinary glyphosate and AMPA concentration
Averaged for two samples from each individual. A) glyphosate, B) AMPA, C) natural log of glyphosate, D) natural log of AMPA. AMPA: aminomethylphosphonic acid

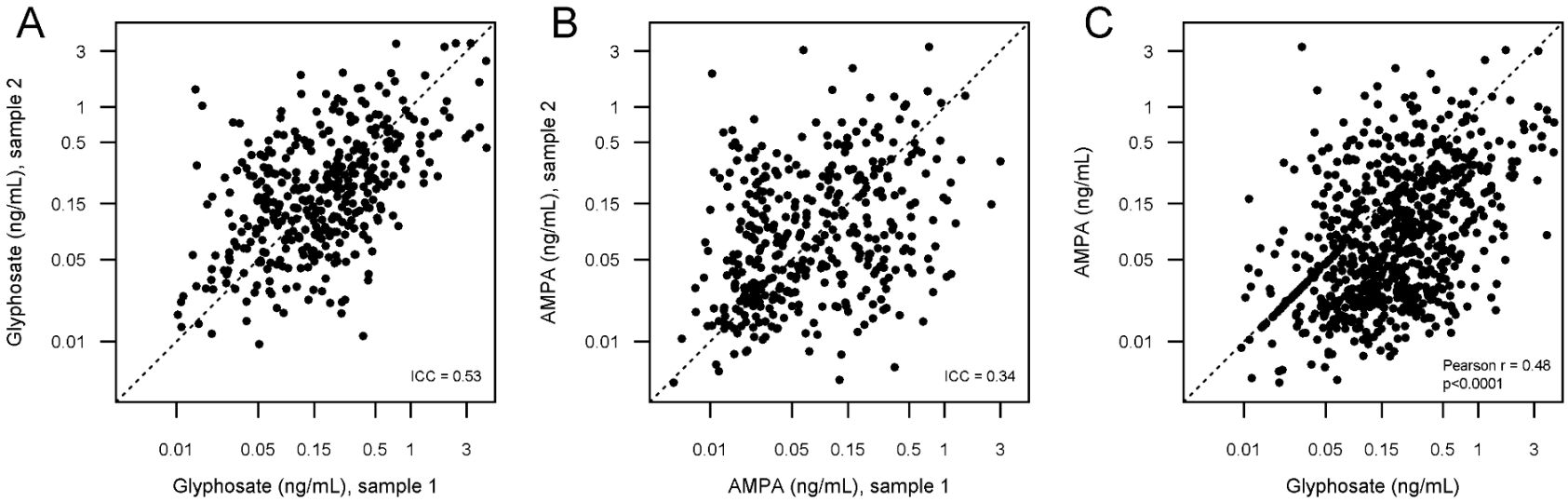

Figure 3.3. Scatter plots of between- and within-sample glyphosate and AMPA concentrations
Between-sample concentrations, A) glyphosate, B) AMPA, from n = 388 individuals with paired urine samples. C) Correlation between glyphosate and AMPA concentrations within-samples (n = 780 samples). All concentrations are transformed with the natural logarithm and divided by the urinary creatinine concentration to account for per-sample differences in urine concentration.

Women who were not of white, Asian, or Hispanic race/ethnicity had lower concentrations of both glyphosate ($p = 0.004$) and AMPA ($p = 0.005$). Women who reported "often or always" eating organic food had lower glyphosate ($p = 0.05$, not significant) and AMPA ($p = 0.01$). Lower diet quality was associated with elevated AMPA ($p = 0.03$), but not glyphosate ($p = 0.99$). Those who met the Physical Activity Guidelines for Americans (≥150 minutes of moderate-intensity exercise per week) had lower glyphosate ($p = 0.03$) and AMPA ($p = 0.08$, not significant) levels. Former smokers had marginally lower AMPA concentrations ($p = 0.04$). None of the other variables examined (age, BMI, alcohol consumption, or herbicide use) were associated with urinary glyphosate or AMPA.

DNA methylation analysis

Neither glyphosate nor AMPA was associated with differences in epigenetic age acceleration in the univariate model, but AMPA was significantly ($p = 0.02$) associated in the fully adjusted model (Table 3.2). In the adjusted model, a one-unit higher AMPA concentration (natural logarithm-transformed) was associated with greater epigenetic age acceleration of 0.49 years.

Twenty-four probes associated with urinary glyphosate concentration and two with urinary AMPA were identified in ≥90% of subsamples and considered differentially methylated (Table 3.3). Seventeen of the 24 probes (71%) were hypomethylated with higher glyphosate (Figure 3.4), while both AMPA-associated probes were hypermethylated. Glyphosate-associated probes with the smallest median p-values were located within the *PEX26*, *SF3B2*, and *CHMP1A* genes. The largest effect sizes (methylation difference between glyphosate tertiles) were observed in probes within the *VMO1*, *KCP*, and *QARS1* genes.

Table 3.2. Association of epigenetic age acceleration with the natural logarithm of glyphosate and AMPA concentration

	Glyphosate			AMPA		
	Coefficient (95% CI)	p	Adjusted R^2	Coefficient (95% CI)	p	Adjusted R^2
Univariate	-0.13 (-0.49, 0.24)	0.50	-0.0014	0.22 (-0.11, 0.55)	0.20	0.0017
Fully Adjusted	-0.35 (-0.77, 0.079)	0.11	0.086	0.49 (0.091, 0.89)	0.02	0.086

The fully adjusted model is adjusted for age, race/ethnicity, body mass index, smoking status, alcohol consumption, self-reported organic eating, diet quality (Healthy Eating Index), and urinary creatinine levels, in addition to batch, position on chip, and estimated white blood cell type proportions. AMPA: aminomethylphosphonic acid

Table 3.3. Differentially methylated probes (DMPs) associated with urinary glyphosate and AMPA concentration identified by resampling-based method

Probe	Frequency	Median p	Median FDR q	Delta-β	Mean methylation (β) by glyphosate or AMPA tertile in training set			Chromosome	Position (hg19)	Gene(s)	Location within gene(s)	CpG island	Predicted chromatin state	On 450k array
					Tert. 1	Tert. 2	Tert. 3							
Glyphosate														
cg13499896	1000	1.74E-8	0.006	-0.71	97.57	97.33	97.19	22	18,571,657	*PEX26*	3'UTR	Open Sea	Str. Enhancer	N
cg24576174	1000	4.24E-8	0.010	-0.60	98.21	98.07	97.96	11	65,828,083	*SF3B2*	Body	Open Sea	Trans. Elong.	N
cg00355690	996	5.64E-8	0.011	-1.89	93.39	93.14	92.86	16	89,722,417	*CHMP1A*	Body	N_Shore	Wk. Enhancer	Y
cg26833395	1000	5.76E-8	0.012	-0.11	0.97	0.94	0.91	20	30,102,102	*HM13*	Promoter	Island	Wk. Promoter	Y
cg06993862	987	2.62E-7	0.023	-2.25	87.87	86.59	86.36	3	49,141,001	*QARS1*	Body	N_Shore	Str. Enhancer	N
cg13310154	978	4.56E-7	0.029	-0.11	1.44	1.41	1.36	17	55,910,843	-	-	Open Sea	Insulator	Y
cg16601151	959	5.92E-7	0.031	-0.11	1.11	1.09	1.05	16	19,079,341	*COQ7*	Body	Island	Act. Promoter	Y
cg26787244	965	6.96E-7	0.034	-0.55	97.56	97.43	97.30	11	644,011	-	-	Island	Wk. Transcribed	Y
cg07261978	950	9.52E-7	0.038	-2.91	88.50	87.86	86.80	8	82,043,566	-	-	Open Sea	Wk. Enhancer	N
cg19029576	954	1.01E-6	0.040	-1.84	65.01	63.76	63.53	9	140,330,686	*ENTPD8*	Body	Island	Wk. Transcribed	N
cg25629796	971	1.12E-6	0.041	0.35	97.40	97.59	97.66	17	11,940,467	*MAP2K4*	Body	Open Sea	Wk. Transcribed	N
cg01430385	954	1.20E-6	0.042	-1.43	92.02	91.72	91.30	10	1,033,349	*GTPBP4*	Promoter	N_Shore	Act. Promoter	Y
cg02519806	948	1.50E-6	0.045	-0.26	2.04	1.97	1.91	4	119,810,036	*SYNPO2*	Promoter	Open Sea	Psd. Promoter	Y
cg04915788	959	1.53E-6	0.044	-1.04	96.06	95.84	95.43	12	25,156,888	-	-	Open Sea	Wk. Enhancer	Y
cg13170005	957	1.75E-6	0.048	-4.77	16.07	13.99	13.78	17	4,690,632	*VMO1*	Promoter	N_Shore	Repressed	Y
cg26483235	938	2.16E-6	0.050	-1.20	96.33	96.33	95.87	8	62,870,065	-	-	Open Sea	Heterochromatin	Y
cg07318309	913	2.29E-6	0.051	0.19	98.74	98.81	98.85	5	156,537,633	*HAVCR2*	Promoter	Open Sea	Wk. Enhancer	N
cg07509511	913	2.45E-6	0.051	-1.74	94.42	94.28	93.80	2	65,297,852	*CEP68*	Body	Open Sea	Trans. Elong.	N
cg11062848	913	2.57E-6	0.052	2.08	93.19	93.41	93.83	6	13,616,537	*NOL7*	Body	S_Shore	Act. Promoter	Y
cg18722557	917	2.80E-6	0.054	0.23	1.86	1.86	1.94	10	123,734,694	*NSMCE4A*	Promoter	Island	Act. Promoter	Y
cg20531550	923	2.81E-6	0.054	-3.11	79.39	76.46	77.03	7	128,550,902	*KCP*	Promoter	Island	Wk. Transcribed	Y
cg05730283	933	2.96E-6	0.055	0.42	2.18	2.21	2.34	2	62,422,808	*B3GNT2*	Promoter	Island	Act. Promoter	Y
cg20540608	901	2.97E-6	0.053	1.53	6.71	7.12	7.22	5	60,241,003	*NDUFAF2; ERCC8*	Promoter	Island	Act. Promoter	Y
cg25597976	929	3.00E-6	0.055	1.05	93.80	94.02	94.34	8	74,750,772	*UBE2W*	Body	Open Sea	Wk. Enhancer	Y
AMPA														
cg22171277	938	1.25E-7	0.038	0.32	1.00	1.06	1.09	2	70,363,119	-	-	Open Sea	Str. Enhancer	N
cg23261070	946	1.55E- 7	0.041	0.13	1.17	1.21	1.21	17	6,918,037	*C17orf49*	Promoter	Island	Act. Promoter	Y

Frequency = number of subsamples out of 1,000 in which DMP was identified as significant. Chromatin state is predicted by ChromHMM from ENCODE data for the GM12878 cell line. Delta-β: methylation difference per unit glyphosate or AMPA in units of β values; AMPA: aminomethylphosphonic acid; Tert: tertile, Str: strong, Wk: weak, Act: active, Psd: poised; Trans. Elong.: transcription elongation

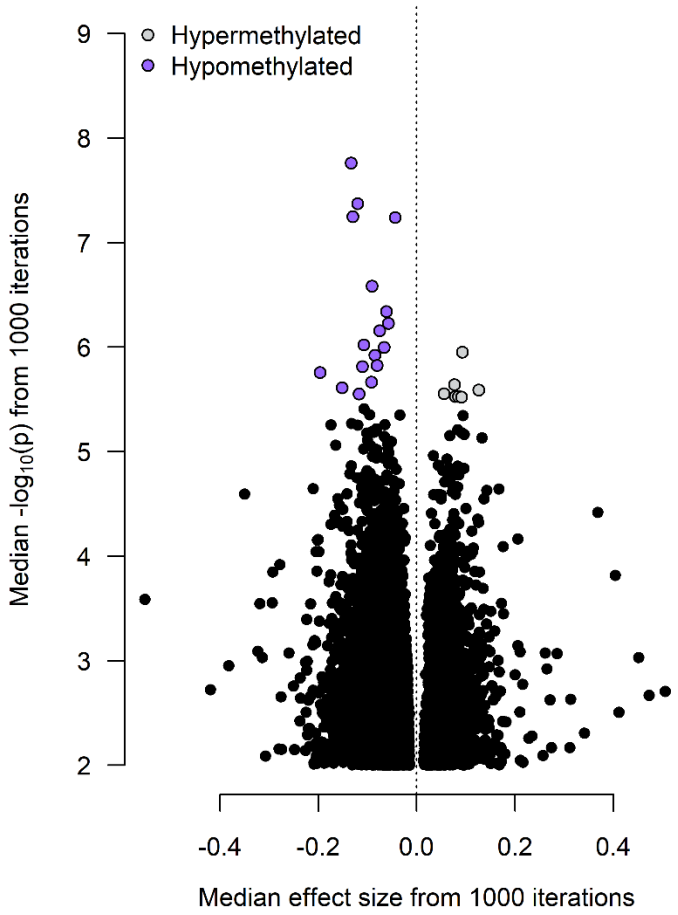

Figure 3.4. Volcano plot of results from probe-level differential methylation analysis for glyphosate

Biological and functional analyses were performed on only glyphosate-associated probes because there were only two AMPA-associated probes (Figure 3.5). Compared to all other probes on the array, a greater proportion of glyphosate-associated DMPs were located within CpG islands (33.3% vs. 19.2%), although this difference was not statistically significant (Fisher p = 0.21). Glyphosate-associated DMPs were significantly enriched for enhancer regions (29.2% vs. 12.6%, Fisher p = 0.02) and depleted for inactive/heterochromatin regions (4.2% vs. 36.8%, Fisher p = 0.0004). Three gene ontology terms were enriched in the 20 annotated genes containing glyphosate-associated probes: DNA metabolic process (biological process), intracellular organelle part (cellular

component), and organelle part (cellular component), all with p<0.01. No KEGG or IPA pathways were enriched in glyphosate-associated genes.

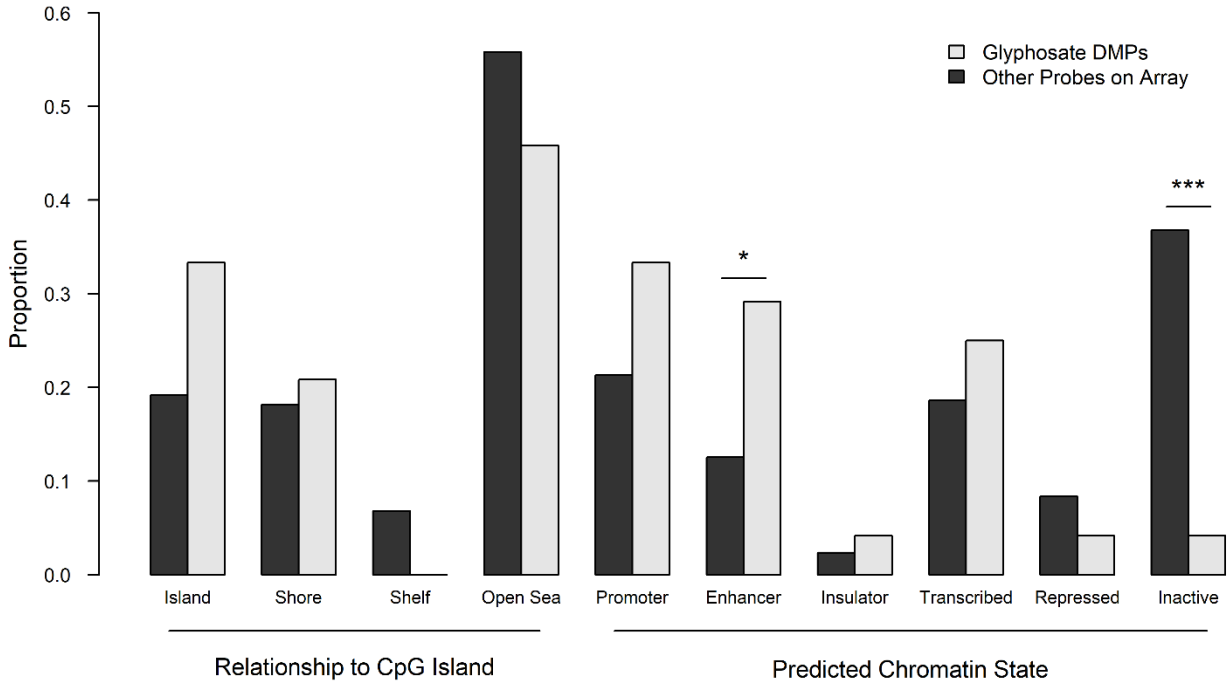

Figure 3.5. Enrichment analysis for genomic context of glyphosate-associated differentially methylated probes (DMPs)
Asterisks denote contexts that are statistically significantly different between DMPs associated with urinary glyphosate levels compared to other probes on the array according to Fisher's exact test (* indicates $p < 0.05$, *** $p < 0.001$).

Four regions for glyphosate and none for AMPA were significant in ≥90% of subsamples and considered differentially methylated (Table 3.4). Three regions were significantly associated with AMPA at a relaxed threshold of >60% of subsamples and are also included in Table 3.4. The glyphosate-associated regions were all located within gene promoters. Three were hypomethylated with greater glyphosate (*MSH4*, *KCNA6*, and *ABAT*) and the other was hypermethylated (*NDUFAF2*/*ERCC8*). The top AMPA-associated regions were all hypomethylated. Two were located within gene bodies (*RNF39*, *TRIM31*) and one within a gene promoter (*ESR1*).

Table 3.4. Differentially methylated regions (DMRs) associated with urinary glyphosate and AMPA concentration identified by a resampling-based method

Chro-mo-some	Region start (hg19)	Region end (hg19)	Width (bases)	Number of CpGs in region	Freq-uency	FDR q	Average delta-β for probes in region	Gene(s)	Location within gene(s)
Glyphosate									
1	76,262,302	76,262,984	682	9	979	0.023	-1.88	*MSH4*	Promoter
5	60,240,926	60,241,324	398	5	918	0.074	0.58	*NDUFAF2; ERCC8*	Promoter
12	4,918,169	4,919,230	1061	10	976	0.011	-5.60	*KCNA6*	Promoter
16	8,806,359	8,806,756	397	9	925	0.034	-4.98	*ABAT*	Promoter
AMPA									
6	30,039,380	30,039,524	144	8	606	0.42	-13.02	*RNF39*	Body
6	30,071,496	30,071,612	116	5	636	0.025	-5.25	*TRIM31*	Body
6	152,126,736	152,126,938	202	5	616	0.030	-3.63	*ESR1*	Promoter

Frequency = number of subsamples out of 1,000 in which DMR was identified as significant. FDR q values were combined for all CpGs in region by Stouffer's method. Delta-β: methylation difference per unit glyphosate or AMPA in units of β values. AMPA: aminomethylphosphonic acid

Methylation index

15 probes were selected by the LASSO operator from the 24 DMPs for inclusion in the final glyphosate methylation index (Appendix A, Table A7). The index was significantly correlated with urinary glyphosate in the training (Pearson's R = 0.63, p <0.0001, Figure 3.6A) and validation (R = 0.35, p = 0.006, Figure 3.6B) sets. In the validation set, the methylation index was significantly associated with glyphosate tertile (ANOVA p = 0.01, Figure 3.6C) and showed excellent discrimination between the top and bottom tertiles of urinary glyphosate (AUC = 0.75, Figure 3.6D). A modified index trained using only probes included on the HumanMethylation450 BeadChip (the previous version of the Illumina methylation array), which selected 11 probes in the final model, achieved similar performance (Appendix B, Figure B2). The AMPA index was not significantly correlated with urinary AMPA in either the training or validation sets, likely due to the use of only two DMPs for prediction (data not shown).

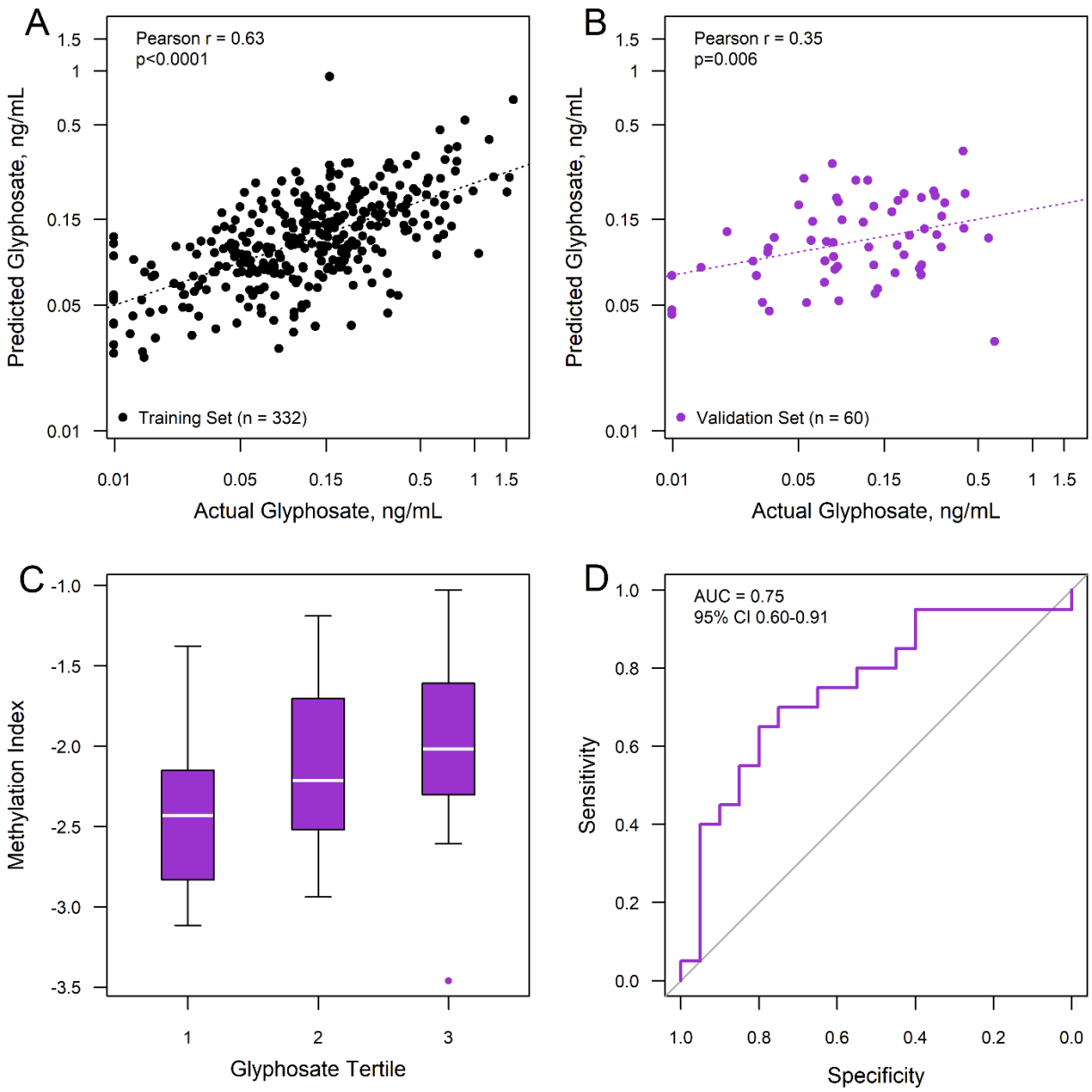

Figure 3.6. Performance of methylation index using 15 CpG sites to predict the natural logarithm of urinary glyphosate concentration

Data shown for the training set (A) and the validation set (B, C). Panel D shows the classification performance of the methylation index in the validation set for classifying the top vs. the bottom tertile of urinary glyphosate.

3.4 Discussion

This is the first study examining associations between urinary glyphosate and AMPA levels and DNA methylation. Higher AMPA, but not glyphosate, was associated with greater

epigenetic age acceleration, a phenomenon which has previously been linked to the risk of many diseases. We identified 24 CpG sites whose methylation was associated with glyphosate and two associated with AMPA. There were four regions associated with glyphosate, within the promoters of *MSH4*, *KCNA6*, *ABAT*, and *NDUFAF2*/*ERCC8*, and an association between *ESR1* promoter hypomethylation and elevated AMPA. Finally, using 15 CpG sites, we developed a methylation index which was significantly associated with glyphosate concentration in an internal validation set.

The significant and replicable differential DNA methylation associated with urinary glyphosate and AMPA informs the hypothesis that these compounds may have biological effects in humans, but the mechanisms by which glyphosate and AMPA could impact human health remain unclear. Although humans and other animals do not possess the shikimate pathway inhibited by glyphosate[197–199], conversion of glyphosate to AMPA has been observed in multiple types of bacteria[155]. Previous studies have shown glyphosate-induced microbiome changes in honeybees[200,201], birds[202], and rats[55–57,174]. Since the microbiome can have significant and wide-ranging impacts on human health[203], it is possible that glyphosate and/or AMPA could influence various aspects of human health via perturbations of the microbiome, changes which could be reflected in differential peripheral blood DNA methylation[204,205].

In our study, AMPA was associated with epigenetic age acceleration, which has been previously associated with other environmental exposures[206–209] and risk of many diseases, including cancer[210–212], obesity and metabolic syndrome[213–215], and all-cause mortality[216–218]. The fact that this relationship was present only for AMPA supports the hypothesis that glyphosate and AMPA have distinct effects on the human body.

Furthermore, the genes whose methylation was associated with glyphosate and AMPA included those involved in various biological pathways related to cancer. *SF3B2* is involved in RNA splicing and DNA repair[219] and associated with various types of cancer[220–222]. A germline mutation in the mismatch repair gene *MSH4*[223] was described in a family with a high incidence of nervous system tumors[224]. *ERCC8* is also involved in DNA repair via transcription-coupled nucleotide excision repair and double-strand break repair[225–227]. Hypomethylation at *TRIM31*, which has been shown to promote progression in a variety of tumor types[228–230], was associated with AMPA. The AMPA-associated hypomethylation at the *ESR1* promoter is fascinating given the potential link between AMPA and breast cancer risk[50,58] and the potential for glyphosate-induced endocrine disruption[173]. Promoter hypomethylation is usually associated with increased gene expression, which was observed for the *ESR1* product ERα in breast cancer cells exposed to glyphosate[58]. However, it should be noted that the relationship between AMPA and *ESR1* promoter hypomethylation was only present after relaxing the original threshold for significance.

This is also the largest report of urinary glyphosate and AMPA levels in the non-agricultural setting in the United States, complementing previous studies in the general population in other countries[69,231,232] and in limited settings in the U.S.[43,44,154,233]. The majority of women (>80%) had detectable glyphosate and AMPA in at least one urine sample, highlighting the near-ubiquitous exposure to these compounds, although the small convenience sample may not be representative of the general population. The levels observed among participants in our study (median of 0.12 ng/mL for glyphosate and 0.06 ng/mL for AMPA) were slightly lower than those previously reported, likely due to differences in the study population and detection assays. Our LC-MS/MS assay had a

considerably higher analytic sensitivity (LOD/LOQ 0.014/0.041 ng/mL for glyphosate and 0.013/0.040 ng/mL for AMPA) than most assays previously described[141,143].

This study utilizes the most recent methylation array chip and best practices for methylation array data filtering, normalization, and analysis alongside a resampling-based approach intended to improve the stability and reproducibility of the results[190–193]. We previously demonstrated that this method is capable of identifying differentially methylated sites associated with smoking with a minimal number of false positives compared to a traditional epigenome-wide association approach[188].

However, the lack of an external validation cohort is a limitation of the study. To our knowledge, no other studies have measured both urinary glyphosate and AMPA and DNA methylation, so an external dataset is not currently available. The successful performance of the methylation index for glyphosate concentration in the internal validation set suggests that the glyphosate-associated differential methylation may be replicated in other cohorts with a similar participant profile. Our study cohort consisted of postmenopausal women aged 45-66, so results may not be generalizable to other populations or those residing outside of California. A further limitation of our study is the lack of gene expression data. Without this, we cannot state whether the differential DNA methylation identified in our study translates to differences in gene expression which could confer functional impacts in the body, nor can we infer the temporal relationship between DNA methylation differences and glyphosate exposure given the cross-sectional nature of the study.

DNA methylation is highly dependent on tissue type. Given the ambiguity of glyphosate's impact on health and the systemic nature of exposure, blood likely represents the best available tissue at this time, especially considering its availability for epidemiologic

study. A major concern regarding the use of blood in epigenome-wide association studies is its heterogeneous cell composition, which may be impacted by disease states or exposure to pro-inflammatory compounds. However, we saw no evidence of differences in white blood cell composition by glyphosate or AMPA levels, and all analyses were adjusted for these proportions.

In humans, glyphosate and AMPA have short half-lives of approximately 5-10 hours[59,60], which makes accurate assessment of long-term exposure challenging. Our choice to use two spot urine samples from a 10-day period is intended to balance the need for a more complete picture of typical exposure with feasibility. The moderate between-sample ICCs of 0.53 for glyphosate and 0.34 for AMPA in our study participants highlight the pitfalls of relying on a single urine sample and suggest that more samples per individual may be needed to provide a better estimate of long-term exposure. However, because the dates of spot urine samples were not known in advance to the study participants, the glyphosate and AMPA measurements should reflect their usual habits.

This study identified differential DNA methylation associated with the herbicide glyphosate and its metabolite AMPA and developed a methylation index that accurately predicted urinary glyphosate concentration and tertile in an internal validation sample. Glyphosate- and AMPA-associated methylation occurred near genes associated with cancer (*SF3B2, MSH4*, and *TRIM31*) and endocrine disruption (*ESR1*), and AMPA was associated with greater epigenetic age acceleration. These results suggest that exposure to these common chemicals affects the epigenome, informing the hypothesis that glyphosate and/or AMPA exposure could elevate the risk for disease, including cancer. Further studies are warranted to replicate our results, determine the functional impact of glyphosate- and

AMPA-associated differential DNA methylation, and explore whether DNA methylation could serve as a biomarker of long-term glyphosate exposure.

Chapter 4: Association of Mammographic Density with Blood DNA Methylation

4.1 Introduction

Breast cancer is the most common cancer among women in the United States, with 266,120 new cases estimated in 2018[234]. The extent of radiologically dense fibroglandular tissue appearing on a mammogram, known as mammographic density, varies between women and is largely heritable[71,72], but also influenced by an individual's lifestyle[73–75] and exposure to exogenous hormones or other drugs[76,77]. Increased mammographic density is a strong risk factor for breast cancer, with women in the highest density category exhibiting more than a 4-fold increase in breast cancer risk compared to women in the lowest density category[78–80]. A number of studies have shown that mammographic density mediates the associations between breast cancer risk factors such as history of breast biopsy, nulliparity, age at first birth, and hormone therapy, and breast cancer risk[81–83]. However, the molecular mechanisms underlying mammographic density and how density influences cancer risk are largely unknown.

