Title
Differentiation of Exposure and Disease-Related Biomarkers Associated with Colorectal Cancer

Author
Perttula, Kelsi Michel

Publication Date
2017

Peer reviewed|Thesis/dissertation

Differentiation of Exposure and Disease-Related Biomarkers Associated with Colorectal Cancer

by

Kelsi Michel Perttula

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Health Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Stephen M. Rappaport, Chair

Professor S. Katharine Hammond

Professor Alan Hubbard

Fall 2017

Differentiation of Exposure and Disease-Related Biomarkers Associated with Colorectal Cancer

Abstract

Differentiating Biomarkers of Exposure and Disease

Associated with Colorectal Cancer

By

Kelsi Perttula

Doctor of Philosophy in Environmental Health Sciences

University of California, Berkeley

Professor Stephen M. Rappaport, Chair

Chronic diseases such as cardiovascular disease, diabetes, and cancer are the leading causes of death among developed and developing countries, and account for approximately 75 percent of deaths worldwide. With the sequencing of the human genome and subsequent genomic studies, we now know genetic factors alone are responsible for a relatively small portion of these diseases. Specifically, cancer risk attributed to genetic factors is typically about eight percent. Thus, the vast majority of cancer risk likely lies within the realm of exposures (non-genetic factors) or a combination of genetic factors and exposures. The collection of exposures over an individual's lifetime comprise the concept of the *exposome*, an epidemiological complement to the genome. The exposome is defined by measurement of both endogenous (inflammation, lipid peroxidation, microbiota) and exogenous (air pollutants, pesticides, drugs, diet, etc.) exposures within an individual.

Much exposure data is from non-individualized sources, such as air quality monitors or other spatial-temporal data, which have limited use in epidemiology. Individual exposure assessment consists largely of self-reported dietary and lifestyle data from interviews or questionnaires. In recent years, advances in analytical chemistry have permitted the simultaneous detection of thousands of molecules in biological fluids including urine, whole blood, plasma, and serum.

High resolution liquid chromatography mass spectrometry (LCMS) is a powerful technique to measure the accurate masses of molecules in biological fluids for high-throughput epidemiological studies. Chapter 1 of this dissertation details a method for the analysis of lipophilic molecules in plasma using specimens from 158 healthy volunteer subjects. The resulting data revealed levels of lipids and other molecules that differed between smoking and nonsmoking, white and black, and male and female subjects. A modified version of this LCMS method was used in the analysis of serum from subjects in a nested case-control study, described in Chapters 2 and 3.

Colorectal cancer (CRC) accounts for one fourth of all cancer deaths worldwide and less than about 15 percent of CRC risk is attributable to genetic factors alone. To investigate possible influences of exposures on CRC risk, serum from 190 subjects in the European Prospective Investigation of Cancer and Nutrition (EPIC) were extracted for lipophilic molecules and

analyzed with high resolution LCMS. These prospective samples – collected up to 22 years prior to diagnosis - offered a unique opportunity to differentiate between CRC biomarkers related to disease causes and those that result from disease progression. Chapter 2 describes the testing of one class of lipids, ultra-long chain fatty acids (ULCFAs), that had been reported as a probable protective factor of CRC. Paired case-control differences were assessed with respect to the time period from when the serum was collected (study enrollment) to when the case was diagnosed. Since, case-control differences decreased with increasing time prior to case diagnosis, ULCFAs were likely depleted by cancer progression rather than by protective exposures.

Many of the features in LCMS profiling are unannotated (identity unknown) chemicals. Rather than relying on hypothesis-driven analyses of only known compounds, data-driven analyses of reliably detectable features can result in the generation of new hypotheses of possible disease-causing exposures. This untargeted methodology, used in Chapters 1 and 3, makes lipidomic and other exposure-related profiling a powerful tool in exposure assessment. In Chapter 3, the untargeted analysis of features in EPIC CRC serum samples revealed potentially relevant molecules associated with CRC causes and disease progression. As opposed to traditional *p*-value-centric analysis used in Chapters 1 and 2, a combination of regularized regression, random forest, and t-tests were employed in the feature selection for this untargeted analysis.