Epigenetic changes, such as DNA methylation, may serve as an intermediate phenotype for risk factors and their associated diseases[176,177,235]. As such, DNA methylation may provide a powerful tool for understanding the etiology of breast cancer. It is well-established that tumor tissue frequently exhibits changes in DNA methylation compared to normal tissue, such as global hypomethylation and hypermethylation of the promoters of tumor suppressor genes[1–4,6]. A developing avenue of research involves examining DNA methylation in surrogate tissues, such as blood. A number of studies have shown that

epigenome-wide DNA methylation in white blood cells, especially at non-promoter regions, is associated with decreased breast cancer risk (reviewed in [10]).

Epigenetic signatures associated with specific cancer risk factors may provide a better understanding of how these risk factors promote the development of breast cancer. White blood cell DNA methylation has provided epigenetic signatures for risk factors such as aging[17,18], lifetime estrogen exposure[235], body mass index (BMI)[20,236], alcohol[33,237], and smoking[14,15]. The goal of this study was to identify DNA methylation loci associated with mammographic density in postmenopausal women. Discovering and validating such loci may advance our understanding of the biological mechanism linking mammographic density and breast cancer risk.

4.2 Methods

Study recruitment

Study participants were postmenopausal women between the ages of 45-66 with no personal history of breast cancer or mastectomy and for whom a mammogram report indicating mammographic density was available within the last 3 years (N = 385). Recruitment, questionnaires, and specimen collection and processing were described previously[144]. Women were considered postmenopausal if they had not had a menstrual cycle for 1 year or more. Participants answered questionnaires about their breast health history and provided a blood sample for DNA methylation analysis. The study was approved by the University of California, Irvine (UCI) Institutional Review Board, HS #2016-3127.

Assessment of mammographic density and covariates

Each participant completed questionnaires which provided age at menarche, age at menopause, parity, age at first birth, use of hormone replacement therapy (HRT), height,

weight, smoking status, and alcohol intake. The number of years since menopause was calculated by subtracting the age at menopause from the age at blood draw. HRT use was categorized as current user, former user, and never user. BMI was calculated from self-reported height and weight by dividing the weight in kilograms by the height in meters squared. Participants who smoked cigarettes regularly at the time of blood draw were categorized as current smokers, those who had smoked regularly for 6 months or more but did not currently smoke were categorized as former smokers, and all others were categorized as never smokers. Alcohol use was assessed with two questions: the frequency of alcohol consumption (never, once per month or less, 2-4 times per month, 2-3 times per week, and 4 times or more per week) and the number of drinks per occasion (1-2, 3-4, 5 or more). The responses were categorized as zero, 1 or fewer, 2-6, and 7 or more drinks per week.

Mammographic breast density was obtained from participants' most recent mammogram report. Mammographic density was classified according to the Breast Imaging Reporting and Data System (BI-RADS), with possible categories: A, almost entirely fatty; B, scattered fibroglandular densities; C, heterogeneously dense; and D, extremely dense[238]. For all data analyses, density was treated as an ordinal variable, with category A being the lowest and category D being the highest.

DNA extraction and methylation profiling

Peripheral blood samples were obtained from each study participant and stored at 4°C until processing. Within 6 hours of collection, samples were centrifuged at 2000 rpm for 10 minutes, and DNA was extracted from the buffy coat using the QIAamp DNA Blood Maxi Kit (QIAGEN, Hilden, Germany). Extracted DNA was quantitated by Synergy HT microplate

reader (BioTek, Winooski, VT, USA) using the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) and stored at -80°C.

1.5 µg of DNA was sent to the University of Southern California Molecular Genomics Core in three batches for methylation profiling. Genomic DNA was bisulfite converted using the Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA, USA) and then DNA methylation at over 850,000 CpG sites was measured using the Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA, USA). Laboratory staff were blinded to the mammographic density status of the samples. All samples passed internal controls and were included in the final data analysis.

Methylation data pre-processing

Figure 4.1 summarizes the data pre-processing and analysis pipeline. Methylation array data were pre-processed according to recommended steps for Illumina methylation BeadChip data[128]. All data processing and analysis was performed in R, version 3.5.1[149].

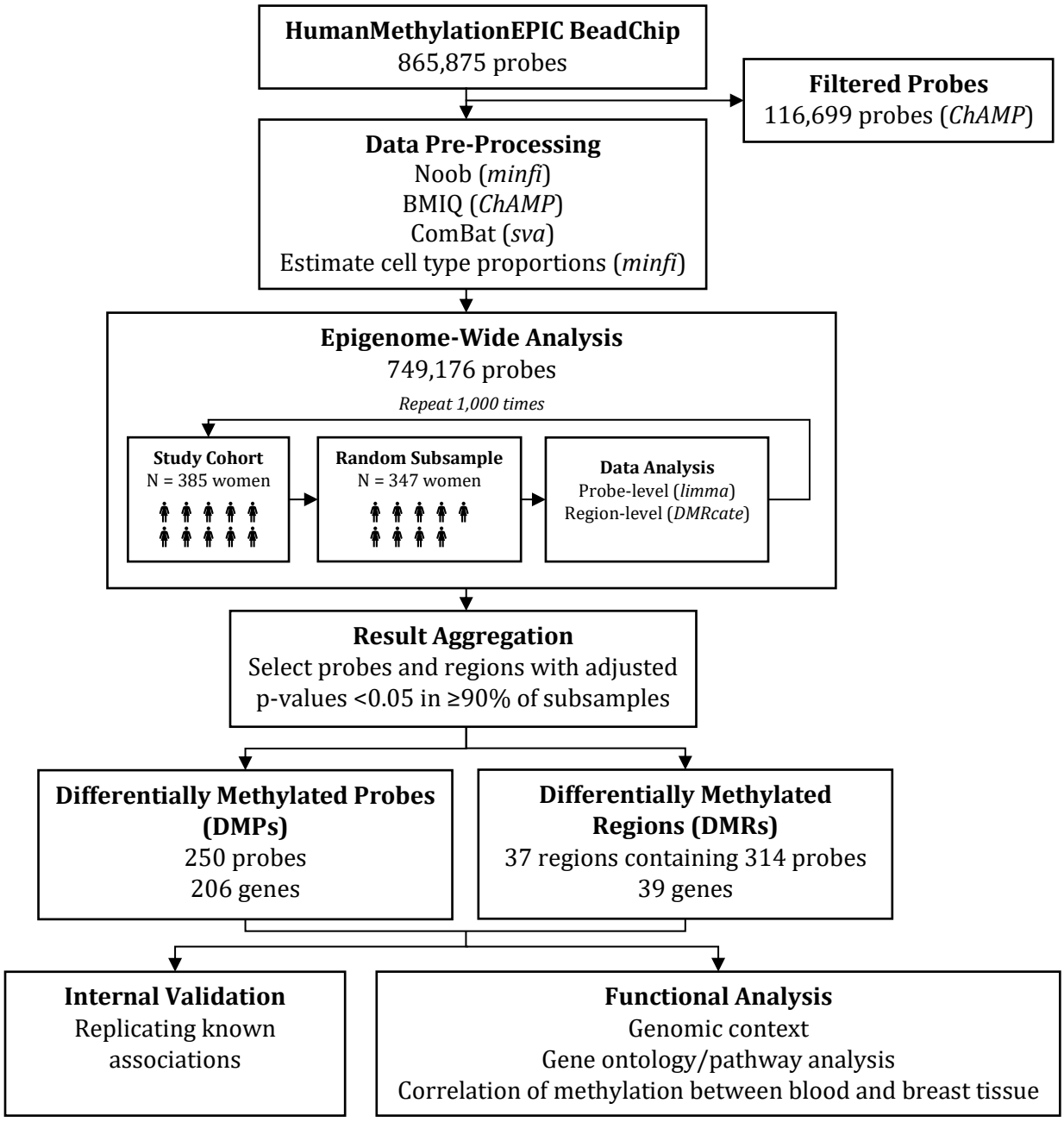

Figure 4.1. DNA methylation data pre-processing and analysis pipeline

First, the raw intensities were background corrected and adjusted for dye bias using the noob normalization procedure within the minfi package, version 1.28.3[181]. Probes with a detection p-value >0.05 in a sample were considered missing for that sample. Low-quality probes were filtered and removed from further analysis for all samples, including probes

that: 1) were missing in at least 20% of samples (n = 648); 2) had SNPs with global minor allele frequency >1% within 5 base pairs of the target sequence or mapping problems with the probe sequence (n = 99,109)[179]; 3) hybridize to multiple locations (n = 15)[180]; or 4) were located on the sex chromosomes (n = 16,927). Beta mixture quantile normalization (BMIQ) was applied to correct for type II probe bias[182]. Previous empirical data suggests that the combined noob and BMIQ methods are the most effective for methylation array normalization[239]. Except for noob normalization, filtering and normalization were completed using the ChAMP package, version 2.10.2[183]. To address batch effects, the ComBat procedure was applied using sva, version 3.30.1, adjusting for batch and position on chip[184,185].

Finally, the proportions of white blood cell types (monocytes, CD8T, CD4T, B cells, natural killer cells, and neutrophils) were estimated with minfi, which implements the reference-based method described by Houseman[186] (Appendix A, Table A8). Methylation was quantified using β values, representing the percent methylation at the site, and M values, the logit transformation of the β value. Because of its more desirable statistical properties, the M value was used for differential methylation analysis[187], but β values were used for reporting of results to simplify interpretation.

Genome-wide average methylation and epigenetic age acceleration

First, genome-wide DNA methylation was estimated by averaging the methylation levels (β values) of all probes on the array. Linear regression was used to examine associations between genome-wide average methylation and mammographic density. The model was adjusted for age, race/ethnicity, BMI, HRT use, parity, time since menopause, alcohol use, smoking status, batch, position on chip, and estimated white blood cell

proportions. These covariates were selected a priori based on known relationships with either DNA methylation, mammographic density, or both. Analyses were repeated for genome-wide average methylation stratified by genomic context, described below.

Epigenetic age was calculated using Horvath's epigenetic clock[17]. Epigenetic age acceleration, or the difference between chronological age and epigenetic age, was compared for women with varying mammographic densities. Linear regression was used to determine whether there was a statistically significant relationship between epigenetic age acceleration and mammographic density.

Probe- and region-level analyses

For probe-level analyses, linear models with the M value at each site as the dependent variable and mammographic density as the predictor were implemented using the limma package, version 3.38.3[145]. Models were adjusted for the same covariates as the genome-wide average methylation analysis above. The threshold for genome-wide significance was set at a false discovery rate (FDR) of $q < 0.05$.

Regional analysis was conducted using DMRcate[189], version 1.18.0, using the same model as for the probe-level analysis. This approach identifies regions of differential methylation in adjacent probes using a kernel smoothing method. We used a threshold of FDR $q < 0.05$ for genome-wide significance and the package developer-suggested values of 1000 bases for λ and 2 for the scaling factor C. To rank the relative significance of the regions, q-values for the probes contained in the region were combined using Stouffer's method.

For both probe- and region-level analysis, a resampling-based approach was applied in order to improve the reproducibility and stability of the list of differentially methylated loci (Figure 4.1). Because of the high dimensionality of microarray data, the results from a

single experiment are highly sensitive to small differences in the sample[190–193]. By aggregating results from a large number of subsamples of the main study sample, we can identify methylation differences that are stably associated with our outcome of interest. This approach has only recently been applied to DNA methylation microarray data[240,241]. Further information about the development and selection of parameters for this method can be found in Appendix C.

The study cohort was randomly sampled without replacement 1000 times to select subsamples consisting of 90% of the study participants. Each subsample was used to perform probe- and region-level analysis as described above using the UCI High Performance Computing Cluster. Results were aggregated to identify probes and regions that were consistently associated with mammographic density. Probes that were called as differentially methylated in ≥90% of subsamples were selected as differentially methylated probes (DMPs). After combining overlapping regions, regions that were associated with mammographic density in ≥90% of subsamples were selected as differentially methylated regions (DMRs).

Internal validation

To demonstrate the validity of our approach, we used the same method to identify DMPs associated with smoking status in our study cohort. We then compared our list of DMPs to results from previous epigenome-wide association studies of smoking.

Biological and functional analyses

DMPs were annotated using the Illumina manifest (version B4) to identify associated genes and genomic context. Probes were also mapped to ChromHMM data from ENCODE[242] to determine predicted chromatin state; data from the GM12878 lymphoblastic cell line was

used as it is the tissue most similar to leukocytes. Chromatin states from the 15-state ChromHMM model were pooled into six categories as followed: promoter (active, weak, or poised promoters), enhancer (strong or weak enhancers), transcribed (transcriptional transition, transcriptional elongation, or weak transcribed), repressed, inactive (heterochromatin and repetitive regions), and insulators. DMPs were examined for enrichment of certain genomic locations, location relative to CpG islands, or chromatin states, compared to the background of all non-DMP probes included on the array. Proportions were compared using a chi-squared test. We used DAVID[243] to examine enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and gene ontology terms in genes associated with DMPs, and assessed enrichment of pathways using Ingenuity Pathway Analysis software, version 57662101 (Release Date: 2020-09-15).

Correlation with breast tissue

To determine the correlation of methylation at DMPs and DMRs between blood cells and breast tissue, we used DNA methylation data from paired samples from the Susan G. Komen Tissue Bank[244]. Details of the study cohort and specimen processing were described previously[245]. Briefly, 40 women provided blood and breast tissue specimens at two time points spaced at least a year apart, for a total of 160 specimens. DNA methylation for these specimens was characterized using the Illumina Infinium HumanMethylation450 BeadChip array[245]. We calculated Pearson correlations to determine the relationship between methylation in paired blood and breast tissue at each time point at each DMP and all CpG sites located within DMRs.

4.3 Results

Cohort characteristics

Cohort characteristics are presented in Table 4.1. Race/ethnicity was associated with density, with a higher proportion of Asian women in the heterogeneously dense or extremely dense category. BMI was inversely associated with density ($p < 0.001$), consistent with previous literature[74]. A higher proportion of current HRT users had extremely dense breasts, also consistent with previous literature[76]. Other variables of interest, including age, smoking status, alcohol use, and reproductive history, were not associated with mammographic density.

Genome-wide average methylation

Genome-wide average DNA methylation was calculated by taking the average of all probes on the array after filtering out low-quality probes and probes on the sex chromosomes (mean 62.15%, range 61.10-63.25%). Average methylation was higher in women with higher mammographic density (Figure 4.2), increasing 0.034% per density category (95% CI 0.001-0.067, $p = 0.04$) (Table 4.2). The adjusted R^2 of 0.40 suggests that the variables included in the model explain a moderate amount of the overall variance in genome-wide average DNA methylation. To further characterize this relationship, analyses were repeated after stratifying probes based on their relationship to CpG islands and chromatin state (Table 4.2). The association between mammographic density and genome-wide DNA methylation was present in probes located in CpG islands and shores (Figure 4.2) but not those in CpG shelves or distal to CpG islands. Mammographic density was also associated with hypermethylation at predicted promoters, enhancers, transcribed regions,

repressed regions and insulators (Figure 4.2), but not in inactive/heterochromatin regions, gene bodies, or intergenic regions.

Table 4.1. Cohort characteristics by BI-RADS mammographic density category

	BI-RADS Mammographic Density					
	A: Almost entirely fatty	B: Scattered fibroglandular densities	C: Heterogeneously dense	D: Extremely dense	Total	p
Total	42 (10.9)	107 (27.8)	160 (41.6)	76 (19.7)	385	
Age	58.0 (54.0, 60.0)	56.0 (53.5, 60.0)	58.0 (55.0, 61.0)	56.0 (54.0, 60.0)	385	0.25
Race/ethnicity						
White	31 (77.5)	70 (65.4)	99 (63.5)	50 (65.8)	250	
Asian	1 (2.5)	7 (6.5)	19 (12.2)	15 (19.7)	42	**0.04**
Hispanic	5 (12.5)	22 (20.6)	31 (19.9)	10 (13.2)	68	
Other	3 (7.5)	8 (7.5)	7 (4.5)	1 (1.3)	19	
BMI	33.6 (27.3, 38.8)	27.4 (24.4, 31.8)	24.4 (22.1, 28.2)	22.1 (20.3, 24.6)	385	**<0.001**
Smoking status						
Never	26 (63.4)	76 (71.0)	115 (71.9)	63 (82.9)	280	0.13
Former	13 (31.7)	24 (22.4)	37 (23.1)	13 (17.1)	87	
Current	2 (4.9)	7 (6.5)	8 (5.0)	0 (0.0)	17	
Alcohol use (drinks per week)						
None	13 (31.7)	26 (24.3)	45 (28.1)	21 (27.6)	105	
1 or fewer	18 (43.9)	47 (43.9)	66 (41.3)	27 (35.5)	158	0.81
2-6	8 (19.5)	19 (17.8)	26 (16.3)	17 (22.4)	70	
7 or more	2 (4.9)	15 (14.0)	23 (14.4)	11 (14.5)	51	
Age at menarche	12.0 (12.0, 13.0)	13.0 (12.0, 13.0)	13.0 (12.0, 14.0)	13.0 (12.0, 14.0)	384	0.54
Parity						
0	7 (16.7)	18 (16.8)	37 (23.1)	22 (28.9)	84	0.21
≥1	35 (83.3)	89 (83.2)	123 (76.9)	54 (71.0)	301	
Age at first birth (parous women)	28.0 (21.0, 31.0)	28.0 (23.0, 32.0)	27.0 (23.0, 31.0)	29.0 (27.0, 31.0)	301	0.18
Age at menopause	48.0 (42.3, 52.0)	50.0 (47.0, 53.0)	50.0 (46.0, 53.0)	50.5 (48.0, 52.3)	382	0.11
HRT use						
Never	29 (69.0)	78 (72.9)	99 (61.9)	37 (48.7)	243	**0.003**
Former	5 (11.9)	17 (15.9)	28 (17.5)	10 (13.2)	60	
Current	8 (19.0)	12 (11.2)	33 (20.6)	29 (38.2)	82	

Values are frequency (percent) for categorical variables and median (interquartile range) for continuous variables. P-values are from Fisher's exact test (categorical variables) and Kruskal-Wallis test (continuous variables). Missing data: 6 for race/ethnicity, 1 for smoking status, 1 for alcohol use, 1 for age at menarche, 3 for age at menopause. BMI: body mass index; HRT: hormone replacement therapy

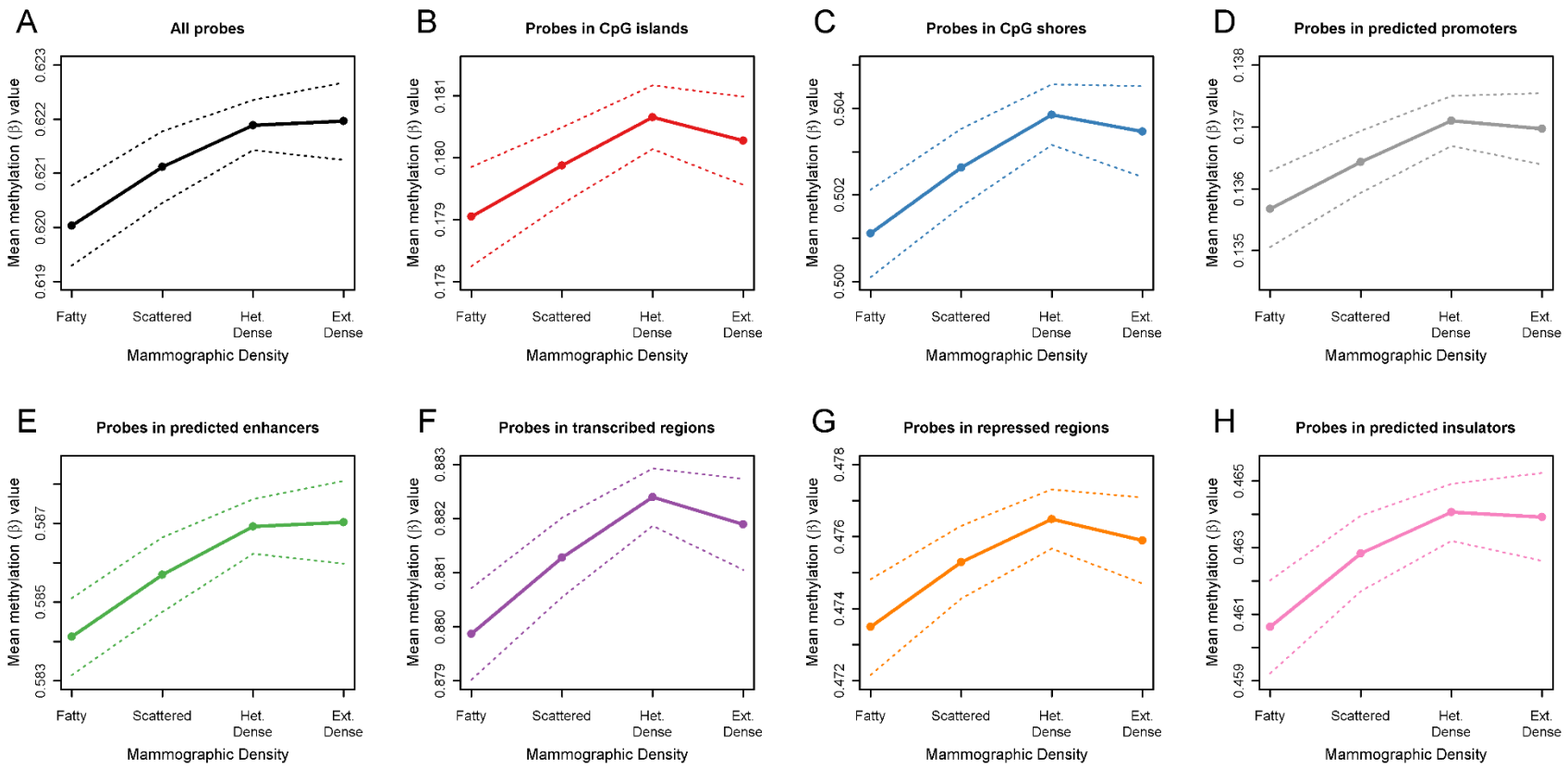

Figure 4.2. Relationship between mammographic density and genome-wide average methylation
Shown for all probes on HumanMethylationEPIC BeadChip (A), for probes in CpG Islands (B) and CpG shores (C), and for probes by chromatin state predicted by ChromHMM from ENCODE data for the GM12878 cell line: promoters (D), enhancers (E), transcribed (F), repressed (G) and insulators (H). Dashed lines show 95% confidence intervals.

Table 4.2. Association of mammographic density with genome-wide average methylation, stratified by genomic context

	Estimate	95% CI	p	Adjusted R^2
All Probes	0.034	(0.001, 0.067)	**0.04**	0.40
CpG Island				
Island	0.050	(0.021, 0.080)	**<0.001**	0.54
Shore	0.054	(0.009, 0.099)	**0.02**	0.45
Shelf	0.021	(-0.015, 0.057)	0.26	0.40
None	0.023	(-0.017, 0.064)	0.26	0.42
Relationship to Gene				
Gene Body	0.031	(-0.001, 0.063)	0.06	0.43
Intergenic	0.030	(-0.015, 0.074)	0.19	0.35
Chromatin State				
Promoter	0.040	(0.016, 0.065)	**0.001**	0.49
Enhancer	0.055	(0.012, 0.098)	**0.01**	0.51
Transcribed	0.033	(0.003, 0.063)	**0.03**	0.62
Repressed	0.069	(0.015, 0.124)	**0.01**	0.41
Inactive	0.013	(-0.033, 0.060)	0.58	0.39
Insulator	0.073	(0.011, 0.135)	**0.02**	0.34

Model adjusted for age, race/ethnicity, BMI, HRT use, parity, time since menopause, alcohol use, smoking status, batch, position on chip, and cell type proportions. Chromatin states are predicted by ChromHMM from ENCODE data for the GM12878 cell line. BMI: body mass index; HRT: hormone replacement therapy

Epigenetic age acceleration

Chronologic age was moderately correlated with epigenetic age in the study cohort (r = 0.64, p <0.001, Figure 4.3). The mean epigenetic age acceleration was -2.9 years (95% CI -3.3, -2.5), indicating a trend of epigenetic age deceleration. There was no association between mammographic density and epigenetic age acceleration, both in the crude model (p = 0.77) and when adjusted for age, race/ethnicity, BMI, HRT use, parity, time since menopause, alcohol use, smoking status, batch, position on chip, and estimated white blood cell proportions (p = 0.71).

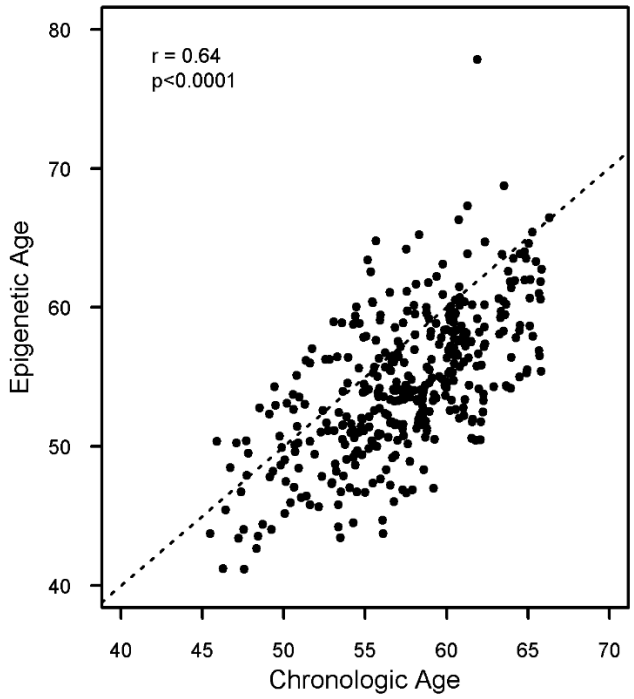

Figure 4.3. Epigenetic age acceleration among study participants

Internal validation of resampling methodology

We sought to test the validity of our resampling-based methodology by identifying probes associated with cigarette smoking in our dataset. Since relationships between DNA methylation at specific sites and smoking exposure have been well-characterized across many studies[14,15,246–249], this exposure serves as a useful positive control.

We identified 40 probes associated with smoking status (current/former/never) in ≥90% of subsamples (Table 4.3). The majority have been previously associated with smoking, either for the specific probe itself (N = 21, 53%) or for other probes within the same gene (N = 14, 35%). Only 5 (13%) of our smoking-associated probes had never been previously associated with smoking. All but one of these probes were exclusive to the Illumina EPIC array, which was used in just one of the previous studies of smoking and DNA methylation[248].

Table 4.3. CpG sites significantly associated with smoking status in ≥90% of subsamples

Probe	Frequency	Median p	Median q	Delta-β, current vs. never	Mean methylation (β): Never smokers	Mean methylation (β): Former smokers	Mean methylation (β): Current smokers	Chromosome	Position (hg19)	Gene	On 450k array	Previously associated with smoking	Reference(s)
cg12876356	906	6.8E-7	0.011	-4.52	89.3	88.0	82.5	1	92,946,825	*GFI1*	Y	Y (this site)	14,246,247,249
cg18316974	923	4.6E-7	9.5E-3	-2.66	96.1	95.4	92.1	1	92,947,035	*GFI1*	Y	Y (this site)	14,246,247,249
cg04535902	987	5.5E-8	1.9E-3	-3.94	90.4	88.9	84.8	1	92,947,332	*GFI1*	Y	Y (this site)	14,246,247,249
cg22976567	934	6.3E-7	0.012	-2.39	21.3	19.9	18.5	1	156,074,182	*LMNA*	N	Y (this gene)	231,232
cg03329539	918	7.1E-7	0.012	-2.98	39.1	37.9	35.3	2	233,283,329		Y	Y (this site)	246,247,249
cg06644428	970	3.6E-7	7.9E-3	-3.77	10.0	7.9	5.8	2	233,284,112		Y	Y (this site)	14,246,247,249
cg21566642	1000	2.6E-19	9.8E-14	-9.88	60.2	56.4	48.3	2	233,284,661		Y	Y (this site)	14,246–249
cg01940273	1000	1.2E-15	3.0E-10	-6.45	66.5	64.1	58.4	2	233,284,934		Y	Y (this site)	14,246–249
cg05435212	1000	3.8E-9	2.1E-4	-0.67	4.0	3.6	3.3	3	27,632,180		Y	N	
cg18754985	982	1.4E-7	3.8E-3	-0.67	97.7	97.5	97.0	3	98,237,750	*CLDND1*	Y	Y (this site)	246–248
cg19859270	1000	2.1E-11	2.3E-6	-0.94	97.6	97.3	96.1	3	98,251,294	*GPR15*	Y	Y (this site)	246–249
cg04180924	999	6.3E-9	3.1E-4	-0.45	98.6	98.4	97.9	3	98,272,064		N	Y (this site)	248
cg02978227	1000	1.9E-8	7.8E-4	-1.33	96.1	95.5	94.6	3	98,292,027		N	Y (this site)	248
cg08712461	920	5.6E-7	0.010	-0.87	97.1	96.8	96.1	3	152,075,029	*MBNL1*	N	Y (this gene)	247
cg03358636	997	3.3E-8	1.2E-3	-4.18	78.8	76.8	74.0	3	197,474,006	*KIAA0226*	Y	Y (this site)	247,249
cg05575921	1000	1.2E-25	8.7E-20	-8.74	86.7	84.1	74.1	5	373,378	*AHRR*	Y	Y (this site)	14,246–249
cg26703534	999	9.5E-10	6.3E-5	-3.07	75.8	74.8	70.9	5	377,358	*AHRR*	Y	Y (this site)	14,246–250
cg21161138	1000	3.3E-10	2.5E-5	-3.60	77.0	76.2	71.5	5	399,360	*AHRR*	Y	Y (this site)	9,14,246–249
cg14466441	1000	3.1E-10	2.4E-5	-1.47	95.9	95.3	94.3	6	11,392,193		N	N	
cg14753356	951	6.2E-7	0.011	-3.93	46.4	43.9	41.1	6	30,720,108		Y	Y (this site)	246,247,249
cg24859433	940	3.9E-7	8.0E-3	-1.76	93.5	92.7	90.8	6	30,720,203		Y	Y (this site)	246,247,249
cg00185804	967	4.7E-7	9.1E-3	-5.71	58.5	55.4	52.0	8	22,907,738	*TNFRSF10B*	N	Y (this gene)	247
cg01494348	981	1.1E-7	3.3E-3	9.31	9.5	12.1	21.0	8	144,660,395	*NAPRT1*	Y	Y (this gene)	246

Table continues

Table 4.3. CpG sites significantly associated with smoking status in ≥90% of subsamples (continued)

Probe	Frequency	Median p	Median q	Delta-β, current vs. never	Mean methylation (β): Never smokers	Mean methylation (β): Former smokers	Mean methylation (β): Current smokers	Chromosome	Position (hg19)	Gene	On 450k array	Previously associated with smoking	Reference(s)
cg00713939	920	5.6E-7	0.010	4.56	5.0	6.2	11.1	8	144,660,590	*NAPRT1*	Y	Y (this gene)	246
cg19357499	954	3.2E-7	7.1E-3	8.07	5.0	6.8	15.8	8	144,660,631	*NAPRT1*	Y	Y (this gene)	246
cg26768182	908	1.1E-6	0.016	-2.36	89.0	88.1	86.0	9	134,272,679		N	N	
cg00798781	1000	2.4E-8	9.6E-4	-0.64	97.2	96.9	96.7	10	32,661,244	*EPC1*	N	N	
cg26190483	901	1.2E-6	0.016	-0.24	98.9	98.8	98.7	11	2,689,695	*KCNQ1*	Y	Y (this gene)	247
cg14391737	1000	1.5E-12	2.4E-7	-6.62	41.9	38.5	35.5	11	86,513,429	*PRSS23*	N	Y (this gene)	246–249
cg00475490	1000	1.3E-10	1.1E-5	-3.68	10.3	8.6	6.6	11	86,517,110	*PRSS23*	N	Y (this gene)	246–249
cg10413305	958	5.0E-7	9.9E-3	-2.57	35.3	33.8	33.1	11	122,540,748	*UBASH3B*	N	Y (this gene)	247
cg20765665	991	1.5E-7	4.0E-3	-4.88	56.4	53.2	52.5	13	72,375,892	*DACH1*	N	N	
cg25845814	972	3.1E-7	7.1E-3	-1.80	90.8	90.3	88.5	14	74,224,613	*MIDEAS*	N	Y (this gene)	247
cg18110140	973	2.2E-7	5.3E-3	-4.64	53.2	50.4	48.1	15	75,350,380		N	Y (this site)	248
cg19572487	972	2.0E-7	4.9E-3	-4.05	55.9	54.3	50.8	17	38,476,024	*RARA*	Y	Y (this site)	246–249
cg17739917	1000	1.1E-12	1.8E-7	-5.85	43.4	41.4	36.5	17	38,477,572	*RARA*	N	Y (this gene)	246–249
cg21698310	998	1.0E-8	4.8E-4	-2.17	12.5	11.2	10.4	17	48,224,718	*PPP1R9B*	Y	Y (this site)	247,249
cg21911711	1000	1.3E-10	1.1E-5	-3.57	85.8	84.4	81.5	19	16,998,668	*F2RL3*	N	Y (this gene)	14,15,246–249
cg03636183	1000	2.5E-11	2.7E-6	-4.92	70.8	69.1	63.8	19	17,000,585	*F2RL3*	Y	Y (this site)	14,15,246–249
cg05086879	970	1.0E-7	3.1E-3	-3.16	88.1	87.6	82.5	22	39,861,490	*MGAT3*	N	Y (this gene)	246,247

Frequency: number of subsamples out of 1,000 in which DMP was identified as significant. Delta-β: methylation difference between current and former smokers in units of β values

Probe-specific analyses

We identified 250 DMPs significantly associated with mammographic density in ≥90% of subsamples after adjustment for age, race/ethnicity, BMI, HRT use, parity, time since menopause, alcohol use, smoking status, batch, position on chip, and estimated cell type proportions (Table 4.4). The majority (N = 192, 77%) were hypermethylated in women with higher mammographic density (Figure 4.4). Probes with the smallest median p-values were located within the *HDLBP*, *TGFB2*, *CCT4*, *PAX8*/*PAX8-AS1*, and *TACC2* genes. The largest effect sizes (delta-β) were observed in probes within the *HLA-DRB1*, *HLA-DRB5*, *PCDHA1*/*PCDHA2*/*PCDHA3*, *TMEM176A*/*TMEM176B*, and *PDGFD* genes. Boxplots for the top four intragenic DMPs by p-value are presented in Figure 4.5 and for the top four intragenic DMPs by delta-β in Figure 4.6.