In Chapter 4, the lipophilic data from the healthy volunteer samples of Chapter 1 are studied once again. Using a method similar to the regularized regression technique described in Chapter 3, we determined which lipids were associated with levels of adductomic biomarkers (another methodology developed in our laboratory), which had also been measured in plasma from the same healthy volunteers. Analysis of the combined data from these two OMIC datasets found interesting correlations between particular lipids and adducts in these samples.

Dedication

This dissertation is dedicated to the amazing family that and grew during the procurement of this degree. My husband and best friend Drew, our sons Ari and Asher, and our sweet dog Apollo. I am eternally grateful for them.

Table of Contents.

List of Figures

List of Tables

Acknowledgments

I could not have completed this dissertation without the support and guidance of countless fellow students, professors, and staff within Environmental Health Sciences and the greater School of Public Health (SPH). Many friends, family members, and others I may have inadvertently omitted.

This work emanates from the work of my brilliant advisor, Stephen Rappaport. He has not only dedicated his career to studying the relationships between our environment and public health, but has contributed new conceptual and methodological tools to define and measure our exposures in association to disease onset. With his long academic and scientific career, he has never been short on advice or ideas to navigate around any challenge. Laboratory work never goes perfectly, but when things did not go as planned, he was instrumental in helping me find my way to the best solution. I am forever grateful for the incredible opportunity to conduct research using the innovative techniques developed in the Rappaport lab, and for the continuous funding throughout my program.

The other members of my dissertation and qualifying exam committees were endlessly helpful and kind. I succeeded only with the mentoring and enormous support from Kathie Hammond. Her compassion and expertise with the dissertation writing process kept me on track. Her wisdom and experience were fortifying.

Alan Hubbard's expertise in biostatistics was essential for the analyses in the first two chapters. His thoughtful encouragement as a committee member also made the completion of this process go more smoothly.

Martyn Smith, another mentor as well as qualifying exam committee member, challenged me to explain and defend my research under pressure, and provided critical feedback and support throughout this process. I know these experiences made me a better scientist.

Chris Vulpe, also a qualifying exam committee member, inadvertently led me to this program before I was a student. His molecular toxicology classes from over a dozen years ago inspired me to combine my affinity and knowledge of chemistry with applications to human health.

Charlotte Smith gave me wonderful opportunities to assist in the instruction of two SPH courses. The experiences in these classes were among my favorite moments during my graduate student career. Her kindness and guidance have been invaluable during this final year of my PhD program.

Navigating the administrative aspects of this process would have been much more daunting were it not for the helpful and organized Norma Firestone. She gave support, information, and considerate support throughout my entire time in the SPH. She enriches our experiences with the events and programs she organizes and runs for all EHS PhD students.

Our European collaborators were critical for the front-end part of the prospective colorectal cancer study. With access to rare prospective biospecimens from the European Prospective Investigation into Cancer and Nutrition, we were afforded this exceptional opportunity. I am very grateful to Paolo Vineis, Silvia Polidoro, Marc Gunter and Alessio Naccarati for the chance to do this work and for their long-distance feedback.

I am very grateful to chemistry and QB3 faculty Evan Williams as well as QB3 personnel Anthony Iavarone and Rita Nichiporuk. Their outside perspectives and assistance with my mass spectrometry challenges were always refreshing and helpful.

We would not have had the mass spectrometry instrumentation to do much of this work without the generosity and help of Agilent Technologies. I give special thanks to Marcus Miller, whose assistance and expertise was critical and much appreciated on many occasions, and included responsive service on weekends and holidays.

An especially wonderful thing about academic life at UC Berkeley is the incredible abundance of colleagues whose brilliance will always astound me:

Xiaoming Cai was the perfect mentor for me during my first year of graduate school. Her kindness and mass spectrometry expertise were beyond my wildest expectations, and she remains a role model and dear friend.