Table 4.4. CpG sites significantly associated with increasing mammographic density in ≥90% of subsamples

Large table is attached as a supplemental file

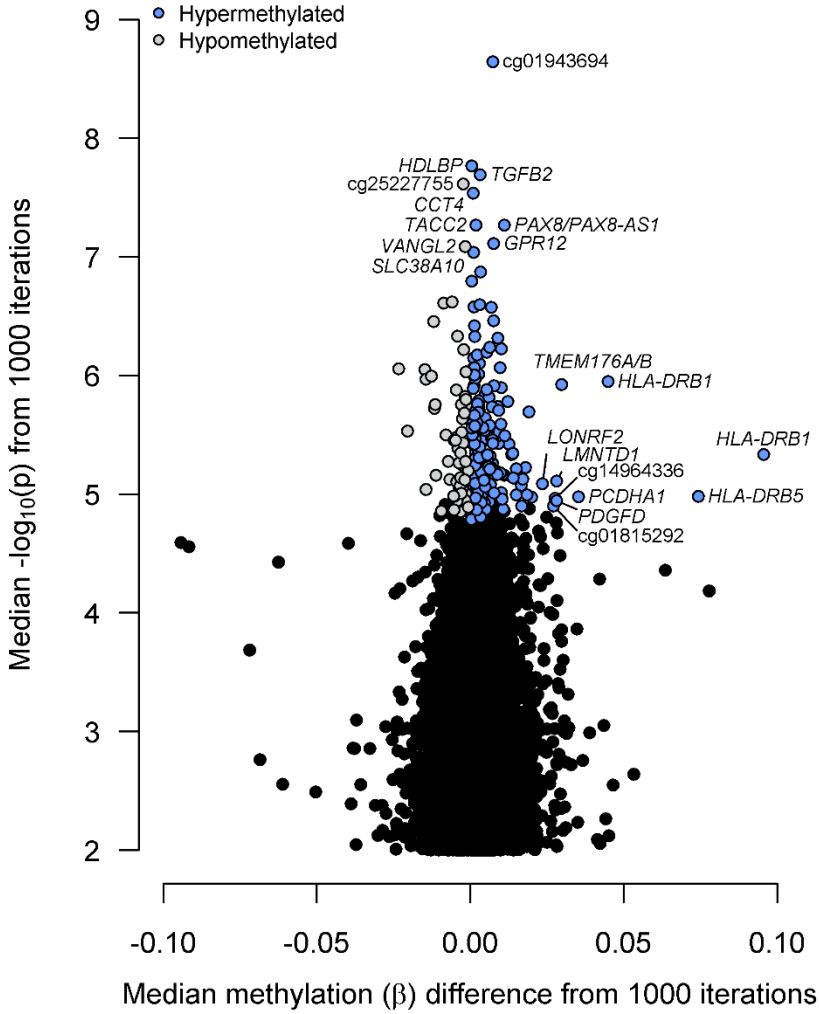

Figure 4.4. Volcano plot of probes associated with mammographic density
Top 10 probes by methylation difference (delta-β) and p-value are labeled with the gene or Illumina identifier (if intergenic).

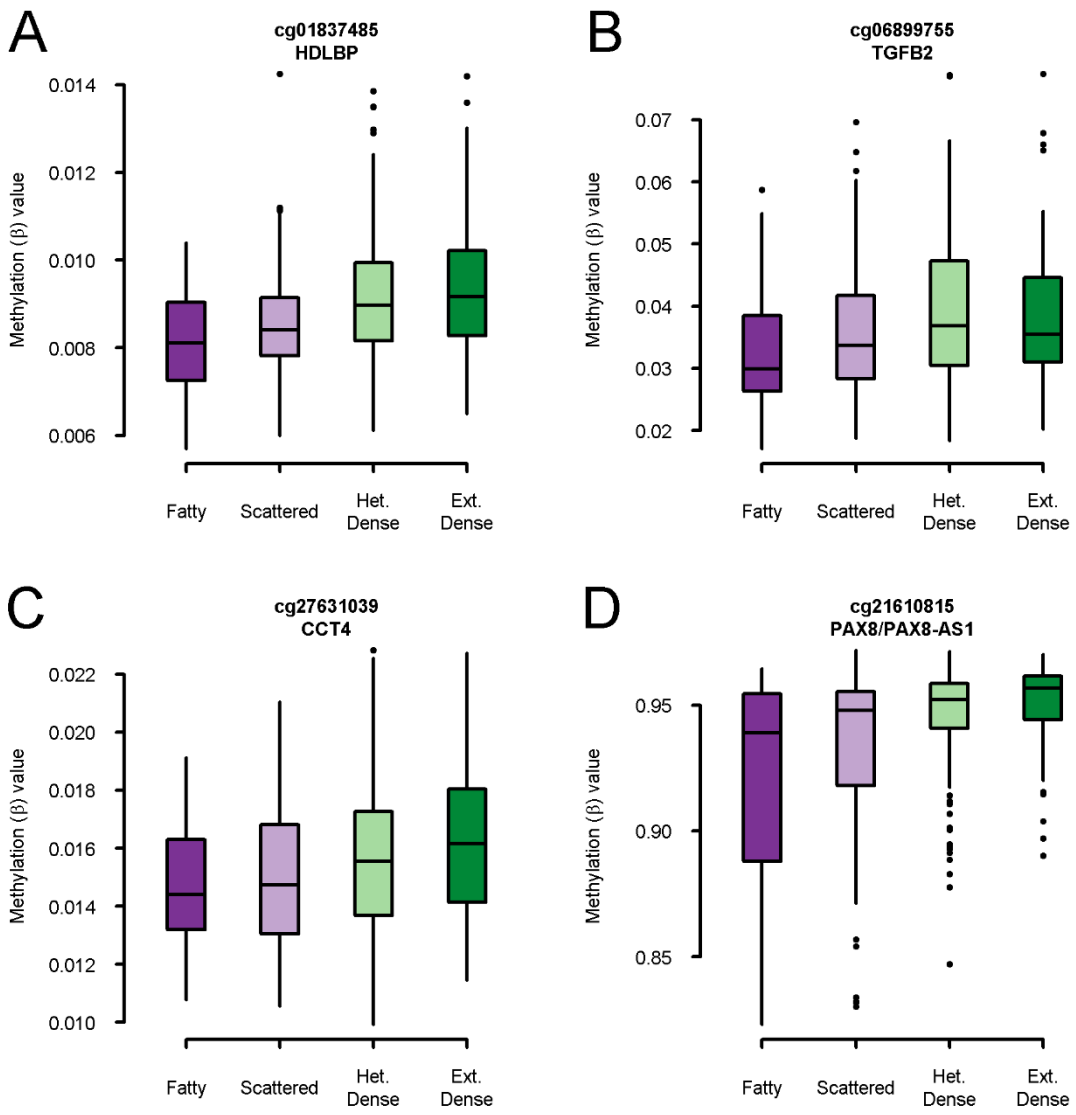

Figure 4.5. Boxplot of methylation (β) value by mammographic density category for top four intragenic probes by p-value

A) cg01837485 in *HDLBP* ($p = 1.7 \times 10^{-8}$); B) cg06899755 in *TGFB2* ($p = 2.0 \times 10^{-8}$); C) cg27631039 in *CCT4* ($p = 2.9 \times 10^{-8}$); D) cg21610815 in *PAX8/PAX8-AS1* ($p = 5.4 \times 10^{-8}$)

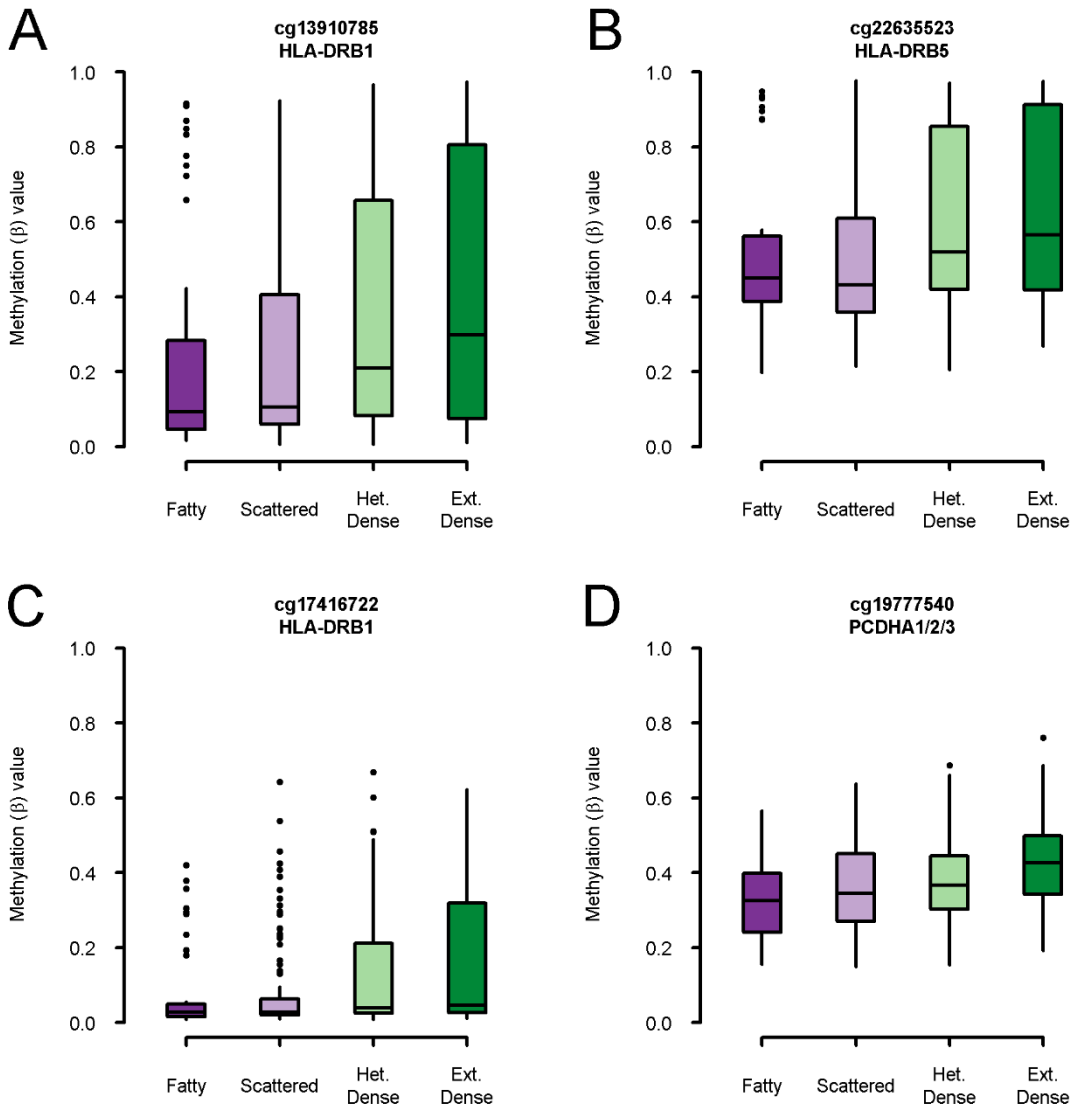

Figure 4.6. Boxplot of methylation (β) value by mammographic density category for top four intragenic probes by effect size (delta-β)
A) cg13910785 in HLA-DRB1 ($p = 4.7 \times 10^{-6}$); B) cg22635523 in HLA-DRB5 ($p = 1.1 \times 10^{-5}$); C) cg17416722 in HLA-DRB1 ($p = 1.1 \times 10^{-5}$); D) cg19777540 in PCDHA1/2/3 ($p = 1.1 \times 10^{-5}$)

Hypermethylated probes were more likely to be located within CpG islands and less likely to be in the open sea, while the opposite was true for hypomethylated probes (Figure 4.7). Furthermore, hypermethylated probes were more likely to be located within predicted promoters and actively repressed regions and less likely to be located in the transcribed and

inactive regions. Hypomethylated probes were more often located in actively transcribed regions and less often in promoter regions.

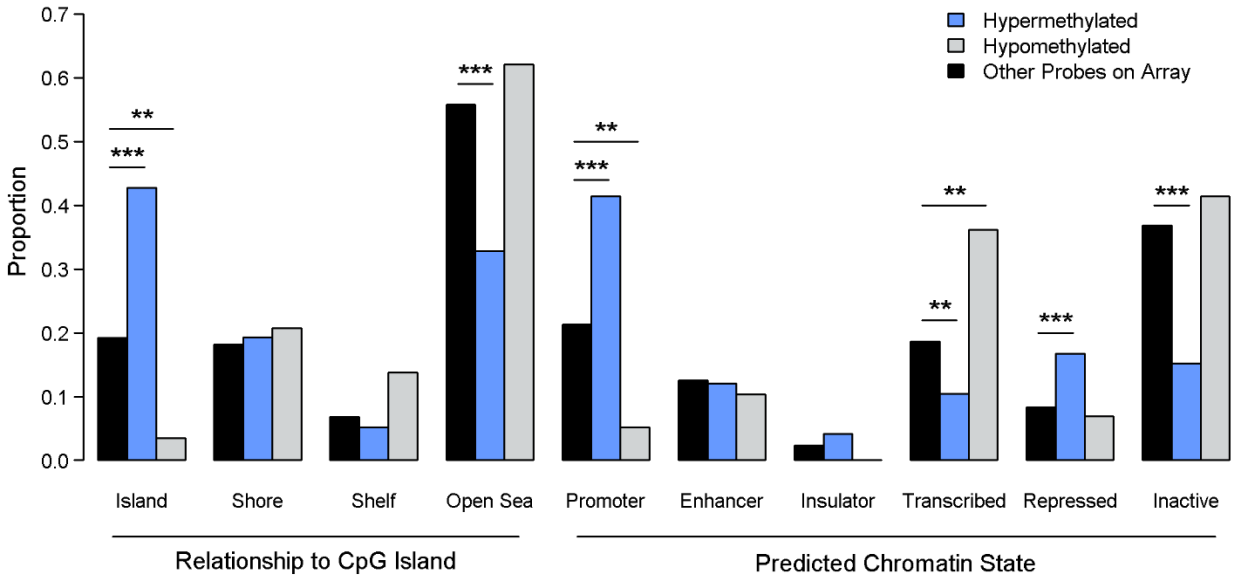

Figure 4.7. Enrichment analysis for genomic context of differentially methylated probes (DMPs) associated with increasing mammographic density
p-values are from X^2 test. *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$

140 of the 250 DMPs are also on the 450k array and could thus be examined for consistency of DNA methylation between paired blood cells and breast tissue. For 13 of the 140 DMPs with data available (9%), methylation in blood samples was statistically significantly correlated with methylation in paired breast tissue after strict Bonferroni correction ($p < 3.6 \times 10^{-4}$, $r \geq 0.39$). DMPs with the greatest correlation between blood and breast tissue included cg17416722 and cg13910785 in *HLA-DRB1* ($r \geq 0.98$), two intergenic probes on chromosomes 4 and 5 (cg01815292 and cg14964336), and cg21610815 in *PAX8/PAX8-AS1* ($r = 0.54$) (Table 4.4).

Regional analyses

There were 37 DMRs significantly associated with mammographic density in ≥90% of subsamples (Table 4.5). 36 (97%) of these regions were hypermethylated in women with higher mammographic density. The smallest FDRs were for regions within the *PAX8/PAX8-AS1* genes, intergenic regions at chr5p15.33 and chr1q21.3, and the *HLA-DRB1* gene; these same regions and another region within the *PF4* promoter had the largest effect sizes. Probes in these five regions also exhibited the strongest correlation between blood and breast tissue ($r \geq 0.5$) among the DMRs. Methylation in 7 DMRs (19%) was significantly correlated between blood and breast tissue after Bonferroni correction ($r \geq 0.36$).

Gene ontology/pathway analysis

Enriched gene ontology (GO) terms, KEGG pathways, and Ingenuity canonical pathways with a significance level of $p < 0.05$ and a minimum of 3 genes were identified (Appendix A, Table A9). For genes mapping to DMPs, 24 GO terms were enriched, including several relating to regulation of transcription or DNA binding. 12 KEGG pathways were enriched, including pathways in cancer, the Ras signaling pathway, non-small cell lung cancer, and thyroid cancer. A further 24 Ingenuity canonical pathways were enriched, including several cancer-related pathways (non-small cell lung cancer signaling, molecular mechanisms of cancer, and glioma signaling) and estrogen receptor signaling. For genes associated with DMRs, six GO terms were enriched, again including transcription- and DNA-binding-related terms.

Table 4.5. Genomic regions significantly associated with increasing mammographic density in ≥90% of subsamples

Chromosome	Start Position (hg19)	End Position (hg19)	Frequency	Combined FDR for Probes in Region	Average delta-β for Probes in Region	Width (bases)	Number of CpGs in Region	Gene(s)	Location within Gene	Average Correlation (Blood vs. Breast Tissue)	Average Correlation p
2	113,992,694	113,994,578	999	1.0E-6	2.52	1884	10	*PAX8; PAX8-AS1*	Promoter/Body; Promoter	0.806	<0.001
5	1,724,892	1,725,823	999	3.7E-5	1.80	931	7	-	Intergenic	0.548	<0.001
1	152,161,237	152,162,025	999	4.0E-5	2.80	788	7	-	Intergenic	0.744	<0.001
6	32,551,749	32,552,453	1000	6.0E-5	-6.70	704	7	*HLA-DRB1*	Body	0.750	<0.001
3	49,459,143	49,460,566	1000	8.6E-5	0.63	1423	14	*AMT; NICN1*	Promoter; 3'UTR	0.130	0.25
16	4,714,229	4,715,122	999	1.9E-4	0.55	893	7	*MGRN1*	Body	-0.016	0.89
14	97,924,757	97,925,676	999	2.1E-4	1.36	919	8	-	Intergenic	0.256	<0.001
4	74,847,710	74,848,016	978	3.2E-4	2.68	306	7	*PF4*	Promoter	0.848	<0.001
5	42,952,113	42,952,369	989	3.8E-4	0.41	256	3	-	Intergenic	0.324	0.003
4	1,294,432	1,295,078	988	7.4E-4	0.73	646	8	*MAEA*	Body	0.020	0.86
6	31,733,619	31,734,580	996	8.2E-4	0.70	961	17	*VWA7*	Body	0.288	0.009
6	157,931,791	157,932,180	926	9.7E-4	1.33	389	5	*ZDHHC14*	Body	0.083	0.462
2	170,589,808	170,589,904	956	1.0E-3	0.87	96	4	*KLHL23; PHOSPO2-KLHL23*	Promoter; Promoter	0.292	0.009
7	150,498,205	150,498,843	989	1.7E-3	1.67	638	5	*TMEM176A; TMEM176B*	Body; Promoter	0.197	0.08
22	46,508,454	46,508,604	948	2.2E-3	0.50	150	5	*MIRLET7BHG; MIRLET7B; MIRLET7A3*	Body; Promoter; Promoter	0.035	0.76
12	50,355,618	50,355,995	961	2.7E-3	1.21	377	3	*AQP5; LOC101927318*	Promoter; Body	0.604	<0.001
12	323,242	323,905	988	3.6E-3	0.38	663	9	*SLC6A12*	Promoter	0.079	0.48
6	32,222,492	32,223,226	965	4.7E-3	0.32	734	6	-	Intergenic	0.410	<0.001
11	94,037,478	94,038,553	992	7.2E-3	1.05	1075	5	*IZUMO1R*	Promoter	0.209	0.06
5	72,746,710	72,747,139	949	7.3E-3	0.31	429	6	-	Intergenic	0.078	0.49
5	156,570,959	156,571,045	916	8.5E-3	0.84	86	2	*MED7*	Promoter	0.095	0.40
5	138,861,106	138,861,241	924	8.7E-3	0.57	135	2	*TMEM173*	Body	0.017	0.88

Table continues

Table 4.5. Genomic regions significantly associated with increasing mammographic density in ≥90% of subsamples (continued)

Chromosome	Start Position (hg19)	End Position (hg19)	Frequency	Combined FDR for Probes in Region	Average delta-β for Probes in Region	Width (bases)	Number of CpGs in Region	Gene(s)	Location within Gene	Average Correlation (Blood vs. Breast Tissue)	Average Correlation p
10	28,034,669	28,035,208	994	9.3E-3	0.18	539	11	*MKX; MKX-AS1*	Promoter; Body	-0.006	0.96
5	146,257,862	146,259,003	1000	9.7E-3	0.25	1141	19	*PPP2R2B*	Promoter/Body	0.069	0.54
22	31,002,739	31,003,283	900	1.1E-2	0.21	544	12	*TCN2; PES1*	Promoter; Promoter	0.068	0.45
22	17,849,639	17,850,130	944	1.1E-2	0.19	491	4	-	Intergenic	0.078	0.49
15	60,296,138	60,296,996	980	1.3E-2	0.11	858	11	*FOXB1*	Promoter	0.001	0.99
13	33,589,621	33,590,048	972	1.4E-2	0.49	427	5	*KL*	Promoter	0.081	0.38
3	47,620,532	47,621,127	973	2.0E-2	0.47	595	7	*CSPG5*	Promoter	0.206	0.11
8	40,755,165	40,755,559	935	2.2E-2	0.22	394	9	*ZMAT4*	Promoter	0.070	0.54
16	10,277,254	10,277,505	986	2.3E-2	0.41	251	2	*GRIN2A*	Promoter	0.219	0.05
4	55,098,359	55,098,483	914	3.5E-2	0.11	124	2	*PDGFRA*	Promoter	-0.056	0.62
11	82,444,456	82,445,114	964	7.8E-2	0.26	658	6	*FAM181B*	Promoter	0.115	0.31
10	46,168,220	46,168,841	999	0.11	0.36	621	13	*ZFAND4*	Promoter/Body	0.193	0.09
15	37,393,188	37,393,589	968	0.14	0.08	401	8	*MEIS2*	Promoter	-0.009	0.93
7	27,209,463	27,209,582	910	0.17	0.14	119	2	*HOXA9; HOXA10*	Body	-0.039	0.73
6	33,244,976	33,246,390	1000	0.23	0.23	1414	41	*B3GALT4*	Promoter/Body	0.100	0.38

Individual probe FDRs were combined using Stouffer's method.

4.4 Discussion

We performed an epigenome-wide association study of mammographic density using the Illumina Infinium MethylationEPIC BeadChip array. First, we identified a pattern of hypermethylation at CpG islands and gene promoters in women with elevated mammographic density. Second, we identified 250 CpG sites for which DNA methylation was significantly associated with mammographic density using a resampling-based method. The top sites were located within genes associated with cancer, including *HDLBP*, *TGFB2*, *CCT4*, and *PAX8*, and were more likely to be located in regulatory regions of the genome. Finally, we identified 37 regions in which DNA methylation was associated with mammographic density, including within the promoters of *PAX8* and *PF4*, a gene involved in the regulation of angiogenesis. Overall, our results paint a picture of epigenetic dysregulation associated with mammographic density, which may impact transcription of cancer-related genes.

There was increased genome-wide DNA methylation associated with mammographic density, particularly among probes within CpG islands and predicted promoters. This pattern was robust to sensitivity analysis. This is consistent with the pattern of promoter hypermethylation that is commonly associated with cancer[250], including breast cancer[9]. Some previous studies have associated genome-wide hypomethylation measured with the 450k array with risk of breast cancer[8,9], but other studies[251] and meta-analyses[252,253] do not support those findings, and a study directly examining the impact of mammographic density on genome-wide methylation did not find a relationship[254]. Epigenetic age acceleration has been associated with breast cancer risk[19,210]; however, consistent with a previous study[254], we did not observe an increase in epigenetic age acceleration with mammographic density.

The top intragenic DMPs by p-value were located within the *HDLBP*, *TGFB2*, *CCT4*, and *PAX8* genes. *HDLBP* codes for the protein vigilin, which is strongly expressed in normal breast epithelium[255] and is a candidate tumor suppressor[256] that has been shown to impair proliferation of breast cancer cells in vitro[257]. *TGFB2* encodes an isoform of TGF-β, a key regulator of cell development and proliferation with both tumor-suppressor and oncogenic functions[258], including in breast cancer[259]. Epigenetic dysregulation of TGF- β signaling is a feature of breast cancer cells[260] and suppression (although not by DNA methylation) of TGF-β2 expression has been previously described in breast cancer cells and tumor samples[261]. Hypermethylation of *TGFB2* has been associated with poorer survival in pancreatic cancer[262] and with prostate cancer progression[263]. *CCT4* encodes a subunit of chaperonin-containing TCP-1 or CCT which supervises folding of proteins, including those critical for the development of cancer[264]. Little data is available regarding the role of CCT4 in carcinogenesis, although one study found a relationship between CCT4 expression and overall survival in breast cancer[265]. Other CCT subunits, particularly CCT2 and CCT3, are overexpressed in a significant proportion of breast cancers and their expression is associated with increased invasiveness[266], proliferation[267], and poor overall survival[265,268] in breast cancer. PAX8 is commonly expressed in tumors of Mullerian origin[269], and some studies have identified PAX8 expression in a significant proportion of metastatic breast cancers[270] and breast cancer cell lines[271]. Another found evidence that reduced expression of an isoform of the *PAX8* antisense RNA *PAX8-AS1* was associated with poor survival in breast cancer[272]. Interestingly, a CpG site in the closely related and breast-cancer associated[273,274] *PAX6* gene was also hypermethylated in women with high mammographic density.

In addition to hypermethylation in the *PAX8* promoter, our regional analysis identified a hypermethylated region at the *PF4* promoter associated with mammographic density. *PF4* encodes a chemokine, CXCL4, which can inhibit the proliferation and migration of endothelial cells[275] and suppress angiogenesis[276]. Its expression is dysregulated in many cancers[275], including breast cancer[277]. One study found silencing of *PF4* associated with promoter hypermethylation in multiple myeloma patients[278]. Since it is often assumed that regional analysis identifies differential methylation with a greater likelihood of impacting gene expression[279], our results suggest that *PF4* and *PAX8* expression may be associated with elevated mammographic density, a question that should be studied further.

Overall, the top DMPs and DMRs point to epigenetic dysregulation in genes associated with breast carcinogenesis. It is important to note that none of these genes have clear mechanistic links to mammographic density, but rather to breast cancer risk. This may suggest that the differential methylation described in this study is part of the biological mechanism linking mammographic density and breast cancer development. Without follow-up data on breast cancer outcomes, it is impossible to state if methylation at these sites is a prelude to or mediator of breast cancer development, but methylation at 9 of the 206 genes located near the 250 DMPs and one of the 37 regions has been associated with breast cancer risk in previous studies (*TACC2*[280], *PDGFD*[251], *ZFAND4*[251], *HSCB*[280], *MAD1L*[281], *PRDM16*[251], *C6orf141*[251], *SLITRK5*[251], *NCOR2*[251]). Further studies are warranted to understand the role of DNA methylation in the relationship between mammographic density and breast cancer.

A recent epigenome-wide association study of mammographic density using the Illumina Infinium HumanMethylation450 BeadChip array did not show any loci associated with mammographic density after adjusting for multiple testing[84]. Several factors may

explain these disparate results. First, we used the updated MethylationEPIC array which covers an additional 400,000 CpG sites compared to the 450k array used in the previous study. Also, we used the qualitative BI-RADS density categorization, rather than the quantitative measures used in the previous study (percent density, dense area, and non-dense area). These measures are generally correlated[282,283], but nonetheless reflect different approaches to measuring mammographic density. Finally, we used the false discovery rate adjustment for multiple comparisons, while the previous study used the stricter Bonferroni correction. If we had used the same statistical approach, we would have identified just 7 DMPs, including those near the *HDLBP*, *TGFB2*, *CCT4*, *PAX8*, and *TACC2* genes, and two intergenic probes at chr6p22.3 and chr20q11.23.