Lauren Petrick, despite usually working on different projects, always found ways to help everyone in the lab, and she inspires me with her incredible intellect and work practices. Her mentorship and friendship throughout my second, third, and fourth years were critical to my experience in this program. As a new professor, I know she will continue to make great contributions to the understanding of environmental factors on disease and public health and I will continue to rely on her as an amazing friend and fellow scientist.

I gratefully acknowledge the profound contribution of the late Will Edmands, first our post-doctorate, then a scientist in our lab. I will forever remember our collaboration and the support I got during the long days and weeks during our multiple data collections throughout my second and third years. His creation and implementation of automating and data analysis tools enabled our lab to reach new high-throughput analytical abilities. His brilliance in bioinformatics and programming leaves an indelible mark on the world, with the useful R packages he wrote that assists my lab-mates and I with our data analysis. I will also remember his wry and quick wit that would often make me laugh.

Courtney Schiffman, our laboratory's congenial and brilliant biostatistician, has a unique ability to combine her detailed understanding of bioinformatic "-OMIC" methods with her profound knowledge in statistics and statistical programming. We are always confident in the quality of data analysis and consultation she and her advisor in Biostatistics, Sandrine Dudoit, provide.

I was fortunate enough to be in a lab that was and is filled with incredibly talented scientists. In addition to Xiaoming, Lauren, Will, and Courtney, my labmates past and present include Hasmik Grigoryan, Luca Regazzoni, Katie Hall, Samantha Lu, Yukiko Yano, Sa Liu, and Henrik Carlsson. All of my colleagues provided profound technical expertise, comradery and/or creative suggestions each step of the way.

The recent EHS PhD graduate Sarah Daniels provided endless support and love during critical moments of pregnancy, childbirth, and illness. I cannot imagine having survived these milestones like these without her. I have never met someone with a bottomless ability to give to her community, friends, and family. My family and I are so grateful for having her in our lives.

Longtime dear friends, especially Holly Rivlin, and fellow PhD students Lauren Bausch, Aviva Goldmann, and Rachel Ruderman gave me enormous comfort with their decades of friendship. The sustained friendship and love from the Orduña family lent help wherever they could. My family and I are so lucky to have these lifelong friends.

I fondly look back on the handful of teachers and mentors who encouraged an early interest and scholarship in science. Namely, my fifth-grade teacher Bonnie Buss (who encouraged my learning and understanding of the periodic table of elements), my seventh-grade zoology teacher Mr. "Rocky" Rothschiller (tough classes and early morning dissections), my 10^{th}-11^{th} grade teacher, the late Kathleen Rose (who always made me feel loved, smart, and important), and my grandpa Herb Hooper (dentist, and other family scientist who quizzed me on said elements). Their inspiration throughout my childhood and adolescence were fueled my desire to go as far as I can with my work.

To Diane Garcia-Gonzales, through good, great, and trying times, I could not be more fortunate to have you in my cohort from day one. Supporting each other was an essential component of my graduate life experience. With three (usually non-ideal) childbirths, family illnesses, two giant moves, and one marriage ceremony, our lives have transformed over our time as PhD students. But your limitless generosity, caring, epidemiology knowledge, and love remained constant.

I am grateful for my extended family. Aunts, uncles, great-grandparents, cousins, and more whose energy, friendship, and devotion propped me up, especially in difficult moments. I've been fortunate enough to have immense parent and grandparent love and support throughout this experience; my children could not have better grandparents. The last 20 months of work would not have been possible without the talent and love from our dear Nana Lila. I will eternally be in awe of and thankful for my perfectly kind-hearted and loving husband Drew. I could not be prouder of our sons, Ari and Asher. "If we lay a strong enough foundation; we'll pass it on to you, we'll give the world to you; and you'll blow us all away" – Lin-Manuel Miranda

Chapter 1: Untargeted lipidomic profiling of human plasma reveals differences due to race, gender and smoking status