The only other study of mammographic density and DNA methylation to date was performed using prediagnostic blood samples from women who subsequently developed breast cancer[85]. This study used a case-control design, matching women with BI-RADS density category A (low density) to women with BI-RADS density category C or D (high density), and measured DNA methylation using the MethylationEPIC BeadChip. No probes were significant after false discovery rate correction, but analysis of DNA methylation at promoter regions identified 140 genes whose promoter methylation was associated with mammographic density. Examination of these genes revealed overlap with our DMPs and DMRs, suggesting genes that may be reliably associated with mammographic density in distinct populations. Hypermethylation of probes in the promoters of *SLITRK5* and *TMEM176A* and regions in the promoters of *MKX* and *MEIS2* were associated with mammographic density in our study; these gene promoters were also hypermethylated in the analysis by Caini et al. *SLITRK5* hypomethylation has been previously associated with

breast cancer risk[251]. The other study also identified hypermethylation at the promoter of the *PF4V1* gene, which encodes a variant of the PF4 protein. Hypermethylation at the *PF4* promoter was associated with elevated mammographic density in our study. A further cluster of mammographic density-related genes may be found in the platelet-derived growth factor family: in our study, a probe in *PDGFD* and a region in *PDGFRA* were associated with mammographic density, in the Caini et al. study, the *PDGFC* promoter was hypermethylated in women with elevated mammographic density, and *PDGFD* hypermethylation has been associated with breast cancer risk[251]. Despite the differences in analytic approach, the results of this study suggest further avenues for exploration of the relationship between mammographic density and DNA methylation and identify loci whose methylation may mediate the relationship between mammographic density and breast cancer risk.

We undertook several important steps to improve the validity of our results. We implemented a resampling-based method to improve the stability of results from epigenome-wide association studies. In 1,000 subsamples of our study cohort, we identified 33,015 unique CpG sites associated with mammographic density; however, the vast majority of these sites (79%) were found in ≤5% of subsamples. The number of sites significant at FDR <0.05 varied greatly for each subsample, from a minimum of 123 to a maximum of 17,209. These results highlight the sensitivity of EWAS methods to even small differences in the study cohort, and emphasize the importance of screening lists of DMPs for false positives.

The sample size of 385 women should be sufficient to identify DMPs and DMRs associated with mammographic density; epigenome-wide studies of other disease risk factors have used cohorts of a similar size[14,15,18,20,235,237]. Our resampling method is designed to improve the specificity of the results and reduce the number of false positives identified,

which are a major hurdle to replication of epigenome-wide association results. The results of the positive control (identifying loci associated with cigarette smoking) suggest that the method can efficiently identify true positive associations while limiting the number of spurious findings. Using our method versus a "traditional" epigenome-wide association approach with the entire study cohort, we compared the proportion of DMPs with previously-established relationships with smoking in other, larger cohort studies. Our study was not designed or powered to examine relationships between smoking and DNA methylation, so it is a reasonable assumption that most results inconsistent with previous literature are false positives. 35/40 (87.5%) DMPs from the resampling method and 164/297 (55.2%) from the traditional method had been previously associated with smoking. These results suggest that resampling-based methods can facilitate identification of differential methylation signals with a high degree of sensitivity and specificity, even in the absence of an external validation dataset.

Blood is a convenient tissue to use to measure DNA methylation and develop biomarkers; however, the extent to which it can be used as a surrogate tissue for breast tissue is unknown. To examine this question, we leveraged a dataset from a previous study[245] which measured DNA methylation in both white blood cells and breast tissue from 40 women. We calculated the correlation between paired blood and breast tissue DNA methylation among our DMPs and DMRs and found that, where data were available, about 9% of DMPs were significantly correlated between blood and breast tissue after multiple testing correction. Similarly, 19% of DMRs overall were statistically significantly correlated between blood and breast tissue, including regions located in the *PF4* and *PAX8* promoters. These data show that, at least for some loci associated with mammographic density,

differences in DNA methylation in blood may reflect similar differences in the breast tissue. In particular, blood DNA methylation at regions associated with mammographic density was more likely to be well-correlated with breast tissue DNA methylation, which supports the common assumption that DMRs have greater functional significance than individual probes[279].

There are several important limitations of this study. First, this is a cross-sectional study and cannot determine the temporal relationship between the differential methylation observed and mammographic density. Although we observed differential methylation in active regulatory regions and near genes with possible connections to breast cancer development, we did not examine differences in gene expression to determine whether these differences had a functional impact. This is an important consideration, especially given the small effect sizes observed for some DMPs and DMRs, which are unlikely to have biological impact. In addition, because our study population consisted entirely of postmenopausal women aged 45-66, our results may not be generalizable to all women, for example, premenopausal women or women substantially older. We used a qualitative measure of mammographic density rather than the quantitative percent density. We chose to use the qualitative BI-RADS density category because it is the most commonly used clinical measurement of mammographic density in the United States, but it may suffer from decreased interobserver reliability, especially for the middle two categories[284–286], which could have reduced our power to detect differential DNA methylation. However, a sensitivity analysis with the cohort limited to women with category A or category D BI-RADS density (n = 118) yielded similar results to the main analysis, suggesting that this limitation did not majorly impact our results. Lastly, we did not validate our results in an independent cohort;

however, our identification of previously-established smoking-associated probes suggests that the resampling-based method we used yields results that are stable and replicable.

Our study identifies differential DNA methylation associated with mammographic density, an important risk factor for breast cancer. We leveraged a well-characterized cohort and a novel resampling-based method to improve the stability and reproducibility of our findings. Our results suggest the potential involvement of several genes, including *PAX8*, *PF4*, *HDLBP*, *TGFB2*, and *CCT4*, in the biological mechanisms behind differences in mammographic density between women. Further studies are warranted to explore these potential mechanisms and potential links to breast cancer risk.

Conclusion

In this work, I identify dietary factors associated with herbicide residues in urine and DNA methylation signatures of urinary glyphosate and mammographic density. This is also the largest study to date collecting data on urinary concentrations of glyphosate and AMPA in the non-agricultural setting in the United States and highlights the near-ubiquitous exposure to these compounds in the general population. Differential methylation associated with both urinary glyphosate and mammographic density was located near genes that could promote the development of cancer, providing evidence that DNA methylation could serve as a mediator between exposures and risk for disease. Additionally, I showed that an index based on DNA methylation at 15 CpG sites could accurately predict urinary glyphosate in an internal validation set, taking the first step toward developing a DNA methylation-based biomarker for glyphosate exposure. To achieve these goals, I developed and tested a novel method for epigenome-wide association analysis and showed that this method reduces the number of type I errors in epigenome-wide association studies.

The majority of women in this study (>80%) had detectable glyphosate and AMPA in at least one urine sample. Although the sample may not be representative of the general population, this is generally consistent with the prevalence observed in other studies in the U.S. I found that urinary glyphosate was associated with increased consumption of grains, even among women who reported often or always eating organic grains. Additionally, glyphosate and AMPA were not significantly associated with organic eating habits after adjustment for diet quality, race/ethnicity, and physical activity, despite glyphosate being prohibited in organic agriculture. This could be due to inconsistent dietary habits, where self-reported frequent organic eaters still consume a significant quantity of conventional

foods, or the unexpected presence of glyphosate and AMPA in the organic food supply. Soy and alcohol consumption were associated with elevated urinary AMPA, but not glyphosate, suggesting distinct pathways of exposure for these related compounds. Further study is needed to verify these results and confirm foods likely to be contaminated with glyphosate and AMPA residues. However, these results are an essential first step towards understanding dietary and lifestyle factors that may be predictive of an individual's exposure to glyphosate and AMPA, which will be important for future epidemiologic studies that attempt to clarify the effect of the herbicide on human health.

The epigenome-wide association study of glyphosate and AMPA described in Chapter 3 is another important step towards understanding the impact of these compounds on the epigenome and, by extension, their potential health effects. The sites and regions associated with glyphosate and AMPA concentrations were located near genes associated with cancer (*SF3B2*, *MSH4,* and *TRIM31*) and endocrine disruption (*ESR1*). Exposure to AMPA was also associated with increased epigenetic age acceleration, a phenomenon which has been associated with the risk of many diseases. These results provide some evidence of possible impacts of glyphosate and AMPA exposure on the body, but should be interpreted with caution given the lack of gene expression data to verify the functional impact of the differences in DNA methylation and the cross-sectional nature of the study. Still, as the first epigenome-wide association study of glyphosate and AMPA, it sheds light on a possible mechanism for adverse health effects. Additionally, the development of a validated methylation-based index of glyphosate exposure is a first step towards the development of an effective biomarker of past glyphosate exposure. Such a biomarker would be an important tool for longitudinal studies of glyphosate exposure and disease risk, and this study provides

evidence that a DNA methylation biomarker of glyphosate could be effective.

Chapter 4 describes differential DNA methylation associated with elevated mammographic density, a strong risk factor for breast cancer. The mechanism by which elevated mammographic density increases breast cancer risk is not well-understood, but evidence from studies of DNA methylation, smoking, and lung cancer risk suggests that DNA methylation could shed light on genes or pathways that may be involved in conferring risk for disease. In this study, the top sites associated with mammographic density were located within genes associated with cancer, including *HDLBP*, *TGFB2*, *CCT4*, *PAX8*, and *PF4*, and were more likely to be located in regulatory regions of the genome, suggesting that the differential DNA methylation could have functional impacts. The method applied to identify these sites may reduce the number of false positive findings. In fact, our results overlap significantly with one of the only other studies of DNA methylation markers associated with mammographic density, and suggest that methylation at *SLITRK5, TMEM176A, MKX,* and *MEIS2* may be reliably associated with mammographic density. Because the other study was limited to women who later developed breast cancer, these loci are of particular interest as possible mediators of the relationship between mammographic density and breast cancer risk. Thus, the work presented in Chapter 4 is an important step forward in characterizing the impact of mammographic density on the epigenome. Further studies are warranted to better understand the role of DNA methylation in the relationship between mammographic density and breast cancer, especially given the cross-sectional nature of our study and lack of follow-up data on breast cancer in the study population.

Another important finding of this work is the development and validation of a resampling-based method for epigenome-wide association studies, which is described

further in Appendix C. The high dimensionality and technical noise present in DNA methylation array data leads to a high proportion of false positive results whose associations with the exposure or outcome of interest are not replicated in further studies with separate study populations. Resampling the study population and combining results from many of these subsamples can improve the stability and replicability of the results, as outlined in the positive control analysis described in Chapter 4. I compared the results from a traditional epigenome-wide approach to my resampling-based approach for an epigenome-wide association study of smoking. Results which were inconsistent with previous epigenome-wide association studies of smoking are more likely to be false positive results. Nearly half (44.8%) of the sites identified by the traditional approach had not been previously associated with smoking, compared to only 12.5% from the resampling method. I also show that lists of DMPs from varying subsamples of a study population can have limited overlap, with over 80% of DMPs being identified in only a single subsample. This further highlights the limitations of the traditional approach and its sensitivity to fluctuations in the study population. Studies which utilize resampling could produce more stable lists of differentially methylated sites that are more likely to be replicated in future studies.

In this work, I conduct a comprehensive analysis of dietary factors associated with glyphosate and AMPA exposure in a U.S. population and apply novel methods to understand the impact of glyphosate exposure and mammographic density on the epigenome. The results of the studies presented here inform our understanding of the mechanisms by which these factors can influence disease risk and lay the groundwork for the development of methylation-based biomarkers for glyphosate exposure.

References

1. Varley KE, Gertz J, Bowling KM, et al. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*. 2013;23(3):555-567. doi:10.1101/GR.147942.112

2. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA methylation in cancer and aging. *Cancer Res*. 2016;76(12):3446-3450. doi:10.1158/0008-5472.CAN-15-3278

3. Esteller M, Corn PG, Baylin SB, Herman JG. A gene hypermethylation profile of human cancer. *Cancer Res*. 2001;61(8):3225-3229. http://www.ncbi.nlm.nih.gov/pubmed/11309270.

4. Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet*. 2000;16(4):168-174. doi:10.1016/S0168-9525(99)01971-X

5. Herman JG, Baylin SB. Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N Engl J Med*. 2003;349(21):2042-2054. doi:10.1056/NEJMra023075

6. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet*. 2006;7(1):21-33. doi:10.1038/nrg1748

7. Xu Z, Bolick SCE, DeRoo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide association study of breast cancer using prospectively collected Sister Study samples. *JNCI J Natl Cancer Inst*. 2013;105(10):694-700. doi:10.1093/jnci/djt045

8. van Veldhoven K, Polidoro S, Baglietto L, et al. Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. *Clin Epigenetics*. 2015;7(1):67. doi:10.1186/s13148-015-0104-2

9. Severi G, Southey MC, English DR, et al. Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. *Breast Cancer Res Treat*. 2014;148(3):665-673. doi:10.1007/s10549-014-3209-y

10. Ennour-Idrissi K, Dragic D, Durocher F, Diorio C. Epigenome-wide DNA methylation and risk of breast cancer: a systematic review. *BMC Cancer*. 2020;20(1). doi:10.1186/s12885-020-07543-4

11. Dietrich D, Kneip C, Raji O, et al. Performance evaluation of the DNA methylation biomarker SHOX2 for the aid in diagnosis of lung cancer based on the analysis of bronchial aspirates. *Int J Oncol*. 2012;40(3):825-832. doi:10.3892/ijo.2011.1264

12. Warren JD, Xiong W, Bunker AM, et al. Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Med*. 2011;9(1):133. doi:10.1186/1741-7015-9-133

13. A stool DNA test (Cologuard) for colorectal cancer screening. *JAMA*. 2014;312(23):2566. doi:10.1001/jama.2014.15746

14. Shenker NS, Polidoro S, van Veldhoven K, et al. Epigenome-wide association study in

the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet*. 2013;22(5):843-851. doi:10.1093/hmg/dds488

15. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet*. 2011;88(4):450-457. doi:10.1016/j.ajhg.2011.03.003

16. Baglietto L, Ponzi E, Haycock P, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *Int J Cancer*. 2017;140(1):50-61. doi:10.1002/ijc.30431

17. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):1-20. doi:10.1186/gb-2013-14-10-r115

18. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*. 2013;49(2):359-367. doi:10.1016/J.MOLCEL.2012.10.016

19. Ambatipudi S, Horvath S, Perrier F, et al. DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *Eur J Cancer*. 2017;75:299-307. doi:10.1016/j.ejca.2017.01.014

20. Dick KJ, Nelson CP, Tsaprouni L, et al. DNA methylation and body-mass index: A genome-wide analysis. *Lancet*. 2014;383(9933):1990-1998. doi:10.1016/S0140-6736(13)62674-4

21. Geurts YM, Dugué P-A, Joo JE, et al. Novel associations between blood DNA methylation and body mass index in middle-aged and older adults. *Int J Obes*. 2018;42(4):887-896. doi:10.1038/ijo.2017.269

22. Lu S, Niu Z, Chen Y, et al. Repetitive element DNA methylation is associated with menopausal age. *Aging Dis*. 2018;9(3):435. doi:10.14336/AD.2017.0810

23. Levine ME, Lu AT, Chen BH, et al. Menopause accelerates biological aging. *Proc Natl Acad Sci*. 2016;113(33):9327-9332. doi:10.1073/pnas.1604558113

24. Demetriou CA, Chen J, Polidoro S, et al. Methylome analysis and epigenetic changes associated with menarcheal age. Yan W, ed. *PLoS One*. 2013;8(11):e79391. doi:10.1371/journal.pone.0079391

25. Lu X, Fraszczyk E, van der Meer TP, et al. An epigenome-wide association study identifies multiple DNA methylation markers of exposure to endocrine disruptors. *Environ Int*. 2020;144:106016. doi:10.1016/j.envint.2020.106016

26. Curtis SW, Cobb DO, Kilaru V, et al. Exposure to polybrominated biphenyl (PBB) associates with genome-wide DNA methylation differences in peripheral blood. *Epigenetics*. 2019;14(1):52-66. doi:10.1080/15592294.2019.1565590

27. Furlong MA, Paul KC, Yan Q, et al. An epigenome-wide association study of ambient pyrethroid pesticide exposures in California's central valley. *Int J Hyg Environ Health*.

2020;229. doi:10.1016/j.ijheh.2020.113569

28. Lee MK, Xu CJ, Carnes MU, et al. Genome-wide DNA methylation and long-term ambient air pollution exposure in Korean adults. *Clin Epigenetics*. 2019;11(1). doi:10.1186/s13148-019-0635-z

29. Gondalia R, Baldassari A, Holliday KM, et al. Methylome-wide association study provides evidence of particulate matter air pollution-associated DNA methylation. *Environ Int*. 2019;132. doi:10.1016/j.envint.2019.03.071

30. Sayols-Baixeras S, Fernández-Sanlés A, Prats-Uribe A, et al. Association between long-term air pollution exposure and DNA methylation: The REGICOR study. *Environ Res*. 2019;176. doi:10.1016/j.envres.2019.108550

31. Shenker NS, Ueland PM, Polidoro S, et al. DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*. 2013;24(5):712-716. doi:10.1097/EDE.0b013e31829d5cb3

32. Teschendorff AE, Yang Z, Wong A, et al. Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. *JAMA Oncol*. 2015;1(4):476-485. doi:10.1001/jamaoncol.2015.1053

33. Liu C, Marioni RE, Hedman AK, et al. A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry*. 2018;23(2):422-433. doi:10.1038/mp.2016.192

34. Zhang Y, Yang R, Burwinkel B, Breitling LP, Brenner H. F2RL3 methylation as a biomarker of current and lifetime smoking exposures. *Environ Health Perspect*. 2014;122(2):131-137. doi:10.1289/EHP.1306937

35. Joubert BR, Felix JF, Yousefi P, et al. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *Am J Hum Genet*. 2016;98(4):680-696. doi:10.1016/j.ajhg.2016.02.019

36. Zhang Y, Elgizouli M, Schöttker B, Holleczek B, Nieters A, Brenner H. Smoking-associated DNA methylation markers predict lung cancer incidence. *Clin Epigenetics*. 2016;8(1):127. doi:10.1186/s13148-016-0292-4

37. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun*. 2015;6(1):10192. doi:10.1038/ncomms10192

38. Zhang Y, Schöttker B, Ordóñez-Mena J, et al. F2RL3 methylation, lung cancer incidence and mortality. *Int J Cancer*. 2015;137(7):1739-1748. doi:10.1002/ijc.29537

39. Benbrook CM. Trends in glyphosate herbicide use in the United States and globally. *Environ Sci Eur*. 2016;28(1):1-15. doi:10.1186/s12302-016-0070-0

40. Duke SO. The history and current status of glyphosate. *Pest Manag Sci*. 2018;74(5):1027-1034. doi:10.1002/ps.4652

41. Gaines TA. The importance of glyphosate in non-GM settings. *Outlooks Pest Manag*. 2018;29(6):255-257. doi:10.1564/V29_DEC_05

42. Wiese A, Schulte M, Theuvsen L, Steinmann H-H. Interactions of glyphosate use with farm characteristics and cropping patterns in Central Europe. *Pest Manag Sci*. 2018;74(5):1155-1165. doi:10.1002/PS.4542

43. Mills PJ, Kania-Korwel I, Fagan J, McEvoy LK, Laughlin GA, Barrett-Connor E. Excretion of the herbicide glyphosate in older adults between 1993 and 2016. *J Am Med Assoc*. 2017;318(16):1610-1611. doi:10.1001/jama.2017.11726

44. Lesseur C, Pirrotte P, Pathak K V, et al. Maternal urinary levels of glyphosate during pregnancy and anogenital distance in newborns in a US multicenter pregnancy cohort. *Environ Pollut*. 2021;280:117002. doi:10.1016/j.envpol.2021.117002

45. Schütze A, Morales-Agudelo P, Vidal M, Calafat AM, Ospina M. Quantification of glyphosate and other organophosphorus compounds in human urine via ion chromatography isotope dilution tandem mass spectrometry. *Chemosphere*. 2021;274:129427. doi:10.1016/j.chemosphere.2020.129427

46. Parvez S, Gerona RR, Proctor C, et al. Glyphosate exposure in pregnancy and shortened gestational length: A prospective Indiana birth cohort study. *Environ Heal*. 2018;17(1). doi:10.1186/s12940-018-0367-0

47. IARC (International Agency for Research on Cancer). *IARC Monographs Volume 112: Evaluation of Five Organophosphate Insecticides and Herbicides*. Lyon; 2015.

48. EPA (Environmental Protection Agency). Draft human health risk assessment in support of registration review. 2017.

49. FAO (Food and Agricultural Organization of the United Nations), WHO (World Health Organization). *Pesticide Residues in Food 2016 - Special Session of the Joint FAO/WHO Meeting on Pesticide Residues*.; 2016.

50. Franke AA, Li X, Shvetsov YB, Lai JF. Pilot study on the urinary excretion of the glyphosate metabolite aminomethylphosphonic acid and breast cancer risk: The Multiethnic Cohort study. *Environ Pollut*. 2021. doi:10.1016/j.envpol.2021.116848

51. Jayasumana C, Gunatilake S, Senanayake P. Glyphosate, hard water and nephrotoxic metals: Are they the culprits behind the epidemic of chronic kidney disease of unknown etiology in Sri Lanka? *Int J Environ Res Public Health*. 2014;11(2):2125-2147. doi:10.3390/ijerph110202125

52. Zhang L, Rana I, Shaffer RM, Taioli E, Sheppard L. Exposure to glyphosate-based herbicides and risk for non-Hodgkin lymphoma: A meta-analysis and supporting evidence. *Mutat Res - Rev Mutat Res*. 2019;781:186-206. doi:10.1016/j.mrrev.2019.02.001

53. Mesnage R, Renney G, Séralini GE, Ward M, Antoniou MN. Multiomics reveal non-alcoholic fatty liver disease in rats following chronic exposure to an ultra-low dose of Roundup herbicide. *Sci Rep*. 2017;7. doi:10.1038/srep39328

54. Ingaramo P, Alarcón R, Muñoz-de-Toro M, Luque EH. Are glyphosate and glyphosate-based herbicides endocrine disruptors that alter female fertility? *Mol Cell Endocrinol*.

2020;518. doi:10.1016/j.mce.2020.110934

55. Mao Q, Manservisi F, Panzacchi S, et al. The Ramazzini Institute 13-week pilot study on glyphosate and Roundup administered at human-equivalent dose to Sprague Dawley rats: Effects on the microbiome. *Environ Heal*. 2018;17(1). doi:10.1186/s12940-018-0394-x

56. Tang Q, Tang J, Ren X, Li C. Glyphosate exposure induces inflammatory responses in the small intestine and alters gut microbial composition in rats. *Environ Pollut*. 2020;261. doi:10.1016/j.envpol.2020.114129

57. Mesnage R, Teixeira M, Mandrioli D, et al. Use of shotgun metagenomics and metabolomics to evaluate the impact of glyphosate or roundup mon 52276 on the gut microbiota and serum metabolome of sprague-dawley rats. *Environ Health Perspect*. 2021;129(1):1-15. doi:10.1289/EHP6990

58. Thongprakaisang S, Thiantanawat A, Rangkadilok N, Suriyo T, Satayavivad J. Glyphosate induces human breast cancer cells growth via estrogen receptors. *Food Chem Toxicol*. 2013;59:129-136. doi:10.1016/j.fct.2013.05.057

59. Faniband MH, Norén E, Littorin M, Lindh CH. Human experimental exposure to glyphosate and biomonitoring of young Swedish adults. *Int J Hyg Environ Health*. 2021;231:113657. doi:10.1016/j.ijheh.2020.113657

60. Connolly A, Jones K, Basinas I, et al. Exploring the half-life of glyphosate in human urine samples. *Int J Hyg Environ Health*. 2019;222(2):205-210. doi:10.1016/j.ijheh.2018.09.004

61. Fagan J, Bohlen L, Patton S, Klein K. Organic diet intervention significantly reduces urinary glyphosate levels in U.S. children and adults. *Environ Res*. 2020;189. doi:10.1016/j.envres.2020.109898

62. FDA (U.S. Food and Drug Administration). *Pesticide Residue Monitoring Program Fiscal Year 2017 Pesticide Report.*; 2017. https://www.fda.gov/food/pesticides/pesticide-residue-monitoring-2017-report-and-data. Accessed August 10, 2021.

63. Kolakowski B, Miller L, Murray A, Leclair A, Bietlot H, van de Riet J. Analysis of glyphosate residues in foods from the Canadian retail markets between 2015 and 2017. *J Agric Food Chem*. 2020;68(18):5201-5211. doi:10.1021/ACS.JAFC.9B07819

64. Zoller O, Rhyn P, Rupp H, Zarn JA, Geiser C. Glyphosate residues in Swiss market foods: monitoring and risk evaluation. *Food Addit Contam Part B, Surveill*. 2018;11(2):83-91. doi:10.1080/19393210.2017.1419509

65. Xu J, Smith S, Smith G, Wang W, Li Y. Glyphosate contamination in grains and foods: An overview. *Food Control*. 2019;106:106710. doi:10.1016/J.FOODCONT.2019.106710

66. Medina-Pastor P, Triacchini G. The 2018 European Union report on pesticide residues in food. *EFSA J*. 2020;18(4):1-103. doi:10.2903/J.EFSA.2020.6057

67. Rubio F, Guo E, Kamp L. Survey of glyphosate residues in honey, corn and soy products. *J Environ Anal Toxicol*. 2014;5(1):249. doi:10.4172/2161-0525.1000249

68. John J, Liu H. Glyphosate monitoring in water, foods, and urine reveals an association between urinary glyphosate and tea drinking: A pilot study. *Int J Environ Health Eng*. 2018;7(1):2. doi:10.4103/IJEHE.IJEHE_5_17

69. Soukup ST, Merz B, Bub A, et al. Glyphosate and AMPA levels in human urine samples and their correlation with food consumption: results of the cross-sectional KarMeN study in Germany. *Arch Toxicol*. 2020;94(5):1575-1584. doi:10.1007/s00204-020-02704-7

70. Stajnko A, Snoj Tratnik J, Kosjek T, et al. Seasonal glyphosate and AMPA levels in urine of children and adolescents living in rural regions of Northeastern Slovenia. *Environ Int*. 2020;143. doi:10.1016/j.envint.2020.105985

71. Boyd NF, Dite GS, Stone J, et al. Heritability of mammographic density, a risk factor for breast cancer. *N Engl J Med*. 2002;347(12):886-894. doi:10.1056/NEJMoa013390

72. Brand JS, Humphreys K, Thompson DJ, et al. Volumetric mammographic density: Heritability and association with breast cancer susceptibility loci. *J Natl Cancer Inst*. 2014;106(12). doi:10.1093/jnci/dju334

73. Soguel L, Durocher F, Tchernof A, Diorio C. Adiposity, breast density, and breast cancer risk: epidemiological and biological considerations. *Eur J Cancer Prev*. 2017;26(6):511-520. doi:10.1097/CEJ.0000000000000310

74. Hart V, Reeves KW, Sturgeon SR, et al. The effect of change in body mass index on volumetric measures of mammographic density. *Cancer Epidemiol Biomarkers Prev*. 2015;24(11):1724-1730. doi:10.1158/1055-9965.EPI-15-0330

75. Ziembicki S, Zhu J, Tse E, Martin LJ, Minkin S, Boyd NF. The association between alcohol consumption and breast density: a systematic review and meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2017;26(2):170-178. doi:10.1158/1055-9965.EPI-16-0522

76. McTiernan A, Martin CF, Peck JD, et al. Estrogen-plus-progestin use and mammographic density in postmenopausal women: Women's Health Initiative randomized trial. *J Natl Cancer Inst*. 2005;97(18):1366-1376. doi:10.1093/jnci/dji279

77. Cuzick J, Warwick J, Pinney E, Warren RML, Duffy SW. Tamoxifen and breast density in women at increased risk of breast cancer. *J Natl Cancer Inst*. 2004;96(8):621-628. http://www.ncbi.nlm.nih.gov/pubmed/15100340.

78. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2006;15(6):1159-1169. doi:10.1158/1055-9965.EPI-06-0034

79. Masala G, Ambrogetti D, Assedi M, Bendinelli B, Caini S, Palli D. Mammographic breast density and breast cancer risk in a Mediterranean population: a nested case-

control study in the EPIC Florence cohort. *Breast Cancer Res Treat*. 2017;164(2):467-473. doi:10.1007/s10549-017-4274-9

80. Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Res*. 2011;13(6):223. doi:10.1186/bcr2942

81. Byrne C, Ursin G, Martin CF, et al. Mammographic density change with estrogen and progestin therapy and breast cancer risk. *J Natl Cancer Inst*. 2017;109(9). doi:10.1093/jnci/djx001

82. Rice MS, Bertrand KA, VanderWeele TJ, et al. Mammographic density and breast cancer risk: A mediation analysis. *Breast Cancer Res*. 2016;18(1). doi:10.1186/s13058-016-0750-0

83. Rice MS, Tamimi RM, Bertrand KA, et al. Does mammographic density mediate risk factor associations with breast cancer? An analysis by tumor characteristics. *Breast Cancer Res Treat*. 2018;170(1):129-141. doi:10.1007/s10549-018-4735-9

84. Li S, Dugué P-A, Baglietto L, et al. Genome-wide association study of peripheral blood DNA methylation and conventional mammographic density measures. *Int J Cancer*. 2019;145(7):1768-1773. doi:10.1002/ijc.32171

85. Caini S, Fiorito G, Palli D, et al. Pre-diagnostic DNA methylation patterns differ according to mammographic breast density amongst women who subsequently develop breast cancer: a case-only study in the EPIC-Florence cohort. *Breast Cancer Res Treat*. June 2021:1-10. doi:10.1007/s10549-021-06273-w

86. Lichtenstein P, Holm N V, Verkasalo PK, et al. 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. doi:10.1056/NEJM200007133430201

87. Foulkes WD. Inherited susceptibility to common cancers. *N Engl J Med*. 2008;359(20):2143-2153. doi:10.1056/NEJMra0802968

88. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science (80-)*. 2015;347(6217):78-81. doi:10.1126/science.1260825

89. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial risk and heritability of cancer among twins in nordic countries. *J Am Med Assoc*. 2016;315(1):68-76. doi:10.1001/jama.2015.17703

90. Forman MR, Winn DM, Collman GW, Rizzo J, Birnbaum LS. Environmental exposures, breast development and cancer risk: Through the looking glass of breast cancer prevention. *Reprod Toxicol*. 2015;54:6-10. doi:10.1016/j.reprotox.2014.10.019

91. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. *Nat Rev Cardiol*. 2015;12(11):627-642. doi:10.1038/nrcardio.2015.152

92. Béjot Y, Reis J, Giroud M, Feigin V. A review of epidemiological research on stroke

and dementia and exposure to air pollution. *Int J Stroke*. 2018;13(7):687-695. doi:10.1177/1747493018772800

93. De Miguel-Díez J, Hernández-Vázquez J, López-De-Andrés A, Álvaro-Meca A, Hernández-Barrera V, Jiménez-García R. Analysis of environmental risk factors for chronic obstructive pulmonary disease exacerbation: A case-crossover study (2004-2013). *PLoS One*. 2019;14(5). doi:10.1371/journal.pone.0217143

94. Dendup T, Feng X, Clingan S, Astell-Burt T. Environmental risk factors for developing type 2 diabetes mellitus: A systematic review. *Int J Environ Res Public Health*. 2018;15(1). doi:10.3390/ijerph15010078

95. Pizzorno J. Environmental toxins and infertility. *Integr Med*. 2018;17(2):8-11. http://www.ncbi.nlm.nih.gov/pubmed/30962779.

96. Cannon JR, Greenamyre JT. The role of environmental exposures in neurodegeneration and neurodegenerative diseases. *Toxicol Sci*. 2011;124(2):225-250. doi:10.1093/toxsci/kfr239

97. Dietz PM, Homa D, England LJ, et al. Estimates of nondisclosure of cigarette smoking among pregnant and nonpregnant women of reproductive age in the United States. *Am J Epidemiol*. 2011;173(3):355-359. doi:10.1093/aje/kwq381

98. Shipton D, Tappin DM, Vadiveloo T, Crossley JA, Aitken DA, Chalmers J. Reliability of self reported smoking status by pregnant women for estimating smoking prevalence: a retrospective, cross sectional study. *BMJ*. 2009;339. doi:10.1136/bmj.b4347

99. Leenen FAD, Muller CP, Turner JD. DNA methylation: conducting the orchestra from exposure to phenotype? *Clin Epigenetics*. 2016;8(1):1-15. doi:10.1186/s13148-016-0256-8

100. Meehan RR, Thomson JP, Lentini A, Nestor CE, Pennings S. DNA methylation as a genomic marker of exposure to chemical and environmental agents. *Curr Opin Chem Biol*. 2018;45:48-56. doi:10.1016/j.cbpa.2018.02.006

101. Ladd-Acosta C, Fallin & MD. DNA methylation signatures as biomarkers of prior environmental exposures. *Curr Epidemiol Reports*. 2019;6(1):1-13. doi:10.1007/s40471-019-0178-z

102. Gao X, Jia M, Zhang Y, Breitling L, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenetics*. 2015;7(1). doi:10.1186/S13148-015-0148-3

103. Philibert R, Hollenbeck N, Andersen E, et al. A quantitative epigenetic approach for the assessment of cigarette consumption. *Front Psychol*. 2015;6:656. doi:10.3389/FPSYG.2015.00656

104. Plusquin M, Guida F, Polidoro S, et al. DNA methylation and exposure to ambient air pollution in two prospective cohorts. *Environ Int*. 2017;108:127-136. doi:10.1016/j.envint.2017.08.006

105. Chi GC, Liu Y, MacDonald JW, et al. Long-term outdoor air pollution and DNA methylation in circulating monocytes: results from the Multi-Ethnic Study of Atherosclerosis (MESA). *Environ Heal*. 2016;15(1):119. doi:10.1186/s12940-016-0202-4

106. Pavanello S, Bollati V, Pesatori AC, et al. Global and gene-specific promoter methylation changes are related to anti-B[a]PDE-DNA adduct levels and influence micronuclei levels in polycyclic aromatic hydrocarbon-exposed individuals. *Int J Cancer*. 2009;125(7):1692-1697. doi:10.1002/ijc.24492

107. White AJ, Chen J, Teitelbaum SL, et al. Sources of polycyclic aromatic hydrocarbons are associated with gene-specific promoter methylation in women with breast cancer. *Environ Res*. 2016;145:93-100. doi:10.1016/j.envres.2015.11.033

108. McPherson K, Steel CM, Dixon JM. Breast cancer—epidemiology, risk factors, and genetics. *BMJ*. 2000;321(7261).

109. Park HL, Tran SM, Lee J, et al. Clinical implementation of a breast cancer risk assessment program in a multiethnic patient population: which risk model to use? *Breast J*. 2015;21(5):562-564. doi:10.1111/tbj.12461

110. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010

111. Benbow JL, Clarke CA, Park HL, et al. Novel implementation of cloud computing and mobile technology to transform data collection and management in cancer epidemiology. In: American Association for Cancer Research (AACR); 2014:302. doi:10.1158/1538-7445.am2014-302

112. Curl C, Beresford S, Hajat A, et al. Associations of organic produce consumption with socioeconomic status and the local food environment: Multi-Ethnic Study of Atherosclerosis (MESA). *PLoS One*. 2013;8(7). doi:10.1371/JOURNAL.PONE.0069778

113. Baudry J, Méjean C, Allès B, et al. Contribution of organic food to the diet in a large sample of French adults (The NutriNet-Santé cohort study). *Nutrients*. 2015;7(10):8615-8632. doi:10.3390/nu7105417

114. Cheng I, Tseng C, Wu J, et al. Association between ambient air pollution and breast cancer risk: The multiethnic cohort study. *Int J Cancer*. 2020;146(3):699-711. doi:10.1002/ijc.32308

115. US Department of Labor Bureau of Labor Statistics. Standard Occupational Classification (SOC) System. https://www.bls.gov/soc/. Accessed January 13, 2020.

116. US Department of Labor E and TA (Organization/Institution). O*NET, the Occupational Information Network. https://www.onetonline.org/. Accessed January 13, 2020.

117. UCI Community and Labor Project, UCLA Labor Center. *Orange County on the Cusp of*

Change.; 2014.

118. Piercy KL, Troiano RP, Ballard RM, et al. The physical activity guidelines for Americans. *J Am Med Assoc*. 2018;320(19):2020-2028. doi:10.1001/jama.2018.14854

119. Subar AF, Kirkpatrick SI, Mittl B, et al. The Automated Self-Administered 24-Hour dietary recall (ASA24): a resource for researchers, clinicians, and educators from the National Cancer Institute. *J Acad Nutr Diet*. 2012;112(8):1134-1137. doi:10.1016/j.jand.2012.04.016

120. Kirkpatrick SI, Subar AF, Douglass D, et al. Performance of the Automated Self-Administered 24-hour Recall relative to a measure of true intakes and to an interviewer-administered 24-h recall. *Am J Clin Nutr*. 2014;100(1):233. doi:10.3945/AJCN.114.083238

121. National Institutes of Health, Epidemiology and Genomics Research Program NCI (Organization/Institution). Diet History Questionnaire, Version 2.0. 2010. https://epi.grants.cancer.gov/dhq2/. Accessed January 13, 2020.

122. Thompson FE, Subar AF, Brown CC, et al. Cognitive research enhances accuracy of food frequency questionnaire reports: Results of an experimental validation study. *J Am Diet Assoc*. 2002;102(2). doi:10.1016/S0002-8223(02)90050-7

123. Subar AF, Thompson FE, Kipnis V, et al. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. *Am J Epidemiol*. 2001;154(12):1089-1099. doi:10.1093/aje/154.12.1089

124. Subar AF, Kipnis V, Troiano RP, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study. *Am J Epidemiol*. 2003;158(1):1-13. doi:10.1093/aje/kwg092

125. National Cancer Institute E and GRP (Organization/Institution). Diet*Calc Analysis Program, Version 1.4.3. 2005. https://epi.grants.cancer.gov/DHQ/dietcalc/. Accessed January 13, 2020.

126. Rhee J, Sampson L, Cho E, Hughes M, Hu F, Willett W. Comparison of methods to account for implausible reporting of energy intake in epidemiologic studies. *Am J Epidemiol*. 2015;181(4):225-233. doi:10.1093/AJE/KWU308

127. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*. 2016;8(3):389-399. doi:10.2217/epi.15.114

128. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief Bioinform*. 2014;15(6):929-941. doi:10.1093/bib/bbt054

129. U.S. Census Bureau. 2013-2017 American Community Survey 5-Year Estimates. Retrieved from data.census.gov. https://data.census.gov/. Accessed January 13,

2020.

130. Schoenborn CA, Adams PF, Peregoy JA. Health behaviors of adults: United States, 2008-2010. *Vital Heal Stat Ser 10 Data from Natl Heal Surv*. 2013;(257).

131. Green JM. The rise and future of glyphosate and glyphosate-resistant crops. *Pest Manag Sci*. 2016;74(5):1035-1039. doi:10.1002/ps.4462

132. Nandula VK, Riechers DE, Ferhatoglu Y, et al. Herbicide metabolism: crop selectivity, bioactivation, weed resistance, and regulation. *Weed Sci*. 2019;67(2):149-175. doi:10.1017/WSC.2018.88

133. Bai S, Ogbourne S. Glyphosate: environmental contamination, toxicity and potential risks to human health via food contamination. *Environ Sci Pollut Res Int*. 2016;23(19):18988-19001. doi:10.1007/S11356-016-7425-3

134. Grandcoin A, Piel S, Baurès E. AminoMethylPhosphonic acid (AMPA) in natural waters: its sources, behavior and environmental fate. *Water Res*. 2017;117:187-197. doi:10.1016/J.WATRES.2017.03.055

135. Borggaard OK, Gimsing AL. Fate of glyphosate in soil and the possibility of leaching to ground and surface waters: a review. *Pest Manag Sci*. 2008;64(4):441-456. doi:10.1002/PS.1512

136. Reddy KN, Cizdziel J V, Martin M. Williams I, Maul JE, Rimando AM, Duke SO. Glyphosate resistance technology has minimal or no effect on maize mineral content and yield. *J Agric Food Chem*. 2018;66(39):10139-10146. doi:10.1021/ACS.JAFC.8B01655

137. Bøhn T, Cuhra M, Traavik T, Sanden M, Fagan J, Primicerio R. Compositional differences in soybeans on the market: glyphosate accumulates in Roundup Ready GM soybeans. *Food Chem*. 2014;153:207-215. doi:10.1016/J.FOODCHEM.2013.12.054

138. Duke SO. Glyphosate degradation in glyphosate-resistant and -susceptible crops and weeds. *J Agric Food Chem*. 2011;59(11):5835-5841. doi:10.1021/JF102704X

139. Cessna AJ, Darwent AL, Kirkland KJ, Townley-Smith L, Harker KN, Lefkovitch LP. Residues of glyphosate and its metabolite AMPA in wheat seed and foliage following preharvest applications. *Can J Plant Sci*. 2011;74(3):653-661. doi:10.4141/CJPS94-117

140. Myers JP, Antoniou MN, Blumberg B, et al. Concerns over use of glyphosate-based herbicides and risks associated with exposures: A consensus statement. *Environ Heal*. 2016;15(1). doi:10.1186/s12940-016-0117-0

141. Gillezeau C, van Gerwen M, Shaffer RM, et al. The evidence of human exposure to glyphosate: a review. *Environ Heal*. 2019;18(1):2. doi:10.1186/s12940-018-0435-5

142. Vandenberg LN, Blumberg B, Antoniou MN, et al. Is it time to reassess current safety standards for glyphosate-based herbicides? *J Epidemiol Community Health*.

2017;71(6):613-618. doi:10.1136/jech-2016-208463

143. Connolly A, Coggins MA, Koch HM. Human biomonitoring of glyphosate exposures: State-of-the-art and future research challenges. *Toxics*. 2020;8(3). doi:10.3390/TOXICS8030060

144. Lucia RM, Huang WL, Alvarez A, et al. Rationale, study design, and cohort characteristics for the Markers for Environmental Exposures (MEE) study. *Int J Environ Res Public Health*. 2020;17(5):1774. doi:10.3390/ijerph17051774

145. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

146. Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg*. 1990;5(1):46-51. doi:10.1080/1047322X.1990.10389587

147. Krebs-Smith SM, Pannucci TRE, Subar AF, et al. Update of the Healthy Eating Index: HEI-2015. *J Acad Nutr Diet*. 2018;118(9):1591-1602. doi:10.1016/j.jand.2018.05.021

148. Curl CL, Beresford SAA, Fenske RA, et al. Estimating pesticide exposure from dietary intake and organic food choices: The multi-ethnic study of atherosclerosis (MESA). *Environ Health Perspect*. 2014;123(5):475-483. doi:10.1289/ehp.1408197

149. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria; 2018. https://www.r-project.org/.

150. Sun Y, Liu B, Du Y, et al. Inverse association between organic food purchase and diabetes mellitus in US adults. *Nutrients*. 2018;10(12):1877. doi:10.3390/nu10121877

151. Petersen SB, Rasmussen MA, Strøm M, Halldorsson TI, Olsen SF. Sociodemographic characteristics and food habits of organic consumers-a study from the Danish National Birth Cohort. *Public Health Nutr*. 2013;16(10):1810-1819. doi:10.1017/S1368980012004119

152. Baudry J, Allès B, Péneau S, et al. Dietary intakes and diet quality according to levels of organic food consumption by French adults: cross-sectional findings from the NutriNet-Santé Cohort Study. *Public Health Nutr*. 2017;20(4):638-648. doi:10.1017/S1368980016002718

153. Eisinger-Watzl M, Wittig F, Heuer T, Hoffmann I. Customers purchasing organic food - do they live healthier? Results of the German National Nutrition Survey II. *Eur J Nutr Food Saf*. 2015;5(1):59-71. doi:10.9734/EJNFS/2015/12734

154. McGuire MK, McGuire MA, Price WJ, et al. Glyphosate and aminomethylphosphonic acid are not detectable in human milk. *Am J Clin Nutr*. 2016;103(5):1285-1290. doi:10.3945/ajcn.115.126854

155. Van Eerd LL, Hoagland RE, Zablotowicz RM, Hall JC. Pesticide metabolism in plants

and microorganisms. *Weed Sci*. 2003;51(4):472-495. doi:10.1614/0043-1745(2003)051[0472:PMIPAM]2.0.CO;2

156. Zhan H, Feng Y, Fan X, Chen S. Recent advances in glyphosate biodegradation. *Appl Microbiol Biotechnol*. 2018;102(12):5033-5043. doi:10.1007/S00253-018-9035-0

157. Anadón A, Martínez-Larrañaga MR, Martínez MA, et al. Toxicokinetics of glyphosate and its metabolite aminomethyl phosphonic acid in rats. *Toxicol Lett*. 2009;190(1):91-95. doi:10.1016/j.toxlet.2009.07.008

158. Nielsen L, Roager H, Casas M, et al. Glyphosate has limited short-term effects on commensal bacterial community composition in the gut environment due to sufficient aromatic amino acid levels. *Environ Pollut*. 2018;233:364-376. doi:10.1016/J.ENVPOL.2017.10.016

159. Reddy KN, Rimando AM, Duke SO, Nandula VK. Aminomethylphosphonic Acid Accumulation in Plant Species Treated with Glyphosate. *J Agric Food Chem*. 2008;56(6):2125-2130. doi:10.1021/jf072954f

160. USDA (United States Department of Agriculture). *Pesticide Data Program Annual Summary, Calendar Year 2011*.; 2013. https://www.ams.usda.gov/reports/pdp-annual-summary-reports. Accessed August 10, 2021.

161. Tittlemier SA, Bestvater L, Carlson J, Kletke J, Izydorczyk M, Fu BX. Fate of glyphosate in wheat during milling and bread production. *Cereal Chem*. 2021;98(1):100-108. doi:10.1002/CCHE.10369

162. Mesnage R, Antoniou M. Computational modelling provides insight into the effects of glyphosate on the shikimate pathway in the human gut microbiome. *Curr Res Toxicol*. 2020;1:25-33. doi:10.1016/J.CRTOX.2020.04.001

163. Klinder A, Shen Q, Heppel S, Lovegrove J, Rowland I, Tuohy K. Impact of increasing fruit and vegetables and flavonoid intake on the human gut microbiota. *Food Funct*. 2016;7(4):1788-1796. doi:10.1039/C5FO01096A

164. Chaudhari S. Weed management in vineyards starts with fall herbicide application. Michigan State University Extension. https://www.canr.msu.edu/news/weed-management-in-vineyards-starts-with-fall-herbicide-application. Accessed August 11, 2021.

165. Cook K. Bottoms Up: Glyphosate Pesticide in Beer and Wine. https://uspirg.org/sites/pirg/files/reports/WEB_USP_Bottoms-Up_REPORT_022619.pdf. Published 2019. Accessed August 10, 2021.

166. Low FL, Shaw IC, Gerrard JA. The effect of Saccharomyces cerevisiae on the stability of the herbicide glyphosate during bread leavening. *Lett Appl Microbiol*. 2005;40(2):133-137. doi:10.1111/J.1472-765X.2004.01633.X

167. Duke SO. Glyphosate: uses other than in glyphosate-resistant crops, mode of action, degradation in plants, and effects on non-target plants and agricultural microbes. *Rev Environ Contam Toxicol*. 2020;(Online ahead of print). doi:10.1007/398_2020_53

168. Duke SO, Scheffler BE, Dayan FE, Dyer WE. Genetic engineering crops for improved weed management traits. *ACS Symp Ser*. 2002;829:52-66. doi:10.1021/bk-2002-0829.ch006

169. Guyton KZ, Loomis D, Grosse Y, et al. Carcinogenicity of tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate. *Lancet Oncol*. 2015;16(5):490-491. doi:10.1016/S1470-2045(15)70134-8

170. Andreotti G, Koutros S, Hofmann JN, et al. Glyphosate use and cancer incidence in the Agricultural Health Study. *J Natl Cancer Inst*. 2018;110(5):509-516. doi:10.1093/jnci/djx233

171. Rappazzo KM, Warren JL, Davalos AD, et al. Maternal residential exposure to specific agricultural pesticide active ingredients and birth defects in a 2003–2005 North Carolina birth cohort. *Birth Defects Res*. 2019;111(6):312-323. doi:10.1002/bdr2.1448

172. Piccoli C, Cremonese C, Koifman RJ, Koifman S, Freire C. Pesticide exposure and thyroid function in an agricultural population in Brazil. *Environ Res*. 2016;151:389-398. doi:10.1016/j.envres.2016.08.011

173. Muñoz JP, Bleak TC, Calaf GM. Glyphosate and the key characteristics of an endocrine disruptor: A review. *Chemosphere*. 2021;270. doi:10.1016/j.chemosphere.2020.128619

174. Dechartres J, Pawluski JL, Gueguen MM, et al. Glyphosate and glyphosate-based herbicide exposure during the peripartum period affects maternal brain plasticity, maternal behaviour and microbiome. *J Neuroendocrinol*. 2019;31(9). doi:10.1111/jne.12731

175. Kurenbach B, Marjoshi D, Amábile-Cuevas CF, et al. Sublethal exposure to commercial formulations of the herbicides dicamba, 2,4-dichlorophenoxyacetic acid, and Glyphosate cause changes in antibiotic susceptibility in Escherichia coli and Salmonella enterica serovar Typhimurium. *MBio*. 2015;6(2). doi:10.1128/mBio.00009-15

176. Martin EM, Fry RC. Environmental influences on the epigenome: exposure-associated DNA methylation in human populations. *Annu Rev Public Health*. 2018;39(1):309-333. doi:10.1146/annurev-publhealth-040617-014629

177. Park HL. Epigenetic biomarkers for environmental exposures and personalized breast cancer prevention. *Int J Environ Res Public Health*. 2020;17(4):1181. doi:10.3390/ijerph17041181

178. Jordahl KM, Phipps AI, Randolph TW, et al. Differential DNA methylation in blood as a mediator of the association between cigarette smoking and bladder cancer risk among postmenopausal women. *Epigenetics*. 2019;14(11):1065-1073. doi:10.1080/15592294.2019.1631112

179. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and

innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res*. 2017;45(4):e22. doi:10.1093/nar/gkw967

180. Nordlund J, Bäcklin CL, Wahlberg P, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol*. 2013;14(9):1-15. doi:10.1186/gb-2013-14-9-r105

181. Fortin J-P, Triche TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 2016;33(4):558-560. doi:10.1093/bioinformatics/btw691

182. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013;29(2):189-196. doi:10.1093/bioinformatics/bts680

183. Tian Y, Morris TJ, Webster AP, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*. 2017;33(24):3982-3984. doi:10.1093/bioinformatics/btx513

184. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037

185. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-883. doi:10.1093/bioinformatics/bts034

186. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13(1):86. doi:10.1186/1471-2105-13-86

187. Du P, Zhang X, Huang C-C, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(1):587. doi:10.1186/1471-2105-11-587

188. Lucia RM, Huang W-L, Alvarez A, et al. Association of mammographic density with blood DNA methylation. *Epigenetics*. June 2021:1-16. doi:10.1080/15592294.2021.1928994

189. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin*. 2015;8(1):6. doi:10.1186/1756-8935-8-6

190. He Z, Yu W. Stable feature selection for biomarker discovery. *Comput Biol Chem*. 2010;34(4):215-225. doi:10.1016/j.compbiolchem.2010.07.002

191. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet*. 2005;365(9458):488-492. doi:10.1016/S0140-6736(05)17866-0

192. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust

gene list for predicting outcome in cancer. *Proc Natl Acad Sci*. 2006;103(15):5923-5928. doi:10.1073/pnas.0601231103

193. Qiu X, Xiao Y, Gordon A, Yakovlev A. Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*. 2006;7. doi:10.1186/1471-2105-7-50

194. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*. 2017;12(12):2478-2492. doi:10.1038/nprot.2017.124

195. Phipson B, Maksimovic J, Oshlack A. MissMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*. 2016;32(2):286-288. doi:10.1093/bioinformatics/btv560

196. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011;39(5):1-13. doi:10.18637/jss.v039.i05

197. Haslam E. *Shikimic Acid: Metabolism and Metabolites*. Wiley; 1993. https://books.google.com/books/about/Shikimic_Acid.html?id=Y3wvAQAAIAAJ. Accessed May 14, 2021.

198. Kishore GM, Shah DM. Amino acid biosynthesis inhibitors as herbicides. *Annu Rev Biochem*. 1988;57:627-663. doi:10.1146/annurev.bi.57.070188.003211

199. Cerdeira AL, Duke SO. The current status and environmental impacts of glyphosate-resistant crops. *J Environ Qual*. 2006;35(5):1633-1658. doi:10.2134/jeq2005.0378

200. Blot N, Veillat L, Rouzé R, Delatte H. Glyphosate, but not its metabolite AMPA, alters the honeybee gut microbiota. *PLoS One*. 2019;14(4). doi:10.1371/journal.pone.0215466

201. Motta EVS, Raymann K, Moran NA. Glyphosate perturbs the gut microbiota of honey bees. *Proc Natl Acad Sci*. 2018;115(41):10305-10310. doi:10.1073/pnas.1803880115

202. Ruuskanen S, Rainio MJ, Gómez-Gallego C, et al. Glyphosate-based herbicides influence antioxidants, reproductive hormones and gut microbiome but not reproduction: A long-term experiment in an avian model. *Environ Pollut*. 2020;266:115108. doi:10.1016/j.envpol.2020.115108

203. Lynch S V, Pedersen O. The Human intestinal microbiome in health and disease. *N Engl J Med*. 2016;375(24):2369-2379. doi:10.1056/nejmra1600266

204. Xiao C, Fedirko V, Beitler J, et al. The role of the gut microbiome in cancer-related fatigue: pilot study on epigenetic mechanisms. *Support Care Cancer*. 2020;29(6). doi:10.1007/s00520-020-05820-3

205. Sobhani I, Bergsten E, Couffin S, et al. Colorectal cancer-associated microbiota contributes to oncogenic epigenetic signatures. *Proc Natl Acad Sci*. 2019;116(48):24285-24295. doi:10.1073/pnas.1912129116

206. Curtis SW, Cobb DO, Kilaru V, et al. Environmental exposure to polybrominated

biphenyl (PBB) associates with an increased rate of biological aging. *Aging (Albany NY)*. 2019;11(15):5498-5517. doi:10.18632/aging.102134

207. Gensous N, Garagnani P, Santoro A, et al. One-year Mediterranean diet promotes epigenetic rejuvenation with country- and sex-specific effects: a pilot study from the NU-AGE project. *GeroScience*. 2020;42(2):687-701. doi:10.1007/s11357-019-00149-0

208. Lind PM, Salihovic S, Lind L. High plasma organochlorine pesticide levels are related to increased biological age as calculated by DNA methylation analysis. *Environ Int*. 2018;113:109-113. doi:10.1016/j.envint.2018.01.019

209. White AJ, Kresovich JK, Keller JP, et al. Air pollution, particulate matter composition and methylation-based biologic age. *Environ Int*. 2019;132:105071. doi:10.1016/j.envint.2019.105071

210. Kresovich JK, Xu Z, O'Brien KM, Weinberg CR, Sandler DP, Taylor JA. Methylation-based biological age and breast cancer risk. *J Natl Cancer Inst*. 2019;111(10):1051-1058. doi:10.1093/jnci/djz020

211. Dugué PA, Bassett JK, Joo JHE, et al. DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *Int J Cancer*. 2018;142(8):1611-1619. doi:10.1002/ijc.31189

212. Levine ME, Hosgood HD, Chen B, Absher D, Assimes T, Horvath S. DNA methylation age of blood predicts future onset of lung cancer in the Women's Health Initiative. *Aging (Albany NY)*. 2015;7(9):690-700. doi:10.18632/aging.100809

213. Quach A, Levine ME, Tanaka T, et al. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY)*. 2017;9(2):419-446. doi:10.18632/aging.101168

214. McCartney DL, Hillary RF, Stevenson AJ, et al. Epigenetic prediction of complex traits and death. *Genome Biol*. 2018;19(1):136. doi:10.1186/s13059-018-1514-1

215. Nannini DR, Joyce BT, Zheng Y, et al. Epigenetic age acceleration and metabolic syndrome in the coronary artery risk development in young adults study. *Clin Epigenetics*. 2019;11(1):1-9. doi:10.1186/s13148-019-0767-1

216. Perna L, Zhang Y, Mons U, Holleczek B, Saum KU, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin Epigenetics*. 2016;8(1):1-7. doi:10.1186/s13148-016-0228-z

217. Marioni RE, Shah S, McRae AF, et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol*. 2015;16(1):1-12. doi:10.1186/s13059-015-0584-6

218. Chen BH, Marioni RE, Colicino E, et al. DNA methylation-based measures of biological age: Meta-analysis predicting time to death. *Aging (Albany NY)*. 2016;8(9):1844-1865. doi:10.18632/aging.101020

219. Prados-Carvajal R, López-Saavedra A, Cepeda-García C, Jimeno S, Huertas P. Multiple roles of the splicing complex SF3B in DNA end resection and homologous recombination. *DNA Repair (Amst)*. 2018;66-67:11-23. doi:10.1016/j.dnarep.2018.04.003

220. Kawamura N, Nimura K, Saga K, et al. SF3B2-mediated RNA splicing drives human prostate cancer progression. *Cancer Res*. 2019;79(20):5204-5217. doi:10.1158/0008-5472.CAN-18-3965

221. Xu W, Huang H, Yu L, Cao L. Meta-analysis of gene expression profiles indicates genes in spliceosome pathway are up-regulated in hepatocellular carcinoma (HCC). *Med Oncol*. 2015;32(4):1-9. doi:10.1007/s12032-014-0425-6

222. Sugimoto T, Seki N, Shimizu S, et al. The galanin signaling cascade is a candidate pathway regulating oncogenesis in human squamous cell carcinoma. *Genes Chromosom Cancer*. 2009;48(2):132-142. doi:10.1002/gcc.20626

223. Her C, Zhao N, Wu X, Tompkins JD. MutS homologues hMSH4 and hMSH5: Diverse functional implications in humans. *Front Biosci*. 2007;12(3):905-911. doi:10.2741/2112

224. Kim YH, Ohta T, Oh JE, et al. TP53, MSH4, and LATS1 germline mutations in a family with clustering of nervous system tumors. *Am J Pathol*. 2014;184(9):2374-2381. doi:10.1016/j.ajpath.2014.05.017

225. Moslehi R, Tsao HS, Zeinomar N, Stagnar C, Fitzpatrick S, Dzutsev A. Integrative genomic analysis implicates ERCC6 and its interaction with ERCC8 in susceptibility to breast cancer. *Sci Rep*. 2020;10(1). doi:10.1038/s41598-020-77037-7

226. Wu Z, Zhu X, Yu Q, Xu Y, Wang Y. Multisystem analyses of two Cockayne syndrome associated proteins CSA and CSB reveal shared and unique functions. *DNA Repair (Amst)*. 2019;83. doi:10.1016/j.dnarep.2019.102696

227. Pascucci B, Fragale A, Marabitti V, et al. CSA and CSB play a role in the response to DNA breaks. *Oncotarget*. 2018;9(14):11581-11591. doi:10.18632/oncotarget.24342

228. Xiao Y, Deng T, Ming X, Xu J. TRIM31 promotes acute myeloid leukemia progression and sensitivity to daunorubicin through the Wnt/β-catenin signaling. *Biosci Rep*. 2020;40(4). doi:10.1042/BSR20194334

229. Shi G, Lv C, Yang Z, et al. TRIM31 promotes proliferation, invasion and migration of glioma cells through akt signaling pathway. *Neoplasma*. 2019;66(5):727-735. doi:10.4149/neo_2019_190106N21

230. Guo P, Ma X, Zhao W, et al. TRIM31 is upregulated in hepatocellular carcinoma and promotes disease progression by inducing ubiquitination of TSC1-TSC2 complex. *Oncogene*. 2018;37(4):478-488. doi:10.1038/onc.2017.349

231. Connolly A, Leahy M, Jones K, Kenny L, Coggins MA. Glyphosate in Irish adults - A pilot study in 2017. *Environ Res*. 2018;165:235-236. doi:10.1016/j.envres.2018.04.025

232. Conrad A, Schröter-Kermani C, Hoppe HW, Rüther M, Pieper S, Kolossa-Gehring M. Glyphosate in German adults – Time trend (2001 to 2015) of human exposure to a widely used herbicide. *Int J Hyg Environ Health*. 2017;220(1):8-16. doi:10.1016/j.ijheh.2016.09.016

233. Curwin BD, Hein MJ, Sanderson WT, et al. Urinary pesticide concentrations among children, mothers and fathers living in farm and non-farm households in Iowa. *Ann Occup Hyg*. 2007;51(1):53-65. doi:10.1093/annhyg/mel062

234. Noone AM, Howlader N, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ CK (eds). *SEER Cancer Statistics Review, 1975-2015*. Bethesda, MD; 2018. https://seer.cancer.gov/csr/1975_2015/. Accessed February 24, 2019.