Xiaoming Cai, Kelsi Perttula, Sara Kherad Pajouh, Alan Hubbard, Daniel K. Nomura, and Stephen M. Rappaport

University of California, Berkeley; School of Public Health; Program in Environmental Health Sciences

A similar version of this manuscript has been published: Cai X, Perttula K, Pajouh SK, Hubbard A, Nomura DK, et al. (2014) Untargeted Lipidomic Profiling of Human Plasma Reveals Differences due to Race, Gender and Smoking Status. Metabolomics 4:131. doi:10.4172/2153-0769.1000131

1.1 Abstract

Lipidomic profiling can link genetic factors and exposures to risks of chronic diseases. Using untargeted liquid chromatography-Fourier Transform mass spectrometry (LC-FTMS), we explored differences in 3,579 lipidomic features in human plasma from 158 subjects, pooled separately by race, gender and smoking status. Significant associations with race (23 features), smoking status (9 features) and gender (2 features) were detected with analysis of variance (ANOVA)-based permutation tests. Identities of several features were confirmed as plasmalogens (vinyl-ether phospholipids) that were present at 2-fold greater concentrations in black subjects. Putative assignments of other features, based on accurate masses, were more abundant in white subjects, namely, dihomo-γ-linolenoyl ethanolamide (DGLEA), an endogenous endocannabinoid receptor agonist and phosphatidyl choline [PC(16:0/18:1)]. After adjustment for race, multivariable linear regression models showed that gender was significantly associated with levels of plasmalogens and DGLEA and that consumption of animal fat was marginally associated with concentrations of plasmalogens. Interestingly, BMI did not explain additional variability in any race-adjusted model. Since plasmalogens are antioxidants that are generally regarded as health-promoting and DGLEA is an agonist of the cannabinoid receptor, our findings that these molecules differ substantially between black and white Americans and between men and women, could have health implications. The concentration of cotinine was greatly elevated in smoking subjects and 6 features with *m/z* values suggestive of phospholipids or sphingomyelins were present at significantly lower concentrations in smokers.

1.2 Introduction

Since lipids are essential to functional membranes, energy storage and signaling [1,2], lipidomics provides an avenue for linking important biological processes with disease states. Indeed, differences in lipid profiles have been reported in investigations of cancer [3–6], diabetes [7], Alzheimer's disease, [8,9] and cardiovascular disease [10,11]. Such studies increasingly rely on high-resolution mass spectrometry (MS) platforms that can detect thousands of lipidomic features in plasma while simultaneously providing accurate masses for annotation [12].

Given strong associations between blood lipid levels and chronic diseases, it is surprising that baseline lipidomic profiles have not been reported across fundamental population characteristics such as race and gender as well as lifestyle factors such as smoking. Here, we used untargeted Fourier Transform (FT) MS to obtain lipidomic profiles containing over 3,000 features detected in plasma from healthy American subjects stratified by race (black and white), gender and smoking status. Race was the strongest classifying factor (23 significant features) followed by smoking status (9 features) and gender (2 features). Identities assigned to race-discriminating features included several plasmalogens (ether phospholipids containing fatty alcohols with vinyl-ether linkages in the sn-1 position and fatty acids with ester linkages in the sn-2 position) that were more abundant in black subjects. Tentative assignments, based on accurate masses, pointed to greater concentrations in white subjects of an endogenous endocannabinoid receptor agonist and a phosphatidyl choline. Several unidentified features, with masses suggestive of phospholipids or sphingomyelins, were present at lower concentrations in smoking subjects. Since all of these lipids are physiologically important and some have been associated with chronic diseases, our results suggest that young American adults may be predisposed to diseases because of differing lipid concentrations associated with race, gender and smoking.

1.3 Materials and Methods

Reagents
Isopropanol, methanol, chloroform, formic acid, ammonium hydroxide and ammonium formate (10 M, pH7.4) were from Fisher Scientific. Phosphate buffer saline (pH 7.4) was from Invitrogen. Lipid standards of 1-octadecenyl-2-(5Z,8Z,11Z,14Z-eicosatetraenoyl)-sn-glycero-3-phosphoethanolamine PE(P-18:0/20:4) and 1-octadecenyl-2-(4Z,7Z,10Z,13Z,16Z,19Z-docosahexaenoyl)-sn-glycero-3-phosphocholine PC(18:0/22:6) were purchased from Avanti Polar Lipids (Alabaster, AL). Water was purified by a Milli-Q Gradient ultrapure water purification system (Millipore, Billerica, MA). All other chemicals were of analytical grade and used without purification.

Lipid nomenclature
Lipids were named according to Lipid Maps (http://www.lipidmaps.org); e.g. 1-hexadecanoyl-2-octadecenoyl-*sn*-glycero-3-phosphocholine is designated PC(16:0/18:1) and 1-hexadecenyl-2-eicosatetraenoyl-glycero-3-phosphoethanolamine is PE(P-16:0/20:4). When the fatty acid chain could not be determined, the total number of carbons and double bonds of all fatty acyl chains are given, e.g. PE(38:4).

Plasma samples
Blood samples were obtained in heparin from 158 healthy subjects (78 males and 80 females), representing a subset from a previous study conducted by the corresponding author under an approved human-subjects protocol [13]. Within a few hours of collection, plasma was separated from red blood cells by centrifugation. Red cells were washed with an equal volume of PBS, which was added to the plasma and thus reduced plasma concentrations. Plasma samples were frozen and stored at -80 °C for approximately 13 y prior to being aliquoted and pooled by combining aliquots from 4 to 6 subjects stratified by race, gender and smoking status. (Pooling was required by our institutional review board to ensure anonymity of subjects). A quality control sample was prepared by pooling 100 l of each of these 35 pooled samples.

Demographics, smoking and dietary assessment
Demographic characteristics, including race, age, height and weight were obtained with a standardized questionnaire at the time of phlebotomy. Smoking status was based upon current smoking (yes/no). A semi-quantitative food-frequency questionnaire containing 131 items was used to evaluate average daily consumption of fat (animal, vegetable and cholesterol) over the past six months for each individual [14,15]. All dietary-intake values were compiled at the Channing Laboratory, Harvard Medical School [16,17].

Extraction of lipids
Lipids were extracted according as described previously [18]. Briefly, 100 μl of plasma was thawed on ice and then mixed with 3 ml of chloroform:methanol (2:1,v/v) and 900 μl of phosphate buffered saline (PBS). After vortexing, the mixture was centrifuged at 2000×g for 5 min. The bottom layer was collected, dried under N_2, and dissolved in 100 μl chloroform. Extracts were stored at -80 °C before LC-MS analysis.

LC-MS analysis
Liquid chromatography-MS analysis was performed with a Surveyor LC system coupled to an LTQ-FTMS, containing a heated electrospray ionization source (ESI) (Thermo Fisher Scientific,

Waltham, MA). The MS was operated in both ESI+ and ESI- ionization modes with data collected from *m/z* 100 to 1200. For LC separation, a Luna C5 column (4.6×50 mm, 100 Å, 5 μm, Phenomenex, Los Angeles, CA) was selected with column and autosampler temperatures maintained at 25 °C and 4 °C, respectively. The C5 column was selected to elute all potential lipids in the samples, including hydrophobic triacylglycerides and cholesterol esters. Injection volumes were 20 μl and 25 μl for ESI+ ionization and ESI- ionization, respectively. Mobile phases contained 0.1% formic acid for ESI+ ionization and 0.1% ammonium hydroxide for ESI- ionization. The column was eluted with a gradient of mobile phase A (methanol:50 mM ammonium formate 5:95) and mobile phase B (isopropanol:methanol:50 mM ammonium formate 60:35:5) as follows: 100%A for 5 min at 0.1 ml/min; 0-100%B over 15 min at 0.4 ml/min; 100%B for 5 min at 0.5 ml/min; 0-100%A for 5 min at 0.4 ml/min. Blank and QC samples were analyzed after 7 or 8 experimental samples to wash the column and monitor stability.