235. Johansson A, Palli D, Masala G, et al. Epigenome-wide association study for lifetime estrogen exposure identifies an epigenetic signature associated with breast cancer risk. *Clin Epigenetics*. 2019;11(1):66. doi:10.1186/s13148-019-0664-7

236. Wilson LE, Harlid S, Xu Z, Sandler DP, Taylor JA. An epigenome-wide study of body mass index and DNA methylation in blood using participants from the Sister Study cohort. *Int J Obes*. 2017;41(1):194-199. doi:10.1038/ijo.2016.184

237. Philibert R, Dogan M, Noel A, et al. Genome-wide and digital polymerase chain reaction epigenetic assessments of alcohol consumption. *Am J Med Genet Part B Neuropsychiatr Genet*. 2018;177(5):479-488. doi:10.1002/ajmg.b.32636

238. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA et al. *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA; 2013.

239. Liu J, Siegmund KD. An evaluation of processing methods for HumanMethylation450 BeadChip data. *BMC Genomics*. 2016;17(1):469. doi:10.1186/s12864-016-2819-7

240. Gonseth S, Roy R, Houseman EA, et al. Periconceptional folate consumption is associated with neonatal DNA methylation modifications in neural crest regulatory and cancer development genes. *Epigenetics*. 2015;10(12):1166-1176. doi:10.1080/15592294.2015.1117889

241. Connell SP, O'Reilly E, Tuzova A, et al. Development of a multivariable risk model integrating urinary cell DNA methylation and cell-free RNA data for the detection of significant prostate cancer. *Prostate*. 2020;80(7):547-558. doi:10.1002/pros.23968

242. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473(7345):43-49. doi:10.1038/nature09906

243. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211

244. Sherman ME, Figueroa JD, Henry JE, Clare SE, Rufenbarger C, Storniolo AM. The Susan G. Komen for the Cure Tissue Bank at the IU Simon Cancer Center: A unique

resource for defining the “molecular histology” of the breast. *Cancer Prev Res*. 2012;5(4):528-535. doi:10.1158/1940-6207.CAPR-11-0234

245. Sehl ME, Henry JE, Storniolo AM, Ganz PA, Horvath S. DNA methylation age is elevated in breast tissue of healthy women. *Breast Cancer Res Treat*. 2017;164(1):209-219. doi:10.1007/s10549-017-4218-4

246. Ambatipudi S, Cuenin C, Hernandez-Vargas H, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics*. 2016;8(5):599-618. doi:10.2217/epi-2016-0001

247. Joehanes R, Just AC, Marioni RE, et al. Epigenetic signatures of cigarette smoking. *Circ Cardiovasc Genet*. 2016;9(5):436-447. doi:10.1161/CIRCGENETICS.116.001506

248. Hong Y, Choi H-M, Cheong HS, Shin HD, Choi CM, Kim WJ. Epigenome-wide association analysis of differentially methylated signals in blood samples of patients with non-small-cell lung cancer. *J Clin Med*. 2019;8(9):1307. doi:10.3390/jcm8091307

249. Dugué PA, Jung CH, Joo JE, et al. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. *Epigenetics*. 2020;15(4):358-368. doi:10.1080/15592294.2019.1668739

250. Kulis M, Esteller M. *DNA Methylation and Cancer*. Vol 70. Academic Press; 2010. doi:10.1016/B978-0-12-380866-0.60002-2

251. Xu Z, Sandler DP, Taylor JA. Blood DNA methylation and breast cancer: a prospective case-cohort analysis in the Sister Study. *J Natl Cancer Inst*. 2020;112(1):87-94. doi:10.1093/jnci/djz065

252. Tang Q, Cheng J, Cao X, Surowy H, Burwinkel B. Blood-based DNA methylation as biomarker for breast cancer: a systematic review. *Clin Epigenetics*. 2016;8(1):115. doi:10.1186/s13148-016-0282-6

253. Bodelon C, Ambatipudi S, Dugué PA, et al. Blood DNA methylation and breast cancer risk: A meta-analysis of four prospective cohort studies. *Breast Cancer Res*. 2019;21(1). doi:10.1186/s13058-019-1145-9

254. Chen M, Wong EM, Nguyen TL, et al. DNA methylation-based biological age, genome-wide average DNA methylation, and conventional breast cancer risk factors. *Sci Rep*. 2019;9(1). doi:10.1038/s41598-019-51475-4

255. Woo HH, Lee SC, Stoffer JB, Rush D, Chambers SK. Phenotype of vigilin expressing breast cancer cells binding to the 69 nt 3′UTR element in CSF-1R mRNA. *Transl Oncol*. 2019;12(1):106-115. doi:10.1016/j.tranon.2018.09.012

256. Molyneux SD, Waterhouse PD, Shelton D, et al. Human somatic cell mutagenesis creates genetically tractable sarcomas. *Nat Genet*. 2014;46(9):964-972. doi:10.1038/ng.3065

257. Woo H-H, Yi X, Lamb T, et al. Posttranscriptional suppression of proto-oncogene c-

fms expression by vigilin in breast cancer. *Mol Cell Biol*. 2011;31(1):215-225. doi:10.1128/mcb.01031-10

258. Kubiczkova L, Sedlarikova L, Hajek R, Sevcikova S. TGF-β - an excellent servant but a bad master. *J Transl Med*. 2012;10(1):1-24. doi:10.1186/1479-5876-10-183

259. Donovan J, Slingerland J. Transforming growth factor-β and breast cancer: Cell cycle arrest by transforming growth factor-β and its disruption in cancer. *Breast Cancer Res*. 2000;2(2):116-124. doi:10.1186/bcr43

260. Suriyamurthy S, Baker D, Ten Dijke P, Iyengar PV. Epigenetic reprogramming of TGF-β signaling in breast cancer. *Cancers (Basel)*. 2019;11(5):726. doi:10.3390/cancers11050726

261. Hinshelwood RA, Huschtscha LI, Melki J, et al. Concordant epigenetic silencing of transforming growth factor-β signaling pathway genes occurs early in breast carcinogenesis. *Cancer Res*. 2007;67(24):11517-11527. doi:10.1158/0008-5472.CAN-07-1284

262. Huang BZ, Binder AM, Sugar CA, Chao CR, Setiawan VW, Zhang ZF. Methylation of immune-regulatory cytokine genes and pancreatic cancer outcomes. *Epigenomics*. 2020;12(15):1273-1285. doi:10.2217/epi-2019-0335

263. Liu L, Kron KJ, Pethe V V, et al. Association of tissue promoter methylation levels of APC, TGFβ2, HOXD3 and RASSF1A with prostate cancer progression. *Int J Cancer*. 2011;129(10):2454-2462. doi:10.1002/ijc.25908

264. Roh SH, Kasembeli M, Bakthavatsalam D, Chiu W, Tweardy DJ. Contribution of the type II chaperonin, TRiC/CCT, to oncogenesis. *Int J Mol Sci*. 2015;16(11):26706-26720. doi:10.3390/ijms161125975

265. Klimczak M, Biecek P, Zylicz A, Zylicz M. Heat shock proteins create a signature to predict the clinical outcome in breast cancer. *Sci Rep*. 2019;9(1). doi:10.1038/s41598-019-43556-1

266. Showalter AE, Martini AC, Nierenberg D, et al. Investigating Chaperonin-Containing TCP-1 subunit 2 as an essential component of the chaperonin complex for tumorigenesis. *Sci Rep*. 2020;10(1). doi:10.1038/s41598-020-57602-w

267. Qu H, Zhu F, Dong H, Hu X, Han M. Upregulation of CCT-3 induces breast cancer cell proliferation through miR-223 competition and Wnt/β-catenin signaling pathway activation. *Front Oncol*. 2020;10. doi:10.3389/fonc.2020.533176

268. Guest ST, Kratche ZR, Bollig-Fischer A, Haddad R, Ethier SP. Two members of the TRiC chaperonin complex, CCT2 and TCP1 are essential for survival of breast cancer cells and are linked to driving oncogenes. *Exp Cell Res*. 2015;332(2):223-235. doi:10.1016/j.yexcr.2015.02.005

269. Laury AR, Perets R, Piao H, et al. A comprehensive analysis of PAX8 expression in human epithelial tumors. *Am J Surg Pathol*. 2011;35(6):816-826. doi:10.1097/PAS.0b013e318216c112

270. Kilgore MR, Bosch DE, Adamson KH, Swanson PE, Dintzis SM, Rendi MH. Unexpected PAX8 immunoreactivity in metastatic high-grade breast cancer. *Appl Immunohistochem Mol Morphol*. 2019;27(9):637-643. doi:10.1097/PAI.0000000000000707

271. Muratovska A, Zhou C, He S, Goodyer P, Eccles MR. Paired-Box genes are frequently expressed in cancer and often required for cancer cell survival. *Oncogene*. 2003;22(39):7989-7997. doi:10.1038/sj.onc.1206766

272. Yu X, Cao Y, Tang L, Yang Y, Chen F, Xia J. Baicalein inhibits breast cancer growth via activating a novel isoform of the long noncoding RNA PAX8-AS1-N. *J Cell Biochem*. 2018;119(8):6842-6856. doi:10.1002/jcb.26881

273. Moelans CB, Verschuur-Maes AH, van Diest PJ. Frequent promoter hypermethylation of BRCA2, CDH13, MSH6, PAX5, PAX6 and WT1 in ductal carcinoma in situ and invasive breast cancer. *J Pathol*. 2011;225(2):222-231. doi:10.1002/path.2930

274. Zong X, Yang H, Yu Y, et al. Possible role of Pax-6 in promoting breast cancer cell proliferation and tumorigenesis. *BMB Rep*. 2011;44(9):595-600. http://www.bmbreports.org/journal/view.html?spage=595&volume=44&number=9.

275. Ruytinx P, Proost P, Struyf S. CXCL4 and CXCL4L1 in cancer. *Cytokine*. 2018;109:65-71. doi:10.1016/j.cyto.2018.02.022

276. Maurer AM, Zhou B, Han ZC. Roles of platelet factor 4 in hematopoiesis and angiogenesis. *Growth Factors*. 2006;24(4):242-252. doi:10.1080/08977190600988225

277. Solassol J, Rouanet P, Lamy PJ, et al. Serum protein signature may improve detection of ductal carcinoma in situ of the breast. *Oncogene*. 2010;29(4):550-560. doi:10.1038/onc.2009.341

278. Suk HC, Ng MHL, Kin ML, et al. 4q loss is potentially an important genetic event in MM tumorigenesis: Identification of a tumor suppressor gene regulated by promoter methylation at 4q13.3, platelet factor 4. *Blood*. 2007;109(5):2089-2099. doi:10.1182/blood-2006-04-018770

279. Michels KB, Binder AM, Dedeurwaerder S, et al. Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods*. 2013;10(10):949-955. doi:10.1038/nmeth.2632

280. Yang Y, Wu L, Shu XO, et al. Genetically predicted levels of DNA methylation biomarkers and breast cancer risk: data from 228 951 women of European descent. *J Natl Cancer Inst*. 2020;112(3):295-304. doi:10.1093/jnci/djz109

281. Bermejo JL, Huang G, Manoochehri M, et al. Long intergenic noncoding RNA 299 methylation in peripheral blood is a biomarker for triple-negative breast cancer. *Epigenomics*. 2019;11(1):81-93. doi:10.2217/epi-2018-0121

282. Winkel RR, von Euler-Chelpin M, Nielsen M, et al. Inter-observer agreement

according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer. *BMC Cancer*. 2015;15(1):274. doi:10.1186/s12885-015-1256-3

283. van der Waal D, den Heeten GJ, Pijnappel RM, et al. Comparing visually assessed BI-RADS breast density and automated volumetric breast density software: a cross-sectional study in a breast cancer screening setting. *PLoS One*. 2015;10(9):e0136667. doi:10.1371/journal.pone.0136667

284. Alikhassi A, Esmaili Gourabi H, Baikpour M. Comparison of inter- and intra-observer variability of breast density assessments using the fourth and fifth editions of Breast Imaging Reporting and Data System. *Eur J Radiol Open*. 2018;5:67-72. doi:10.1016/j.ejro.2018.04.002

285. Gard CC, Aiello Bowles EJ, Miglioretti DL, Taplin SH, Rutter CM. Misclassification of Breast Imaging Reporting and Data System (BI-RADS) mammographic density and implications for breast density reporting legislation. *Breast J*. 2015;21(5):481-489. doi:10.1111/tbj.12443

286. Nicholson BT, LoRusso AP, Smolkin M, Bovbjerg VE, Petroni GR, Harvey JA. Accuracy of assigned BI-RADS breast density category definitions. *Acad Radiol*. 2006;13(9):1143-1149. doi:10.1016/j.acra.2006.06.005

Appendix A: Supplemental Tables

Table A1. Full text of study questionnaires

ID	Question	Response Choices
Environmental Exposures Questionnaire		
1	Do you eat organic food? ["Organic" food is food that is EITHER labeled "USDA Organic", purchased locally from an "organic farm", grown without pesticides in a home garden, or raised on organic feed without hormones and without antibiotics.]	Seldom or never Sometimes Often or always Don't know or not sure
2	If you eat fresh fruit or drink fruit juice, how often is that fruit or fruit juice organic?	I do not eat the food Seldom or never Sometimes Often or always Don't know or not sure
3	How long have you been eating/drinking organic fruit/fruit juice?	More than 10 years 5-10 years 1-5 years Less than 1 year
4	If you eat fresh vegetables or drink vegetable juice, how often is that vegetable or vegetable juice organic?	I do not eat the food Seldom or never Sometimes Often or always Don't know or not sure
5	How long have you been eating/drinking organic vegetables or vegetable juice?	More than 10 years 5-10 years 1-5 years Less than 1 year
6	If you eat grains (for example, wheat, rice), how often are those grains organic?	I do not eat the food Seldom or never Sometimes Often or always Don't know or not sure
7	How long have you been eating organic grains?	More than 10 years 5-10 years 1-5 years Less than 1 year
8	If you eat meat, how often is that meat organic?	I do not eat the food Seldom or never Sometimes Often or always Don't know or not sure
9	How long have you been eating organic meat?	More than 10 years 5-10 years 1-5 years Less than 1 year

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
10	If you eat eggs, how often are those eggs organic?	I do not eat the food Seldom or never Sometimes Often or always Don't know or not sure
11	How long have you been eating organic eggs?	More than 10 years 5-10 years 1-5 years Less than 1 year
12	If you eat dairy products (for example, milk, cheese), how often are those dairy products organic?	I do not eat the food Seldom or never Sometimes Often or always Don't know or not sure
13	How long have you been eating organic dairy products?	More than 10 years 5-10 years 1-5 years Less than 1 year
14	How often do you eat a meal or snack from a place such as McDonald's, In-N-Out, Subway, Burger King, Chik Fil-A, Wendy's, Arby's, Taco Bell, Pizza Hut, or Kentucky Fried Chicken?	Never 1-5 times/year 6-11 times/year 1-3 times/month 1 time/week 2-3 times/week 4-6 times/week 1 time/day 2 or more times/day
15	Other than the types of eating establishments in the previous question, how often do you eat a meal or snack from other eating establishments?	Never 1-5 times/year 6-11 times/year 1-3 times/month 1 time/week 2-3 times/week 4-6 times/week 1 time/day 2 or more times/day
16	How often do you eat a meal or snack that has been prepared at a deli or grocery store?	Never 1-5 times/year 6-11 times/year 1-3 times/month 1 time/week 2-3 times/week 4-6 times/week 1 time/day 2 or more times/day

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
17	How often do you eat meals that have been home-cooked from basic, simple ingredients (excluding pre-made meals)?	1-5 times/year 6-11 times/year 1-3 times/month 1 time/week 2-3 times/week 4-6 times/week 1 time/day 2 or more times/day
18	Have you ever lived on a farm?	Yes No
19	For how many total years did you live on a farm?	Less than 1 year 1-5 years 6-10 years More than 10 years
20	How old were you when you STARTED living on a farm?	Numeric
21	What is your primary source of drinking water?	Tap water Bottled water Other (please specify) Don't know or not sure
22	(if "other") Please indicate your primary source of drinking water.	Free text
23	Is the tap water you use for drinking or cooking filtered or treated in the home to remove chemicals and minerals? (This includes filtering pitchers like Brita or PUR and filters in your refrigerator or faucet but does NOT include water softeners)	Yes No Don't know or not sure
24	Please list the ZIP code OR city and state OR country of the CURRENT place you live.	Free text
25	How many years have you lived at your CURRENT residence?	Numeric
26	Have you lived at a PREVIOUS place (for at least one year)?	Yes No
27	Please list the ZIP code OR city and state OR country of the PREVIOUS place you lived.	Free text
28	How many years did you live at the past place listed directly above?	Numeric
29	Have you lived at another PREVIOUS place (for at least one year)?	Yes No
30	Please list the ZIP code OR city and state OR country of the PREVIOUS place you lived.	Free text
31	How many years did you live at the past place listed directly above?	Numeric
32	Have you lived at another PREVIOUS place (for at least one year)?	Yes No

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
33	Please list the ZIP code OR city and state OR country of the PREVIOUS place you lived.	Free text
34	How many years did you live at the past place listed directly above?	Numeric
35	Have you lived at another PREVIOUS place (for at least one year)?	Yes No
36	Please list the ZIP code OR city and state OR country of the PREVIOUS place you lived.	Free text
37	How many years did you live at the past place listed directly above?	Numeric
38	What is your current job status? (Mark the one that best describes you. If more than one describes you, mark both.)	Not working Retired Homemaker, raising children/grandchildren, care of others Employed (full-time or part-time) Disabled, unable to work Other (please specify)
39	(if "other") Please describe.	Free text
40	What is/was your occupation?	Free text
41	In the past 7 days, were any chemical products used in your home to control fleas, roaches, ants, termites, or other insects?	Yes No I don't know
42	What kind(s) of insecticides did you use (name of product or take a picture)?	Free text
43	In the past 7 days, were any chemical products used in your lawn or garden to kill weeds?	Yes No I don't know
44	What kind(s) of weedkiller(s) did you use (name of product or take a picture)?	Free text
45	In the past 7 days, did you use any chemical products to control fleas, roaches, ants, termites, or other insects, or weeds, AT WORK?	Yes No I don't know
46	Do you play golf?	Never Occasionally Once per week More than once per week
47	Have you ever smoked regularly for 6 months or more?	Yes No Don't know
48	Now, do you smoke cigarettes every day, some days, or not at all?	Every day Some days Not at all
49	How old were you when you last smoked fairly regularly?	Numeric

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
50	During the years you smoked, how many cigarettes did you usually smoke per day?	Numeric
51	During the past 30 days, on the days that you did smoke, about how many cigarettes did you usually smoke per day?	Numeric
52	Do you have the tendency to flush (get red or pink) in the face immediately after drinking a glass of beer (~6 fluid ounces) or 2/3 glass of wine or a shot of hard liquor?	Yes No Don't know
53	During the first to second year after you started drinking alcohol, did you have a tendency to flush in the face immediately after drinking a glass of beer (~6 fluid ounces) or 2/3 glass of wine or a shot of hard liquor?	Yes No Don't know
54	Have you ever taken any of the following hormone therapies (female hormones prescribed for women after menopause)? (Please check all that apply)	Estrogen only (e.g., Premarin) Progestin only (for example, Provera) Estrogen and progestin combination (for example, Prempro) Estrogen and testosterone combination (for example, Estratest) Natural hormone therapy (Herbal supplements) Topical vaginal estrogen cream Estrogen patch Other hormone therapy Yes - But not sure what kind of hormone replacement therapy No - I have never been on any hormone therapy
55	Are you CURRENTLY taking hormone replacement (female hormones prescribed for women after menopause)? (Please check all that apply)	Estrogen only (e.g., Premarin) Progestin only (for example, Provera) Estrogen and progestin combination (for example, Prempro) Estrogen and testosterone combination (for example, Estratest) Natural hormone therapy (Herbal supplements) Topical vaginal estrogen cream Estrogen patch Other hormone therapy Yes - But not sure what kind of hormone replacement therapy No - I am not currently on any hormone therapy

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
56	For how many total years were you or have you been on hormone replacement therapy?	Less than 5 years For five years or more Don't know
57	Are you CURRENTLY taking any of the following medications? (Please check all that apply)	Tamoxifen (Nolvadex) Raloxifene (Evista) Anastrozole (Arimidex) Hormones for birth control None
58	How many years have you been taking Tamoxifen (Nolvadex)?	Numeric
59	How many years have you been taking Raloxifene (Evista)?	Numeric
60	How many years have you been taking Anastrozole (Arimidex)?	Numeric
61	How many years have you been taking hormones for birth control?	Numeric
62	What is your current weight?	Numeric
Breast Health Questionnaire		
1	What is your gender?	Female Male
2	Are you currently pregnant or possibly pregnant?	No Yes Don't know
3	Are you currently breast feeding?	No Yes
4	When was your last mammogram?	Less than 1 year ago 1 to 2 years ago 2 to 3 years ago 4 or more years ago I've never had a mammogram Don't know
5	Please list medical facility, city and state where your last mammogram was done:	Free text
6	Have you had a clinical breast exam within the last 3 months (done by a doctor or health care provider)?	Yes No Don't know
7	Did your doctor or other health care provider discover a new or unusual lump?	Yes No Don't know
8	Have you noticed any of the following changes in your RIGHT breast in the LAST 3 MONTHS? [Choose all that apply]	Lump (new or unusual) Nipple discharge Pain Other No, I haven't noticed any changes

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
9	Have you noticed any of the following changes in your LEFT breast in the LAST 3 MONTHS? [Choose all that apply]	Lump (new or unusual) Nipple discharge Pain Other No, I haven't noticed any changes
10	Have you noticed any of the following changes in your RIGHT breast that are PRESENT TODAY? [Choose all that apply]	Lump (new or unusual) Nipple discharge Pain Other No, I haven't noticed any changes
11	Have you noticed any of the following changes in your LEFT breast that are PRESENT TODAY? [Choose all that apply]	Lump (new or unusual) Nipple discharge Pain Other No, I haven't noticed any changes
12	The reason for your scheduled mammogram is:	Routine checkup Other
13	Has a doctor ever told you that you have breast cancer or DCIS (ductal carcinoma in situ)? [Choose all that apply]	Yes - invasive breast cancer Yes - ductal carcinoma in situ (DCIS) Yes - but I don't know what kind of breast cancer No Don't know
14	In which breast were you diagnosed with DCIS (ductal carcinoma in situ)?	Left Right Both Don't know
15	How old were you when you were first diagnosed with DCIS (ductal carcinoma in situ)?	Numeric
16	How old were you when you were first diagnosed with DCIS (ductal carcinoma in situ) in your other breast?	Numeric
17	In which breast were you diagnosed with breast cancer?	Left Right Both Don't know
18	How old were you when you were first diagnosed with breast cancer?	Numeric
19	How old were you when you were first diagnosed with breast cancer in your other breast?	Numeric
20	Considering your health over the last month, how would you characterize your health?	Excellent Very Good Good Fair Poor

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
21	Has a doctor ever told you that you have any of the following health conditions? [Choose all that apply]	Alzheimer's disease, dementia Anemia or other blood disease Back Pain Being Overweight (Obesity) Cancer (other than breast cancer) Chronic Obstructive Pulmonary Disease Congestive Heart Failure Depression Diabetes (problem with your blood sugar levels) Heart Attack Heart Disease High Blood Pressure Kidney Disease Liver Disease Lung Disease (e.g., asthma, pulmonary fibrosis, etc.) Osteoarthritis, degenerative arthritis Rheumatoid arthritis Stroke Thyroid Disease Ulcer or Stomach Disease Don't know None Other conditions (not listed)
22	Which type(s) of the following cancers have you had? [Choose all that apply]	Adrenal Cancer Bladder Cancer Bone Cancer Brain Cancer Breast Cancer Cervical Cancer Colon, Rectal, Large Intestine Cancer Esophageal Cancer Gall Bladder Cancer Gastric/Stomach Cancer Small Intestine Cancer Kidney or Renal Cancer Leukemia Lung Cancer Lymphoma - Hodgkins Lymphoma - non-Hodgkins Ovarian, Peritoneal, Fallopian Tube Cancer Pancreatic Cancer

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
		Prostate Cancer
		Renal Pelvis / Ureter Cancer
		Sarcoma
		Skin - Basal / Squamous Cell Cancer
		Skin - Melanoma Cancer
		Testicular Cancer
		Thyroid Cancer
		Uterine (non-Cervical) / Endometrial Cancer
		Unknown Type of Cancer
		Other
23	When were you diagnosed with ovarian cancer? (Date can be estimated if exact date is not known)	Date
24	Have you ever had chemotherapy?	No
		Yes
		Don't know
25	How old were you when you had chemotherapy?	Numeric
26	Why did you have chemotherapy?	Free text
27	Have you ever had radiation therapy on your chest wall prior to the age of 35?	No
		Yes
		Don't know
28	How old were you when you received radiation therapy on your chest wall?	Numeric
29	Why did you have radiation therapy on your chest wall?	Free text
30	Are you adopted?	Yes
		No
		Don't know
31	Do you know the medical history from your biological family?	Yes
		No
		Don't know
Instructions: Please answer the following questions for your blood relatives, including half-relatives and relatives who have died, but not step or adoptive relatives.		
32	How many sisters do you have?	Numeric
33	How many daughters do you have?	Numeric
34	How many maternal aunts (mom's sisters) do you have?	Numeric
35	How many paternal aunts (father's sisters) do you have?	Numeric
36	Which of your blood relatives have ever been diagnosed with breast cancer or ductal carcinoma in situ (DCIS)? [Choose all that apply]	Mother
		Sister(s)
		Daughter(s)
		Maternal grandmother (mother's mother)
		Paternal grandmother (father's mother)
		Maternal aunt(s) - mother's sisters
		Paternal aunt(s) - father's sisters

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
		Any male relatives Don't know None of the above
37	Please specify if your mother was diagnosed with breast cancer or ductal carcinoma in situ (DCIS). [Choose all that apply]	Invasive Breast Cancer Ductal carcinoma in situ (DCIS) I don't know what kind of breast cancer
38	At about what age did your mother have breast cancer? (Please estimate if you are not sure)	Numeric
39	At about what age did your mother have DCIS (ductal carcinoma in situ)? (Please estimate if you are not sure)	Numeric
40	Did your mother have breast cancer or DCIS in BOTH breasts (bilateral)?	Yes No Don't know
41	Please specify if your sister(s) were diagnosed with breast cancer or ductal carcinoma in situ (DCIS). [Choose all that apply]	Invasive Breast Cancer Ductal carcinoma in situ (DCIS) I don't know what kind of breast cancer
42	How many sisters have ever been diagnosed with breast cancer?	Numeric
43	How many sisters have ever been diagnosed with DCIS (ductal carcinoma in situ)?	Numeric
44	At about what age did your sister(s) have breast cancer (if more than one sister had breast cancer, please select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric
45	At about what age did your sister(s) have DCIS (if more than one sister had DCIS, please select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric
46	Did any of your sisters have breast cancer or DCIS in BOTH breasts (bilateral)?	Yes No Don't know
47	Please specify if your daughter(s) were diagnosed with breast cancer or ductal carcinoma in situ (DCIS)? [Choose all that apply]	Invasive Breast Cancer Ductal carcinoma in situ (DCIS) I don't know what kind of breast cancer
48	How many of your daughters have ever been diagnosed with breast cancer?	Numeric
49	How many of your daughters have ever been diagnosed with DCIS (ductal carcinoma in situ)?	Numeric
50	At about what age did your daughter(s) have breast cancer (if more than one daughter had breast cancer, please select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
51	At about what age did your daughter(s) have DCIS (if more than one daughter had DCIS, please select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric
52	Did any of your daughters have breast cancer or DCIS in BOTH breasts (bilateral)?	Yes No Don't know
53	Please specify if your maternal grandmother (mother's mother) was diagnosed with breast cancer or ductal carcinoma in situ (DCIS)? [Choose all that apply]	Invasive Breast Cancer Ductal carcinoma in situ (DCIS) I don't know what kind of breast cancer
54	At about what age did your maternal grandmother have breast cancer? (Please estimate if you are not sure)	Numeric
55	At about what age did your maternal grandmother have DCIS (ductal carcinoma in situ)? (Please estimate if you are not sure)	Numeric
56	Did your maternal grandmother have breast cancer or DCIS in BOTH breasts (bilateral)?	Yes No Don't know
57	Please specify if your paternal grandmother (father's mother) was diagnosed with breast cancer or ductal carcinoma in situ (DCIS)? [Choose all that apply]	Invasive Breast Cancer Ductal carcinoma in situ (DCIS) I don't know what kind of breast cancer
58	At about what age did your paternal grandmother have breast cancer? (Please estimate if you are not sure)	Numeric
59	At about what age did your paternal grandmother have DCIS (ductal carcinoma in situ)? (Please estimate if you are not sure)	Numeric
60	Did your paternal grandmother have breast cancer or DCIS in BOTH breasts (bilateral)?	Yes No Don't know
61	Please specify if your maternal aunt(s) (mother's sisters) were diagnosed with breast cancer or ductal carcinoma in situ (DCIS)? [Choose all that apply]	Invasive Breast Cancer Ductal carcinoma in situ (DCIS) I don't know what kind of breast cancer
62	How many of your maternal aunts (mother's sisters) have been diagnosed with breast cancer?	Numeric
63	How many of your maternal aunts (mother's sisters) have been diagnosed DCIS (ductal carcinoma in situ)?	Numeric
64	At about what age did your maternal aunt(s) have breast cancer (if more than one maternal aunt had breast cancer, select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
65	At about what age did your maternal aunt(s) have DCIS (if more than one maternal aunt had DCIS, select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric
66	Did any of your maternal aunts have breast cancer or DCIS in BOTH breasts (bilateral)?	Yes No Don't know
67	Please specify if your paternal aunt(s) (father's sisters) were diagnosed with breast cancer or ductal carcinoma in situ (DCIS)? [Choose all that apply]	Invasive Breast Cancer Ductal carcinoma in situ (DCIS) I don't know what kind of breast cancer
68	How many of your paternal aunts (father's sisters) have ever been diagnosed with breast cancer?	Numeric
69	How many of your paternal aunts (father's sisters) have been diagnosed with DCIS (ductal carcinoma in situ)?	Numeric
70	At about what age did your paternal aunt(s) have breast cancer (if more than one paternal aunt had breast cancer, select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric
71	At about what age did your paternal aunt(s) have DCIS (if more than one paternal aunt had DCIS, select youngest age of diagnosis)? (Please estimate if you are not sure)	Numeric
72	Did any of your paternal aunts have breast cancer or DCIS in BOTH breasts (bilateral)?	Yes No Don't know
73	How many of your close, blood, female relatives (mother, sisters, daughters, grandmothers, aunts) have ever been diagnosed with ovarian cancer?	Numeric
74	Which of your blood relatives have ever been diagnosed with ovarian cancer? [Choose all that apply]	Mother Sister(s) Daughter(s) Maternal grandmother (mother's mother) Paternal grandmother (father's mother) Maternal aunt(s) - mother's sisters Paternal aunt(s) - father's sisters Don't know None of the above
75	Have any of your close, blood, female relatives (mother, sisters, daughters, grandmothers, or aunts) ever been diagnosed with BOTH breast and ovarian cancers?	Yes No Don't know