The MS was tuned with the following high-abundance lipids in several structural classes: tuning in positive mode employed LPC(16:0), LPE(18:1),PC(36:4),PE(34:1) and TG(58:5) and tuning in negative mode employed FA(16:0), FA(20:4), PI(24:1) and PG(34:1). Several FTMS parameters, namely, mass resolution, maximum injection time, and maximum number of ions collected for each scan, were optimized for sensitivity while maintaining a mass resolution of 100,000. The following settings were used: vaporizer temperature, 280 °C; sheath and auxiliary gases, 35 and 15 (arbitrary units); spray voltage, 3.5 kV; capillary temperature, 350 °C; capillary voltage, 10 V; tube-lens voltage, 120 V; maximum injection time, 1000 ms; maximum number of ions collected for each scan, 5×10^5. Mass calibration was carried out with a standard LTQ calibration mixture (Thermo Scientific, Waltham, MA). For untargeted analyses, a full scan was used for the FTMS with a mass resolution of 100,000, and data were recorded in centroid mode. To study structures of discriminating features, tandem MS/MS analyses were performed with the linear ion trap in low-resolution mode with a CID voltage of 30 V. Accurate masses were calculated using the lipid calculator (http://pharmacology.ucdenver.edu/lipidcalc/) and then extracted with a mass tolerance of 10 ppm in the total ion chromatogram (TIC).

Quantitation of analytes

Because PBS had been added to plasma (as erythrocyte washes) at the time of phlebotomy, volumes of diluted plasma varied across the pooled samples in our investigation. Thus, rather than quantifying peaks of unknown lipidomic features relative to internal standards, quantitation was based on dividing each peak intensity by the sum of all peak intensities detected in each pooled sample [19,20].

Data collection and processing

Data were collected continuously over the 30-min LC separation using Xcalibur software (Thermo Fisher Scientific). The raw data were converted to mzXML data format using proteoWizard software (Spielberg Family Center for Applied Proteomics, Los Angeles, CA). Peak detection, retention time collection and alignment were processed on the XCMS platform (http://xcmsserver.nutr.berkeley.edu/). All data-collection parameters were set to the "HPLC Orbitrap" default values (centwave feature detection, loess non-linear retention time alignment, 0.5 minimum fraction of samples in one group to be a valid group, *P*-value thresholds = 0.05, isotopic ppm error = 5, *m/z* absolute error = 0.015) except the following: maximal tolerated *m/z* deviation in consecutive scans = 3.5 ppm; width of overlapping *m/z* slices (mzwid) = 0.005;

retention time window (bw) = 15 s, minimum peak width = 20, maximum peak width = 80. Lists of retention times (RT), *m/z* values and peak intensities were exported to an Excel spreadsheet for processing. As noted previously, the intensity of each peak was normalized to the sum of total intensities in each sample and was then multiplied by 10,000 for statistical analysis.

Characteristics of lipid classes under our instrumental parameters were determined from LC-FTMS of a training set consisting of 193 common lipid species representing monoacylglycerols (MG), diacylglycerols (DG), triacylglycerols (TG), glycerophosphocholines (PC), glycerophosphoethanolamines (PE), monoglycerophosphocholines (LPC), monoglycerophosphoethanolamines (LPE), sphingomyelin (SM), and cholesterol esters (CE). Mass accuracy, precision, and stability of the method were estimated from repeated analysis of 8 ion peaks representing lipids detected in the quality control sample that covered large ranges of masses, intensities and retention times. Mass accuracies were less than 6 ppm and coefficients of variation of retention times and peak intensities were 0.10%-0.56% and 4.08%-24.47%, respectively.