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
76	Please check all of the statements below that apply to you: [Choose all that apply]	I have 2 or more relatives on the same side of my family who have been diagnosed with uterus (endometrial) and/or colon cancer before age 55 I have a niece or cousin who was diagnosed with breast cancer (or ductal carcinoma in situ (DCIS) before age 55 I have a relative who was diagnosed with any type of cancer before age 20 I have a relative who was diagnosed with leukemia, brain cancer, or a sarcoma before age 45 Don't know None of the above
77	Have you or your relatives ever had genetic testing for breast cancer risk? [Choose all that apply]	Yes - I have Yes - my relatives have No Don't know
78	Were you positive for a gene mutation (e.g., BRCA1, BRCA2, gene variant, etc.):	Yes No Don't know
79	Do you know what type of gene mutation you have? [Choose all that apply]	BRCA1 BRCA2 SNPs P53 PTEN CDH1 CHEK2 Gene variant Other Don't know
80	Were any of your relatives positive for a gene mutation (e.g., BRCA1, BRCA2, gene variant, etc.):	Yes No Don't know
81	Do you know what type of gene mutation your relative(s) have: [Choose all that apply]	BRCA1 BRCA2 P53 PTEN CDH1 SNPs CHEK2 Gene variant Other

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
		Don't know
82	How many times have you been pregnant?	Numeric
83	Have you ever given birth?	Yes
		No
84	How many live births?	Numeric
85	How old were you when your first child was born?	Numeric
86	How old were you when you had your first menstrual period (please estimate if you're not sure)?	Under age 10
		10
		11
		12
		13
		14
		15
		16
		Over 16
		Don't know
87	Have your menstrual periods stopped permanently?	Yes - Periods stopped naturally (menopause)
		Yes - But now have periods induced by hormones
		Yes - Uterus removed by surgery
		Yes - Both ovaries removed by surgery
		Yes - Uterus AND both ovaries removed by surgery
		Yes - Uterus AND one ovary removed by surgery
		Yes - Due to radiation
		Yes - Other reason
		No - my menstrual periods have not stopped
		Not sure - periods less frequent
88	How old were you when your periods stopped?	Numeric
89	Have you ever had a hysterectomy (a surgery to remove your uterus or womb)?	Yes
		No
		Don't know
90	At what age did you have your hysterectomy?	Numeric
91	Have you ever had one, both, or part of an ovary removed (oophorectomy)?	Yes - both ovaries were removed
		Yes - only one ovary was removed
		Yes - only part of one ovary was removed
		Yes - but I don't know whether one or both ovaries were removed
		No - I have never had an oophorectomy

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
		Don't know
92	Were your ovaries removed at the same time?	Yes - same time
		No - different times
		Don't know
93	At what age did you first have an ovary or part of an ovary removed?	Numeric
94	Have you had any of the following breast procedures on your RIGHT breast? [Choose all that apply]	Fine Needle Aspiration (FNA)
		Core biopsy
		Surgical biopsy
		Lumpectomy for cancer
		Mastectomy
		Radiation Therapy
		Breast reconstruction
		Breast reduction
		Implants
		None of the above
		Don't know
95	When did you have your FIRST fine needle aspiration (FNA) on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
96	When did you have your FIRST core biopsy on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
97	When did you have your FIRST surgical biopsy on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
98	When did you have a lumpectomy for cancer on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
99	When did you have a mastectomy on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
100	When did you COMPLETE radiation therapy on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
101	When did you have breast reconstruction on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
102	When did you have breast reduction on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
103	When did you get breast implants in your RIGHT breast? (Date can be estimated if exact date is not known)	Date
104	Have you had any of the following breast procedures on your LEFT breast? [Choose all that apply]	Fine Needle Aspiration (FNA)
		Core biopsy
		Surgical biopsy
		Lumpectomy for cancer
		Mastectomy

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
		Radiation Therapy
		Breast reconstruction
		Breast reduction
		Implants
		None of the above
		Don't know
105	When did you have your FIRST fine needle aspiration (FNA) on your LEFT breast? (Date can be estimated if exact date is not known)	Date
106	When did you have your FIRST core biopsy on your LEFT breast? (Date can be estimated if exact date is not known)	Date
107	When did you have your FIRST surgical biopsy on your LEFT breast? (Date can be estimated if exact date is not known)	Date
108	When did you have a lumpectomy for cancer on your LEFT breast? (Date can be estimated if exact date is not known)	Date
109	When did you have a mastectomy on your LEFT breast? (Date can be estimated if exact date is not known)	Date
110	When did you COMPLETE radiation therapy on your LEFT breast? (Date can be estimated if exact date is not known)	Date
111	When did you have breast reconstruction on your LEFT breast? (Date can be estimated if exact date is not known)	Date
112	When did you have breast reduction on your LEFT breast? (Date can be estimated if exact date is not known)	Date
113	When did you get breast implants in your LEFT breast? (Date can be estimated if exact date is not known)	Date
114	How many breast biopsies (e.g., needle biopsies, surgeries) have you had in total?	0 1 More than 1 Don't know
115	Have any of your breast biopsies (e.g., needle biopsies, surgeries) showed atypical ductal hyperplasia, sometimes called atypia or ADH?	Yes No Don't know
116	Have any of your breast biopsies (e.g., needle biopsies, surgeries) shown LCIS (lobular carcinoma in situ)?	No Yes Don't know
117	Have you had a cyst aspiration on your breast(s)?	Yes, don't know which side Yes, on both breasts Yes, on my left breast Yes, on my right breast No Don't know

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
118	When did you have your FIRST cyst aspiration on your RIGHT breast? (Date can be estimated if exact date is not known)	Date
119	When did you have your FIRST cyst aspiration on your LEFT breast? (Date can be estimated if exact date is not known)	Date
120	When did you have your FIRST cyst aspiration? (Date can be estimated if exact date is not known)	Date
121	What is your height in feet (inches will be asked in the next question)?	3 feet 4 feet 5 feet 6 feet 7 feet 8 feet
122	What is your current height in inches?	0 inches 1 inch 2 inches 3 inches 4 inches 5 inches 6 inches 7 inches 8 inches 9 inches 10 inches 11 inches
123	How often do you have a drink containing alcohol?	Never Once per month or under 2-4 times per MONTH 2-3 times per WEEK 4 or more times per WEEK
124	On a typical day that you have alcohol, how many drinks do you have?	1 or 2 3 or 4 5 or 6 7 to 9 10 or more
125	Think about the WALKING you do outside the home. How often do you walk outside the home for more than 10 minutes without stopping?	Rarely or never 1 - 3 times each month 1 time each week 2 - 3 times each week 4 - 6 times each week 7 or more times each week
126	When you walk outside the home for more than 10 minutes without stopping, for how many minutes do you usually walk?	Less than 20 min 20 - 39 minutes 40 - 59 minutes

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
		1 hour or more
127	What is your usual speed of walking?	Casual strolling or walking (less than 2 miles an hour) Average or normal (2-3 miles an hour) Fairly fast (3-4 miles an hour) Very fast (more than 4 miles an hour) Don't know
Instructions: The following questions ask about four types of exercise: *1) walking* *2) strenuous or very hard exercise (you work up a sweat and your heart beats fast)* *3) moderate exercise (not exhausting)* *4) mild exercise*		
128	Think about the STRENUOUS OR VERY HARD EXERCISE you do (you work up a sweat and your heart beats fast); for example, aerobics, aerobic dancing, jogging, tennis, swimming laps). How many days per week do you do this type of exercise?	none 1 day each week 2 days each week 3 days each week 4 days each week 5 or more days each week
129	How long do you usually exercise like this at one time?	Less than 20 min 20 - 39 minutes 40 - 59 minutes 1 hour or more
130	Think about the MODERATE exercise you do (not exhausting; for example, biking outdoors, using an exercise machine (like a stationary bike or treadmill), calisthenics, easy swimming, popular or folk dancing. How many days per week do you do this type of exercise?	none 1 day each week 2 days each week 3 days each week 4 days each week 5 or more days each week
131	How long do you usually exercise like this at one time?	Less than 20 min 20 - 39 minutes 40 - 59 minutes 1 hour or more
132	Think about the MILD exercise you do; for example, slow dancing, bowling, golf, yoga. How many days per week do you do this type of exercise?	none 1 day each week 2 days each week 3 days each week 4 days each week 5 or more days each week
133	How long do you usually exercise like this at one time?	Less than 20 min 20 - 39 minutes 40 - 59 minutes 1 hour or more

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
134	What is your racial background? [Choose all that apply]	Black or African American White Asian American Indian or Alaska Native Native Hawaiian or Other Pacific Islander Some other race Don't know Prefer not to answer
135	What is your ASIAN background? [Choose all that apply]	Chinese Filipino Asian Indian Japanese Korean Vietnamese Other Asian Don't know
136	What is your PACIFIC ISLANDER background? [Choose all that apply]	Native Hawaiian Samoan Guamanian or Chamorro Other Pacific Islander Don't know
137	Are you of Hispanic, Latino or Spanish origin or ancestry?	No, not of Hispanic, Latino or Spanish origin Yes - Mexican, Mexican American, or Chicano Yes - Puerto Rican Yes - Cuban Yes - other Hispanic, Latino, or Spanish origin Prefer not to answer
138	Do you have any Jewish ancestry in your family?	Yes - only on my father's side Yes - only on my mother's side Yes - on both sides Yes - not sure which side of my family No Don't know Prefer not to answer
139	What best describes your current marital status?	Married Widowed Living with a partner in a marriage-like relationship Never married Divorced

Table continues

Table A1. Full text of study questionnaires (continued)

ID	Question	Response Choices
		Separated
140	How many years of schooling have you had?	Some high school or less
		High school graduate
		Some college or technical school
		College graduate or more

Table A2. Assay characteristics for urinary glyphosate and AMPA measurements. As determined at assay validation prior to measurements performed in study samples. Precision is shown as intraday (interday) precision. AMPA: aminomethylphosphonic acid

Analyte	Coefficient of determination (R^2)	LOD (ng/mL)	LOQ (ng/mL)	Precision (%)				
				At LOD	At LOQ	Quality control		
						Low (0.25 ng/mL)	Medium (0.75 ng/mL)	High (2.5 ng/mL)
Glyphosate	0.9921	0.014	0.041	13.7 (19.6)	5.3 (7.4)	12.2 (7.2)	6.9 (5.3)	2.9 (1.2)
AMPA	0.9951	0.013	0.040	6.3 (4.5)	6.9 (9.7)	9.4 (9.2)	6.8 (3.4)	4.4 (4.2)

Table A3. List of food and beverage items included in additional categories

Corn
Corn flakes, NFS
Corn Puffs
Corn, cooked, NS as to form, NS as to color, NS as to fat added in cooking
Corn, dried, cooked
Corn, white, cooked, from fresh, fat not added in cooking
Corn, white, cooked, from frozen, fat not added in cooking
Corn, white, cooked, from frozen, NS as to fat added in cooking
Corn, yellow and white, cooked, from fresh, fat added in cooking W/ BUTTER, NFS
Corn, yellow and white, cooked, from fresh, fat not added in cooking
Corn, yellow, cooked, from canned, NS as to fat added in cooking
Corn, yellow, cooked, from fresh, fat added in cooking
Corn, yellow, cooked, from fresh, fat added in cooking W/ VEGETABLE OIL, NFS (INCLUDE OIL, NFS)
Corn, yellow, cooked, from fresh, fat not added in cooking
Corn, yellow, cooked, from fresh, NS as to fat added in cooking
Corn, yellow, cooked, from frozen, fat added in cooking W/ VEGETABLE OIL, NFS (INCLUDE OIL, NFS)
Corn, yellow, cooked, from frozen, fat not added in cooking
Corn, yellow, cooked, NS as to form, fat not added in cooking
Cornbread, made from home recipe
Cornbread, prepared from mix
Grits, cooked, corn or hominy, instant, fat not added in cooking
Mixed vegetables (corn, lima beans, peas, green beans, and carrots), cooked, from frozen, fat added in cooking W/ BUTTER, NFS
Mixed vegetables (corn, lima beans, peas, green beans, and carrots), cooked, from frozen, fat not added in cooking
Mixed vegetables (corn, lima beans, peas, green beans, and carrots), cooked, NS as to form, NS as to fat added in cooking
Peas and corn, cooked, fat not added in cooking
Popcorn, air-popped (no butter or no oil added)
Popcorn, air-popped, buttered
Popcorn, flavored
Popcorn, popped in oil, buttered
Popcorn, popped in oil, lowfat
Popcorn, popped in oil, lowfat, low sodium
Popcorn, popped in oil, unbuttered
Popcorn, sugar syrup or caramel-coated
Salty snacks, corn based puffs and twists, cheese puffs and twists, lowfat
Salty snacks, corn or cornmeal base, corn chips, corn-cheese chips
Salty snacks, corn or cornmeal base, corn puffs and twists
corn-cheese puffs and twists
Salty snacks, corn or cornmeal base, nuts or nuggets, toasted
Salty snacks, corn or cornmeal base, tortilla chips
Salty snacks, corn or cornmeal base, tortilla chips, light (baked with less oil)
Salty snacks, corn or cornmeal base, tortilla chips, lowfat, baked without fat
Salty snacks, corn or cornmeal base, tortilla chips, unsalted
Taco shell, corn
Tortilla, corn
Oats
Bread, oat bran, toasted
Bread, oatmeal
Cheerios

Table continues

Table A3. List of food and beverage items included in additional categories
(continued)

Oats, continued
Cookie, oatmeal
Cookie, oatmeal, with raisins
Cracklin' Oat Bran
Granola bar, lowfat, NFS
Granola bar, NFS
Granola bar, reduced sugar, NFS
Granola with Raisins, lowfat, Kellogg's
Granola, homemade
Granola, lowfat, Kellogg's
Granola, NFS
Honey Bunches of Oats Honey Roasted Cereal
Honey Bunches of Oats with Almonds, Post
Honey Nut Cheerios
MultiGrain Cheerios
Nature Valley Chewy Trail Mix Granola Bar
Nature Valley Crunchy Granola Bar
Nature Valley Granola, with fruit and nuts
Nature Valley Sweet and Salty Granola Bar
Oatmeal, cooked, instant, fat not added in cooking
Oatmeal, cooked, instant, made with milk, fat not added in cooking
Oatmeal, cooked, NS as to regular, quick or instant, fat not added in cooking
Oatmeal, cooked, NS as to regular, quick or instant
NS as to fat added in cooking
Oatmeal, cooked, quick (1 or 3 minutes), fat not added in cooking
Oatmeal, cooked, quick (1 or 3 minutes), made with milk, fat not added in cooking
Oatmeal, cooked, regular, fat added in cooking
Oatmeal, cooked, regular, fat not added in cooking
Oatmeal, cooked, regular, made with milk, fat not added in cooking
Oatmeal, NS as to regular, quick, or instant, made with milk, fat not added in cooking
Oats, raw
Quaker Chewy Granola Bar
Quaker Oatmeal Squares (formerly Quaker Oat Squares)
Snack bar, oatmeal
Other soy foods
Beef and rice with soy-based sauce (mixture)
Beef, noodles, and vegetables (including carrots, broccoli, and/or dark-green leafy), soy-based sauce (mixture)
Chicken or turkey and rice with soy-based sauce (mixture)
Chicken or turkey and vegetables (excluding carrots, broccoli, and dark-green leafy (no potatoes)), soy-based sauce (mixture)
Chicken or turkey and vegetables (including carrots, broccoli, and/or dark-green leafy (no potatoes)), soy-based sauce (mixture)
Chicken or turkey teriyaki (chicken or turkey with soy-based sauce)
Chicken or turkey, rice, and vegetables (including carrots, broccoli, and/or dark-green leafy), soy-based sauce (mixture)
Coffee, latte W/ SOY MILK
Edamame
Fish and vegetables (excluding carrots, broccoli, and dark-green leafy (no potatoes)), soy-based sauce (mixture)

Table continues

Table A3. List of food and beverage items included in additional categories
(continued)

Other soy foods, continued
Gravy or sauce, Chinese (soy sauce, stock or bouillon, cornstarch)
Milk, soy, light, ready-to-drink, not baby's
Milk, soy, light, ready-to-drink, not baby's, chocolate
Milk, soy, nonfat, ready-to-drink, not baby's
Milk, soy, ready-to-drink, not baby's
Rice, white, with soy-based sauce, fat added in cooking
Rice, white, with soy-based sauce, fat not added in cooking
Rice, white, with vegetables, soy-based sauce, fat added in cooking
Seaweed, prepared with soy sauce
Shrimp and vegetables (including carrots, broccoli, and/or dark-green leafy (no potatoes)), soy-based sauce (mixture)
Shrimp teriyaki (shrimp with soy-based sauce) (mixture)
Soy sauce
Soy sauce, reduced sodium
Vegetable combination (excluding carrots, broccoli, and dark-green leafy), cooked, with soy-based sauce
Vegetable combination (including carrots, broccoli, and/or dark-green leafy), cooked, with soy-based sauce
Coffee
Blended coffee beverage, made with regular coffee, milk, and ice, sweetened
Café con leche
Café con leche prepared with sugar
Cappuccino
Cappuccino W/ 2% MILK
Cappuccino, decaffeinated
Coffee and cocoa (mocha), made from powdered instant mix, with whitener, presweetened
Coffee, decaffeinated, made from ground
Coffee, decaffeinated, made from powdered instant
Coffee, espresso
Coffee, espresso, decaffeinated
Coffee, Latte
Coffee, latte W/ 1% MILK
Coffee, latte W/ 2% MILK
Coffee, latte W/ SKIM MILK
Coffee, latte W/ SOY MILK
Coffee, latte W/ WHOLE MILK
Coffee, latte, decaffeinated W/ 2% MILK
Coffee, latte, decaffeinated W/ SKIM MILK
Coffee, latte, W/ SOY MILK, decaffeinated
Coffee, made from ground, equal parts regular and decaffeinated
Coffee, made from ground, regular
Coffee, made from liquid concentrate
Coffee, made from powdered instant mix, with whitener and low calorie sweetener
Coffee, made from powdered instant mix, with whitener and sugar, instant
Coffee, made from powdered instant, 50% less caffeine
Coffee, made from powdered instant, regular
Coffee, mocha
Coffee, mocha W/ WHOLE MILK
Coffee, regular, NS as to ground or instant
Coffee, regular, presweetened with sugar, pre-lightened

Table continues

Table A3. List of food and beverage items included in additional categories
(continued)

Tea
Half and Half beverage, half iced tea and half fruit juice drink (lemonade)
Tea, herbal
Tea, herbal, presweetened with low calorie sweetener
Tea, herbal, presweetened with sugar
Tea, hibiscus
Tea, leaf, decaffeinated, presweetened, NS as to sweetener
Tea, leaf, decaffeinated, unsweetened
Tea, leaf, presweetened with low calorie sweetener
Tea, leaf, presweetened with sugar
Tea, leaf, presweetened, NS as to sweetener
Tea, leaf, unsweetened
Tea, made from powdered instant, unsweetened
Tea, NS as to type, decaffeinated, presweetened with sugar
Tea, NS as to type, presweetened with low calorie sweetener
Tea, NS as to type, presweetened with sugar
Tea, NS as to type, presweetened, NS as to sweetener
Tea, NS as to type, unsweetened
Wine
Brandy
Wine cooler
Wine, dessert, sweet
Wine, rice
Wine, table, red
Wine, table, white
Beer
Beer
Beer, lite
Alcoholic malt beverage, sweetened
Spirits
Cocktail, NFS
Gin
Gin and Tonic
Liqueur with cream
Margarita
Martini
Rum
Vodka
Whiskey

Table A4. Sensitivity analysis: association of dietary and demographic/behavioral variables with glyphosate and AMPA levels, excluding poor-quality dietary recalls (remaining N = 461 recalls)

	Glyphosate		AMPA	
	Coefficient (95% CI)	p	Coefficient (95% CI)	p
Dietary Recall Characteristics				
Major Food Groups (Quartile)				
Fruit	-0.01 (-0.11, 0.09)	0.86	-0.10 (-0.22, 0.01)	0.08
Vegetables	-0.04 (-0.14, 0.06)	0.48	-0.02 (-0.13, 0.10)	0.77
Grains	0.14 (0.03, 0.25)	**0.01**	0.08 (-0.04, 0.20)	0.21
Whole Grains	0.12 (0.02, 0.21)	**0.02**	-0.03 (-0.14, 0.08)	0.58
Refined Grains	0.12 (0.005, 0.23)	**0.04**	0.17 (0.05, 0.30)	**0.01**
Protein Foods	-0.11 (-0.21, -0.01)	**0.03**	0.04 (-0.06, 0.15)	0.42
Meat and Poultry	-0.10 (-0.20, -0.002)	**0.048**	0.02 (-0.09, 0.12)	0.78
Dairy	-0.06 (-0.16, 0.04)	0.22	-0.08 (-0.19, 0.03)	0.14
Other Foods and Beverages (Yes/No)				
Legumes	0.20 (-0.04, 0.44)	0.10	-0.03 (-0.30, 0.24)	0.81
Corn	-0.12 (-0.35, 0.10)	0.29	-0.39 (-0.65, -0.14)	**0.003**
Oats	0.51 (0.24, 0.79)	**<0.001**	0.11 (-0.19, 0.42)	0.47
Eggs	-0.05 (-0.28, 0.17)	0.64	0.06 (-0.20, 0.31)	0.67
Soy Protein	-0.07 (-0.34, 0.20)	0.62	0.27 (-0.03, 0.57)	0.08
Soy Protein, Milk, and Other Soy Foods	-0.11 (-0.34, 0.11)	0.32	0.18 (-0.08, 0.43)	0.17
Nuts and Seeds	0.08 (-0.14, 0.31)	0.46	0.04 (-0.21, 0.29)	0.76
Coffee	0.07 (-0.18, 0.31)	0.59	0.01 (-0.25, 0.27)	0.93
Tea	-0.02 (-0.26, 0.21)	0.83	0.10 (-0.15, 0.35)	0.44
Alcohol	0.05 (-0.20, 0.30)	0.70	0.49 (0.23, 0.75)	**<0.001**
Wine	0.05 (-0.21, 0.31)	0.71	0.49 (0.20, 0.77)	**0.001**
Beer	-0.17 (-0.61, 0.27)	0.45	0.31 (-0.18, 0.80)	0.22
Spirits	0.21 (-0.22, 0.64)	0.35	0.22 (-0.26, 0.70)	0.36
Overall Dietary Pattern				
HEI (for individual dietary recall)	-0.0003 (-0.01, 0.01)	0.95	-0.01 (-0.01, 0.003)	0.18
Individual Characteristics				
Age	0.03 (-0.003, 0.05)	0.08	0.01 (-0.02, 0.04)	0.56
Race/Ethnicity				
Non-Hispanic White	Ref		Ref	
Asian	-0.31 (-0.74, 0.12)	0.16	0.07 (-0.37, 0.51)	0.74
Hispanic	0.10 (-0.24, 0.44)	0.58	-0.02 (-0.37, 0.32)	0.89
Other Race	-0.48 (-1.14, 0.17)	0.15	-0.82 (-1.50, -0.14)	**0.02**
Education				
College Graduate	Ref		Ref	
Some College	-0.11 (-0.44, 0.22)	0.52	0.06 (-0.28, 0.39)	0.74
High School Graduate	0.13 (-0.43, 0.69)	0.64	0.14 (-0.44, 0.71)	0.64
Some High School	0.51 (-0.40, 1.42)	0.27	-0.57 (-1.48, 0.34)	0.22

Table continues

Table A4. Sensitivity analysis: association of dietary and demographic/behavioral variables with glyphosate and AMPA levels, excluding poor-quality dietary recalls (remaining N = 461 recalls) (continued)

	Glyphosate		AMPA	
	Coefficient (95% CI)	p	Coefficient (95% CI)	p
Median Annual Household Income				
Less than $100,000	Ref		Ref	
$100,000 - $149,999	-0.01 (-0.31, 0.29)	0.95	0.19 (-0.12, 0.49)	0.23
$150,000 or more	0.04 (-0.33, 0.41)	0.84	0.06 (-0.32, 0.43)	0.77
BMI	-0.001 (-0.02, 0.02)	0.95	0.01 (-0.01, 0.03)	0.57
Physical Activity				
<150 minutes/week	Ref		Ref	
≥150 minutes/week	-0.23 (-0.49, 0.02)	0.07	-0.15 (-0.41, 0.11)	0.26
Smoking Status				
Never	Ref		Ref	
Former	0.04 (-0.25, 0.34)	0.79	-0.17 (-0.47, 0.13)	0.26
Current	-0.24 (-0.88, 0.41)	0.47	-0.26 (-0.92, 0.40)	0.44
Fast Food Meals				
Less than 1 time/month	Ref		Ref	
1-3 times/month	-0.19 (-0.50, 0.11)	0.22	-0.08 (-0.39, 0.24)	0.63
1-3 times/week	0.09 (-0.27, 0.45)	0.63	0.06 (-0.31, 0.43)	0.74
4 or more times/week	0.33 (-0.59, 1.25)	0.49	0.72 (-0.22, 1.67)	0.13
Organic Eating				
Seldom/Never	0.22 (-0.09, 0.52)	0.17	0.24 (-0.08, 0.55)	0.14
Sometimes	0.19 (-0.12, 0.50)	0.23	0.17 (-0.15, 0.48)	0.30
Often/Always	Ref		Ref	
Primary Drinking Water Source				
Bottled	Ref		Ref	
Tap (Filtered)	-0.08 (-0.35, 0.20)	0.60	-0.13 (-0.41, 0.15)	0.37
Tap (Unfiltered)	-0.06 (-0.51, 0.39)	0.80	-0.31 (-0.77, 0.15)	0.18

Adjusted for urinary creatinine, total Kcal, recall-specific Healthy Eating Index (HEI), race/ethnicity, and physical activity level. BMI: body mass index, AMPA: aminomethylphosphonic acid

Table A5. Distribution of estimated cell type proportions from the Houseman method, by glyphosate and AMPA tertile.

	Proportions, Mean (SD)				Proportions, Mean (SD)			
	Glyphosate Tertile 1	Glyphosate Tertile 2	Glyphosate Tertile 3	p	AMPA Tertile 1	AMPA Tertile 2	AMPA Tertile 3	p
CD8T	10.7 (4.2)	10.7 (4.0)	10.3 (3.8)	0.62	10.4 (4.3)	11.0 (4.0)	10.3 (3.7)	0.27
CD4T	15.4 (6.0)	14.7 (5.1)	15.3 (5.6)	0.65	15.3 (5.9)	15.3 (5.7)	14.7 (5.3)	0.35
NK	5.8 (2.4)	5.4 (1.9)	5.6 (2.4)	0.56	5.6 (2.5)	5.7 (2.0)	5.5 (2.3)	0.47
B Cell	6.4 (2.5)	6.4 (2.7)	6.4 (2.2)	0.73	6.5 (3.0)	6.5 (2.1)	6.2 (2.1)	0.50
Monocyte	8.9 (1.8)	8.9 (2.2)	9.1 (2.1)	0.95	9.0 (1.8)	8.8 (1.9)	9.1 (2.3)	0.18
Neutrophil	57.8 (9.6)	58.7 (8.8)	58.4 (8.6)	0.69	58.1 (9.6)	57.6 (8.3)	59.2 (9.0)	0.28

Values are shown as mean proportion (SD). P-values are for each cell type in a linear regression model including all cell types. Results were consistent when glyphosate and AMPA were analyzed as continuous variables. AMPA: aminomethylphosphonic acid

Table A6. Cohort characteristics stratified by randomized set assignment (training or validation).

	Training	Validation	
	N (%) or Mean (SD)		p
Race/Ethnicity			0.44
Asian	37 (11.3%)	6 (10.2%)	
Hispanic	54 (16.5%)	15 (25.4%)	
Other	18 (5.5%)	2 (3.4%)	
White	218 (66.7%)	36 (61.0%)	
Age (years)	56.9 (4.6)	56.8 (4.3)	0.81
BMI (kg/m^2)	26.7 (6.2)	27.4 (8.1)	0.39
Smoking Status			0.029
Never	250 (75.3%)	35 (59.3%)	
Former	68 (20.5%)	21 (35.6%)	
Current	14 (4.2%)	3 (5.1%)	
Alcohol (drinks/week)			0.99
None	92 (27.7%)	15 (25.4%)	
1 or fewer	136 (41.0%)	25 (42.4%)	
2-6	61 (18.4%)	11 (18.6%)	
7 or more	43 (13.0%)	8 (13.6%)	
Organic Eating			0.23
Seldom/Never	109 (32.8%)	15 (25.4%)	
Sometimes	99 (29.8%)	15 (25.4%)	
Often/Always	124 (37.3%)	29 (49.2%)	
HEI	63.7 (12.3)	62.5 (11.2)	0.47
Physical Activity			0.88
≥150 minutes/week	129 (40.4%)	24 (42.1%)	
<150 minutes/week	190 (59.6%)	33 (57.9%)	
Herbicide Use (past week)			0.22
Yes	20 (9.7%)	1 (2.4%)	
No	186 (90.3%)	41 (97.6%)	
Glyphosate Tertile			0.99
Tertile 1	111 (33.4%)	20 (33.3%)	
Tertile 2	110 (33.1%)	20 (33.3%)	
Tertile 3	111 (33.4%)	20 (33.3%)	
AMPA Tertile			0.79
Tertile 1	109 (32.8%)	22 (36.7%)	
Tertile 2	110 (33.1%)	20 (33.3%)	
Tertile 3	113 (34.0%)	18 (30.0%)	

Randomization was stratified by glyphosate tertile. P-values are from ANOVA (continuous variables) or Fisher's exact test (categorical variables). BMI: body mass index; HEI: Healthy Eating Index; AMPA: aminomethylphosphonic acid

Table A7. Final coefficients for the glyphosate methylation index

Probe	Coefficient (Main Index)	Coefficient (Probes on 450k Array Only)
Intercept	43.914	29.672
cg01430385	-7.280	-6.829
cg02519806	-12.143	0
cg04915788	-6.087	-12.243
cg05730283	17.420	0.712
cg06993862	-10.710	N/A
cg07261978	-5.058	N/A
cg11062848	6.133	6.399
cg13170005	-2.366	-2.567
cg13499896	-3.027	N/A
cg19029576	-1.535	N/A
cg20531550	-2.269	-3.405
cg20540608	5.666	7.992
cg25597976	7.812	8.591
cg26483235	-19.675	-15.900
cg26787244	-8.724	-6.970
cg00355690	0	-3.643
cg07318309	0	N/A
cg07509511	0	N/A
cg13310154	0	0
cg16601151	0	N/A
cg18722557	0	0
cg24576174	0	N/A
cg25629796	0	N/A
cg26833395	0	N/A

All 24 glyphosate-associated DMPs are included in the table, but probes with zero coefficients were not selected by the LASSO operator. Probes exclusive to the EPIC array are marked N/A for the 450k index. To obtain the predicted glyphosate level (natural log) for a sample, multiply the β percentage values for each probe by the coefficient, add all values, and add the intercept.

Table A8. Distribution of estimated cell type proportions from the Houseman method, presented as mean (SD), stratified by mammographic density

	Mammographic Density Category				
	A: Almost entirely fatty	**B: Scattered fibroglandular densities**	**C: Heterogeneously dense**	**D: Extremely dense**	**p**
CD8T	9.9 (3.5)	10.7 (4.2)	10.8 (3.9)	10.4 (3.9)	0.20
CD4T	14.1 (4.6)	14.8 (5.4)	15.7 (5.8)	15.1 (5.9)	0.10
NK	4.7 (1.8)	5.6 (2.3)	5.9 (2.4)	5.7 (2.2)	**0.03**
B Cell	5.7 (1.9)	6.5 (2.7)	6.6 (2.5)	6.1 (2.2)	0.21
Monocyte	8.4 (1.7)	8.9 (2.1)	9.1 (1.9)	9.0 (2.0)	0.08
Neutrophil	62.0 (7.4)	58.3 (9.9)	57.0 (8.7)	58.8 (8.8)	0.40

p-values are from a multivariate linear model.

Table A9. Results of gene ontology and pathway analysis for differentially methylated probes (DMPs) and regions (DMRs) associated with increasing mammographic density

Category	Term	N Genes	Genes	p	Fold Enrich-ment
DMPs					
GO: Biological Process	Transcription from RNA polymerase II promoter	15	*POU2F2,ISL1,MEIS1,CTDP1,PAX6,CSRNP2,YBX1,ATOH1,RUNX2,CAMTA1,ZNF768,CDX2,RREB1,PAX8,KAT2A*	<0.001	2.91
GO: Biological Process	Somatic stem cell population maintenance	5	*HES1,VANGL2,CDX2,LIG4,PRDM16*	0.004	7.64
GO: Biological Process	Detection of bacterium	3	*HLA-DRB5,HLA-DRB1,NLRC4*	0.007	22.93
GO: Biological Process	Regulation of transcription, DNA-templated	26	*HES1,ABL1,MEIS1,RXRB,ZNF878,IRF2BP2,PAX6,ATOH1,YBX1,RHOH,YY1AP1,CHEK2,KAT2A,POU2F2,ISL1,ZNF512B,PRDM16,TTF1,HOXA11,TLX3,RUNX2,ZNF768,CDX2,RREB1,DNTTIP2,PAX8*	0.008	1.72
GO: Biological Process	Neuron fate specification	3	*TLX3,NTRK3,ISL1*	0.008	21.29
GO: Biological Process	Negative regulation of neuron differentiation	4	*TLX3,ISL1,MEIS1,PAX6*	0.02	7.10
GO: Biological Process	Neuron migration	5	*ATOH1,TLX3,NTRK3,PAX6,PCM1*	0.02	4.73
GO: Biological Process	Central nervous system development	5	*ATOH1,TLX3,LIG4,PAX6,PAX8*	0.03	4.14
GO: Biological Process	Positive regulation of insulin secretion involved in cellular response to glucose stimulus	3	*BAD,HLA-DRB5,HLA-DRB1*	0.03	10.28
GO: Biological Process	Chemical synaptic transmission	7	*PRKCG,GABBR2,MINK1,TPGS1,GAD2,GRIN2A,SLITRK5*	0.03	2.90
GO: Biological Process	Response to endoplasmic reticulum stress	4	*HYOU1,ATP2A1,ABL1,USP19*	0.04	5.30
GO: Cellular Component	Nucleus	70	*HES1,RANBP3,PRKCG,TACR3,MEIS1,MAD1L1,ACAD9,IRF2BP2,KDM4C,TACC2,ATOH1,YY1AP1,RNF216,SESN1,MAEA,HSCB,SLC3A2,NLRC4,CTDP1,ZNF512B,XPO6,CSRNP2,HOXA11,TLX3,EIF5A,MICAL3,PCDHA2,USP31,DGCR8,NICN1,DHX34,ALYREF,RREB1,NEDD4,PAX8,USP3,TOX,ABL1,IFIH1,PRKAB1,LIG4,RXRB,ZNF878,PAX6,XRRA1,YBX1,CAMTA1,PRPF8,TCF4,KAT2A,NCOR2,POU2F2,ISL1,PRDM16,TENM2,TTF1,METTL16,NUFIP1,RUNX2,LDHC,SRCAP,RFWD3,CDX2,TRAPPC12,MIS18A,HDLBP,CAMKK1,TCF7,DNTTIP2,PTBP1*	0.004	1.33
GO: Cellular Component	Nuclear chromatin	7	*USP3,KDM4C,RUNX2,NCOR2,PAX6,KAT2A,TCF4*	0.01	3.73
GO: Cellular Component	Nucleoplasm	38	*HES1,MED7,RANBP3,PDK2,ABL1,ZMYM5,PRKAB1,LIG4,RXRB,PAX6,KDM4C,YBX1,TRMU,RAD9B,PRPF8,TOX4,CHEK2,CCT4,KAT2A,ANAPC11,POU2F2,NCOR2,ISL1,FANCF,CTDP1,TTF1,RUNX2,MICAL3,RFWD3,CDX2,DGCR8,MIS18A,ALYREF,TCF7,RREB1,SLC39A7,PTBP1,PAX8*	0.02	1.41

Table continues

Table A9. Results of gene ontology and pathway analysis for differentially methylated probes (DMPs) and regions (DMRs) associated with increasing mammographic density (continued)

Category	Term	N Genes	Genes	p	Fold Enrich -ment
GO: Cellular Component	Transcription factor complex	6	*MED7,HOXA11,RUNX2,MEIS1,TCF7,TCF 4*	0.04	3.20
GO: Molecular Function	Transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	10	*YBX1,ATOH1,RUNX2,ISL1,POU2F2,CAM TA1,MEIS1,PAX6,PAX8,TCF4*	<0.001	4.26
GO: Molecular Function	Sequence-specific DNA binding	14	*HES1,POU2F2,ISL1,RXRB,PRDM16,PAX 6,TTF1,CSRNP2,HOXA11,ATOH1,TLX3,C AMTA1,CDX2,TCF7*	0.002	2.72
GO: Molecular Function	Chromatin binding	10	*USP3,YBX1,RUNX2,ISL1,NCOR2,MEIS1,P AX6,TTF1,KAT2A,TCF4*	0.02	2.57
GO: Molecular Function	RNA polymerase II core promoter sequence-specific DNA binding	4	*RUNX2,PAX6,RREB1,PAX8*	0.02	7.05
GO: Molecular Function	bHLH transcription factor binding	3	*RUNX2,ISL1,TCF4*	0.02	13.11
GO: Molecular Function	RNA polymerase II core promoter proximal region sequence-specific DNA binding	9	*YBX1,ATOH1,RUNX2,ISL1,POU2F2,MEIS 1,PAX6,PAX8,TCF4*	0.03	2.55
GO: Molecular Function	Cytokine activity	6	*NODAL,C1QTNF4,IL1F10,TGFB2,CMTM 7,LTB*	0.03	3.43
GO: Molecular Function	Ubiquitin-protein transferase activity	8	*ANAPC11,KLHL29,RFWD3,RNF216,NLR C4,MAEA,PAX6,NEDD4*	0.045	2.44
GO: Molecular Function	Chaperone binding	4	*HES1,HYOU1,HSCB,TBCD*	0.046	4.96
KEGG Pathway	Epstein-Barr virus infection	6	*JAK3,HLA-DRB5,NCOR2,HLA-DRB1,PLCG2,NEDD4*	0.01	4.76
KEGG Pathway	Calcium signaling pathway	7	*GNA14,TACR3,PRKCG,ATP2A1,ITPR1,PL CG2,GRIN2A*	0.01	3.79
KEGG Pathway	Rheumatoid arthritis	5	*ATP6V0E1,HLA-DRB5,HLA-DRB1,TGFB2,LTB*	0.01	5.50
KEGG Pathway	HTLV-I infection	8	*RANBP3,ANAPC11,JAK3,HLA-DRB5,HLA-DRB1,CHEK2,TGFB2,KAT2A*	0.01	3.05
KEGG Pathway	Non-small cell lung cancer	4	*BAD,PRKCG,RXRB,PLCG2*	0.02	6.92
KEGG Pathway	Ras signaling pathway	7	*BAD,PDGFD,PRKCG,ABL1,PLCG2,RALB, GRIN2A*	0.03	3.00
KEGG Pathway	Maturity onset diabetes of the young	3	*HES1,PKLR,PAX6*	0.03	11.18
KEGG Pathway	Thyroid hormone signaling pathway	5	*BAD,PRKCG,RXRB,PLCG2,KAT2A*	0.03	4.21
KEGG Pathway	Thyroid hormone synthesis	4	*PRKCG,ITPR1,TTF1,PAX8*	0.03	5.54
KEGG Pathway	Thyroid cancer	3	*RXRB,TCF7,PAX8*	0.03	10.02
KEGG Pathway	Cell cycle	5	*ANAPC11,ABL1,MAD1L1,CHEK2,TGFB2*	0.04	3.91
KEGG Pathway	Pathways in cancer	9	*BAD,PRKCG,ABL1,RXRB,PLCG2,TCF7,RA LB,TGFB2,PAX8*	0.045	2.22

Table continues

Table A9. Results of gene ontology and pathway analysis for differentially methylated probes (DMPs) and regions (DMRs) associated with increasing mammographic density (continued)

Category	Term	N Genes	Genes	p	Fold Enrich -ment
Ingenuity Canonical Pathways	Synaptic Long Term Depression	8	*CACNA2D2,GNA14,ITPR1,PLCG2,PPP2R 2B,PRKCG,RALB,SLC39A7*	0.02	NA
Ingenuity Canonical Pathways	VDR/RXR Activation	6	*HES1,NCOR2,PRKCG,RUNX2,RXRB,TGFB 2*	0.02	NA
Ingenuity Canonical Pathways	Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	7	*IFIH1,IL1F10,LTB,NLRC4,PLCG2,PRKCG ,TGFB2*	0.02	NA
Ingenuity Canonical Pathways	Calcium-induced T Lymphocyte Apoptosis	5	*ATP2A1,HLA-DRB1,HLA-DRB5,ITPR1,PRKCG*	0.02	NA
Ingenuity Canonical Pathways	Aldosterone Signaling in Epithelial Cells	7	*DNAJB12,DNAJC25,HSCB,ITPR1,NEDD4, PLCG2,PRKCG*	0.02	NA
Ingenuity Canonical Pathways	Non-Small Cell Lung Cancer Signaling	5	*ABL1,BAD,ITPR1,RALB,RXRB*	0.02	NA
Ingenuity Canonical Pathways	Adrenomedullin signaling pathway	8	*BAD,GNA14,IL1F10,ITPR1,PLCG2,RALB, RXRB,SLC39A7*	0.02	NA
Ingenuity Canonical Pathways	T Cell Exhaustion Signaling Pathway	8	*HLA-DRB1,HLA-DRB5,IFNAR1,JAK3,PLCG2,PPP2R2B,RA LB,TCF7*	0.02	NA
Ingenuity Canonical Pathways	Dopamine-DARPP32 Feedback in cAMP Signaling	7	*ATP2A1,CAMKK1,GRIN2A,ITPR1,PLCG2, PPP2R2B,PRKCG*	0.02	NA
Ingenuity Canonical Pathways	Estrogen Receptor Signaling	10	*BAD,GNA14,HES1,JAK3,NCOR2,PLCG2,P RKAB1,PRKCG,RALB,RUNX2*	0.02	NA
Ingenuity Canonical Pathways	Systemic Lupus Erythematosus In B Cell Signaling Pathway	9	*BAD,IFIH1,IFNAR1,IL1F10,LTB,PLCG2,P RKCG,RALB,TGFB2*	0.02	NA
Ingenuity Canonical Pathways	Synaptic Long Term Potentiation	6	*GNA14,GRIN2A,ITPR1,PLCG2,PRKCG,RA LB*	0.02	NA
Ingenuity Canonical Pathways	Erythropoietin Signaling	7	*BAD,IL1F10,ITPR1,LTB,PRKCG,RALB,TG FB2*	0.02	NA
Ingenuity Canonical Pathways	PDGF Signaling	5	*ABL1,JAK3,PDGFD,PLCG2,RALB*	0.02	NA
Ingenuity Canonical Pathways	Human Embryonic Stem Cell Pluripotency	6	*NODAL,NTRK3,PDGFD,TCF4,TCF7,TGFB 2*	0.03	NA
Ingenuity Canonical Pathways	Transcriptional Regulatory Network in Embryonic Stem Cells	4	*CDX2,ISL1,MEIS1,PAX6*	0.03	NA
Ingenuity Canonical Pathways	Neuroinflammation Signaling Pathway	9	*GABBR2,GAD2,GRIN2A,HLA-DRB1,HLA-DRB5,JAK3,PLCG2,PRKCG,TGFB2*	0.03	NA
Ingenuity Canonical Pathways	Factors Promoting Cardiogenesis in Vertebrates	6	*NODAL,PLCG2,PRKCG,TCF4,TCF7,TGFB 2*	0.04	NA
Ingenuity Canonical Pathways	CREB Signaling in Neurons	13	*CACNA2D2,GABBR2,GNA14,GPR12,GPR 137,GRIN2A,ITPR1,NTRK3,PLCG2,PRKC G,RALB,TACR3,TGFB2*	0.04	NA
Ingenuity Canonical Pathways	Molecular Mechanisms of Cancer	10	*ABL1,BAD,CHEK2,GNA14,JAK3,PRKCG,R ALB,RHOH,TCF4,TGFB2*	0.046	NA
Ingenuity Canonical Pathways	Glioma Signaling	5	*ABL1,PDGFD,PLCG2,PRKCG,RALB*	0.046	NA
Ingenuity Canonical Pathways	Protein Ubiquitination Pathway	8	*ANAPC11,DNAJB12,DNAJC25,HSCB,NED D4,USP19,USP3,USP31*	0.046	NA

Table continues

Table A9. Results of gene ontology and pathway analysis for differentially methylated probes (DMPs) and regions (DMRs) associated with increasing mammographic density (continued)

Category	Term	N Genes	Genes	p	Fold Enrich-ment
Ingenuity Canonical Pathways	GPCR-Mediated Nutrient Sensing in Enteroendocrine Cells	5	*CACNA2D2,GNA14,ITPR1,PLCG2,PRKCG*	0.047	NA
Ingenuity Canonical Pathways	HMGB1 Signaling	6	*IL1F10,KAT2A,LTB,RALB,RHOH,TGFB2*	0.049	NA
DMRs					
GO: Biological Process	Visual learning	3	*FOXB1,MEIS2,GRIN2A*	0.003	38.60
GO: Biological Process	Single fertilization	3	*HOXA9,HOXA10,IZUMO1R*	0.004	29.44
GO: Cellular Component	Integral component of plasma membrane	8	*KL,PDGFRA,HLA-DRB1,MAEA,CSPG5,SLC6A12,GRIN2A,AQP5*	0.007	3.32
GO: Cellular Component	Transcription factor complex	3	*MED7,HOXA9,HOXA10*	0.04	9.14
GO: Molecular Function	Sequence-specific DNA binding	5	*HOXA9,HOXA10,FOXB1,MEIS2,MKX*	0.009	5.82
GO: Molecular Function	Ras guanyl-nucleotide exchange factor activity	3	*KL,PDGFRA,GRIN2A*	0.01	15.73
Ingenuity Canonical Pathways	Neuroinflammation Signaling Pathway	3	*GRIN2A,HLA-DRB1,SLC6A12*	0.01	NA

Appendix B: Supplemental Figures

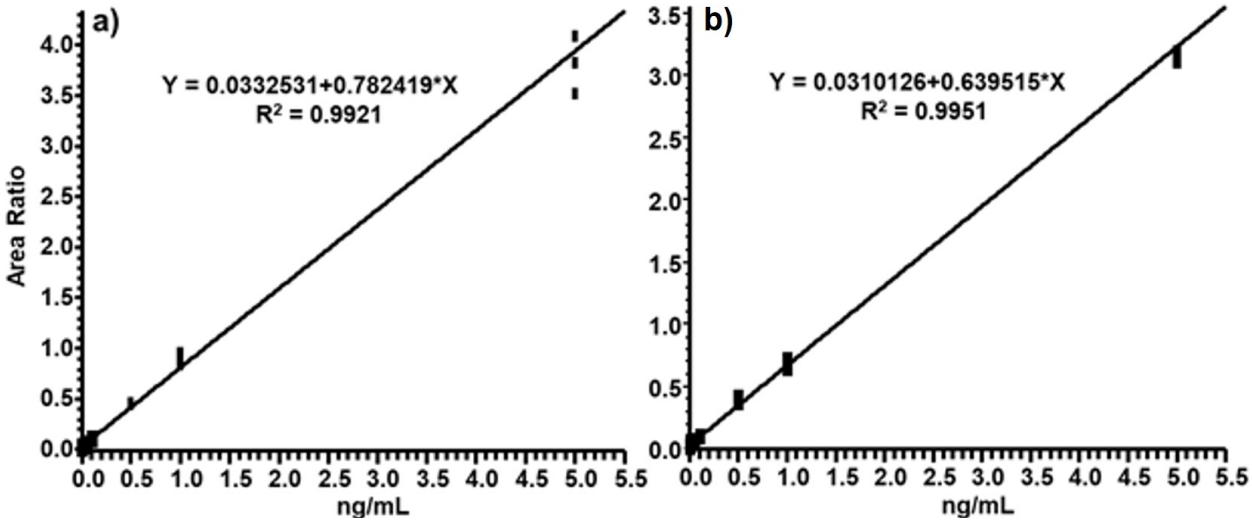

Figure B1. Calibration curve for a) glyphosate and b) AMPA ranging from 0 to 5 ng/mL in urine. The area ratio represents the ratio of variable concentration of glyphosate and AMPA (0 to 5 ng/mL) to their respective internal standards ($^{13}C_2{}^{15}N$-Glyphosate and $D_2{}^{13}C{}^{15}N$-AMPA) with a constant concentration of 6.25 ng/mL spiked in urine. AMPA: aminomethylphosphonic acid

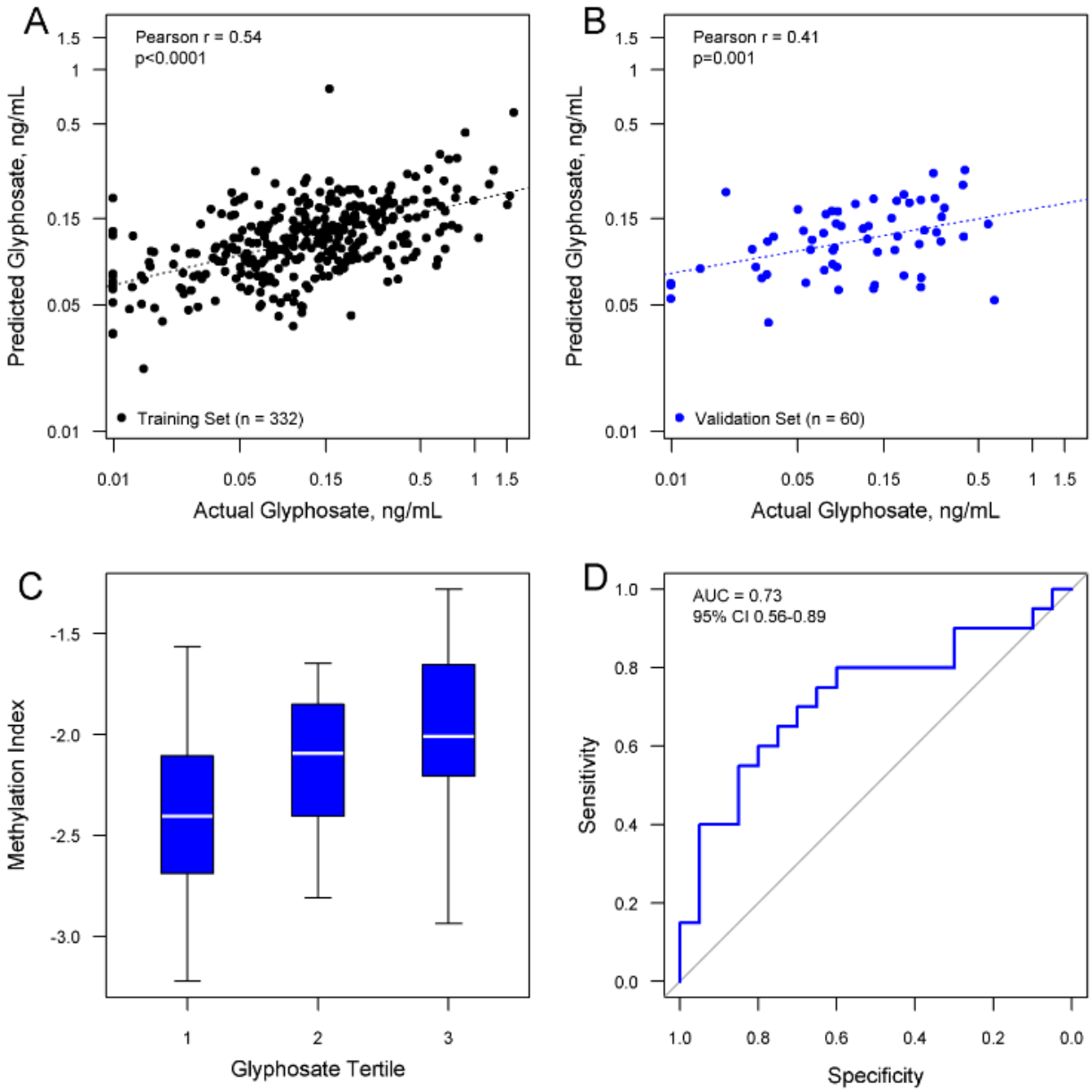

Figure B2. Performance of methylation index using 11 CpG sites on the 450k methylation array to predict the natural logarithm of urinary glyphosate concentration. Data shown for the training set (A) and the validation set (B, C). Panel D shows the classification performance of the methylation index in the validation set for classifying the top vs. the bottom tertile of urinary glyphosate.

Appendix C: Development of Subsampling Methodology

Replication is a major problem in epigenome-wide association studies. Research on other microarray-based studies suggest that results from a single study are highly sensitive to small differences in the study population[190–193]. In this work, I implemented a novel approach to aggregate results from a large number of subsamples of the main study population, in an effort to filter out spurious results and identify methylation differences that are stably associated with each outcome of interest.

Results from the four epigenome-wide association studies described in Chapters 3 and 4 highlight the necessity of this approach and show that it successfully reduces the number of false positive associations. The number of probes meeting the FDR significance threshold in a single subsample of the study population varied widely depending on the subsample (Table C1). For example, in the epigenome-wide association study of mammographic density, the number of FDR-significant probes ranged from 123 to 17,209 depending on subsample used. Furthermore, most of the FDR-significant probes were not robust to permutations in the study population. In the epigenome-wide association study of smoking, 3,443 probes were significantly associated with smoking in at least one subsample. However, 28.3% of these probes were significant in just one subsample of the study population and were not significant in the other 999, indicating that they are likely spurious results (Table C1). In fact, in all four studies, over 85% of the probes associated with the exposure of interest in at least one subsample were identified in less than 10% of the other subsamples.

Table C1. Results from epigenome-wide association studies on 1,000 subsamples of 90% of the study population

	Smoking	Mammographic Density	Glyphosate	AMPA
	Median (Minimum, Maximum)			
Number of FDR-significant probes per subsample	171 (23, 1,258)	1,596 (123, 17,209)	239.5 (11, 2,819)	47 (0, 483)
Probes by number of subsamples in which the probe was FDR-significant, N (%)				
At least one subsample	3,443	33,015	7,362	2,367
Exactly one subsample	976 (28.3%)	7,381 (22.4%)	2,318 (31.5%)	779 (32.9%)
<10% of subsamples	3,045 (88.4%)	28,356 (85.9%)	6,656 (90.4%)	2,193 (92.6%)

Two parameters may affect the stability of results from a resampling-based methylation analysis: τ, the frequency threshold which is used to determine the final list of significant probes or regions, and N, the number of subsamples used in the analysis. In the final analysis presented in this work, I used a value of 0.9 for τ, which includes all probes which were FDR-significant in ≥90% of subsamples as the final set of differentially methylated probes, and a value of 1,000 for N. Here I present data to support the use of these parameter values.

Because smoking has been extensively studied using traditional epigenome-wide approaches, it makes a useful positive control to test the claim that many of the results from a single subsample will be false positives. This study was not designed or powered to test associations between smoking and DNA methylation, so results that are inconsistent with previous epigenome-wide association studies of smoking are likely to be false positive results. In Chapter 4, I present data showing that 35/40 (87.5%) of probes associated with smoking with $\tau = 0.9$ had been associated with smoking in previous studies[14,15,246–249]. Figure C1 shows the proportion of probes previously associated with smoking for varying values of τ. For all probes associated with smoking in at least one subsample, over 50% had not been

associated with smoking in any previous study and may be false positives. The proportion of these purported false positives falls to 30.3% when limited to probes which were FDR-significant in over half of subsamples ($\tau = 0.5$), and steadily falls as the stringency of the threshold is increased. Figure C1 shows that a value of $\tau = 0.9$ ensures that most differentially methylated probes are likely to be true positive associations.

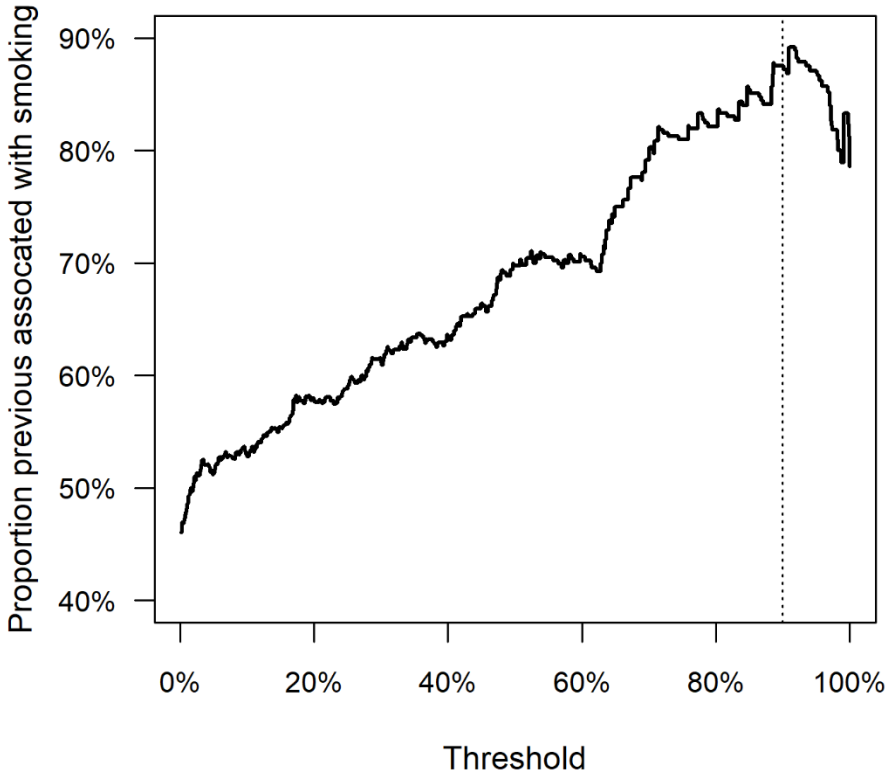

Figure C1. True positive rates at varying values of τ (frequency threshold). The dotted line shows $\tau = 0.9$, the final value used in the epigenome-wide association studies in this work.

An important consideration given the computational burden of repeated analyses is the number of subsamples to use (N). Figure C2 shows the number of differentially methylated probes for increasing number of subsamples in the four studies included in this work. For each study, the list of differentially methylated probes converges after approximately 200 subsamples. Further resampling does not appear to substantially alter

the final list of differentially methylated probes after N = 200. This suggests that the value of N = 1000 used in this work is likely in excess of what is required to achieve a stable list of differentially methylated probes and regions. Future analyses may reduce the computational time required by reducing the value of N.

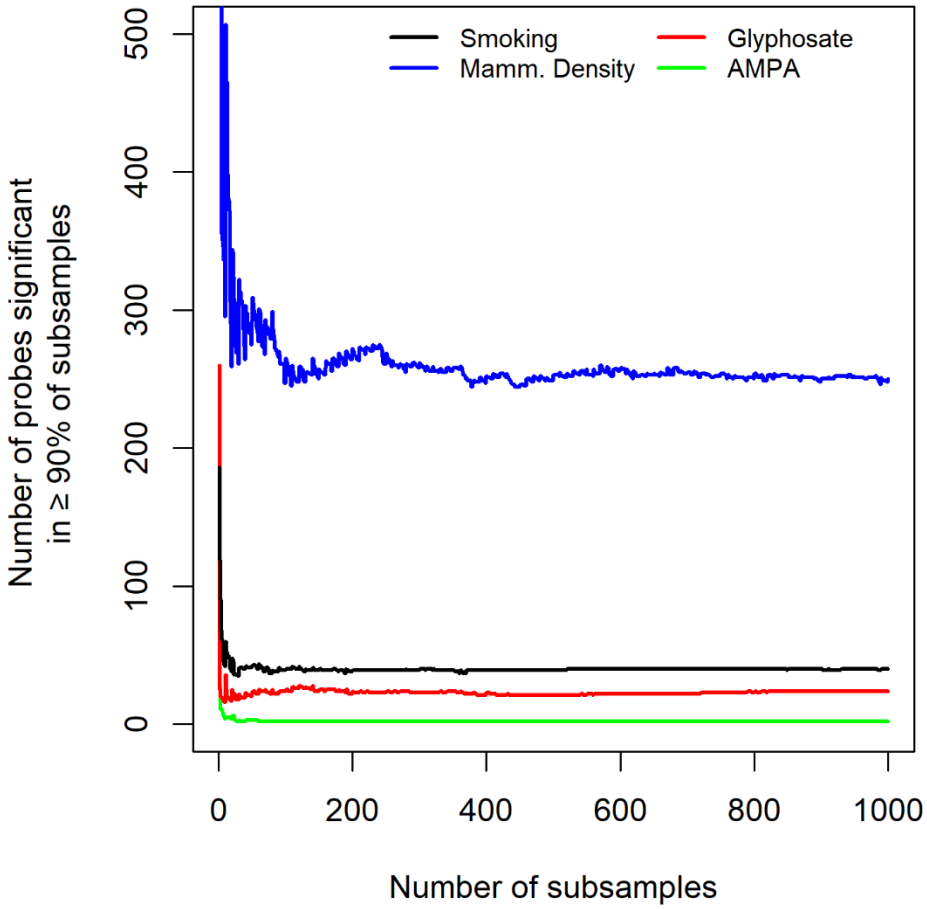

Figure C2. Convergence of list of differentially methylated probes for increasing number of subsamples (N).