Statistical analysis

Because plasma samples were pooled for the current investigation (4 - 6 subjects per pooled sample) mean values were used for statistical analyses. A combination of univariate and multivariate statistical models was used to investigate discriminating features. First, two-tailed Student's *t*-tests and analysis of variance (ANOVA) were performed to screen for discriminating features by race, gender and smoking status. Then significance was determined using a non-parametric permutation test with 10,000 observations [21]. False discovery rates (FDR) were corrected using the Benjamini-Hochberg (BH) method to adjust *P*-values for false discovery involving multiple comparisons [22]. After application of the BH method, 34 significant features were detected.

After putative identification of discriminating features (described below), sources of variation of dihomo-γ-linolenoyl ethanolamide (DGLEA), PC(16:0/18:1), and the sum of 6 plasmalogens [PE(P-16:0/20:4), PE(P-18:1/20:4), PE(P-18:0/20:4), PE(P-18:0/22:6), PE(P-18:0/22:5) and PC(P-18:0/22:6)] were evaluated with multivariable linear models that employed combinations of race, gender, smoking status, BMI and dietary fat (g) as predictor variables. (The sum of plasmalogen levels was used because concentrations of the 6 plasmalogens were highly correlated). Models were constructed using SAS software for Windows (v. 9.3, SAS Institute, Cary, NC).

Structural identification of discriminating features

Preliminary identification relied upon matching accurate masses from FTMS (with a mass tolerance of 10 ppm) with entries in the Human Metabolome Database (HMDB) (http://www.hmdb.ca/), the Structure Database of Lipid Maps (LMSD) (http://www.lipidmaps.org) and the Metabolite and Tandem MS Database (METLIN) (http://metlin.scripps.edu/). Since human plasma rarely contains lipids with odd-numbered fatty acyl chains, matches representing odd-numbered acyl chains were removed. Other filtering rules were constructed based on relative abundances of signals representing molecular ions and their common adducts, as determined from analyses of our training set of 193 lipid species. Additional structural information was derived from MS/MS analysis and comparisons with reference standards.

1.4 Results

Univariate analyses of demographic characteristics and dietary fat
Table 1.1 lists summary statistics for the subjects represented by the 35 pooled plasma samples in the current investigation. Subjects were young, with mean ages of 26 y and 25 y for black and white participants, respectively. The mean BMI for black subjects (28.9 kg/m^2) was significantly greater than that of white subjects (24.1 kg/m^2) and black subjects had significantly higher consumption of all forms of fat. Also, smokers consumed significantly more dietary fats than nonsmokers.

Profiling of plasma lipids

Untargeted lipidomic profiles of the 35 pooled plasma samples and QC samples were obtained by LC-FTMS in both ESI+ and ESI- modes. Many more features were detected in ESI+ mode (*n*=2,862) than in ESI- mode (*n*=717). Figure 1.1A represents a typical base-peak chromatogram of a lipid extract in ESI+ mode. Plasma lipids were mainly located in three time domains: 16-20 min, 20.5-23 min, 24-26 min. The averaged mass spectra of these three time domains, shown in Figure 1.1C, display prominent *m/z* ranges of 300-550, 700-820 and 800-910, respectively. The distribution of *m/z* is partially annotated in a density map (Figure 1.1B) which shows time domains of major lipid classes eluting between 19 and 26 min. Characteristics of these lipid classes were inferred from 193 lipid molecules in the training set. Masses of PCs, PEs, SMs, LPCs, and LPEs were mainly detected as [M+H]+, while MGs, DGs, TGs and CEs were detected as [M+NH_4]+.

Discriminating lipidomic features

To screen for differences associated with race, gender and smoking status, ANOVA models were obtained for each *m/z* feature (2,862 in ESI+ mode and 717 in ESI- mode) and random permutation tests were performed to establish *P*-values. The significance of each feature for a given comparison was determined by its *P*-value after BH correction for false discovery (*P*-values were truncated at 10^{-8}). As summarized in Table 1.2 and Figure 1.2, a total of 34 discriminating features was detected, namely, 23 for race, 9 for smoking status and 2 for gender. These features were concentrated in the *m/z* region between 650 Da and 850 Da, which is the domain of phospholipids and sphingomyelins (Figure 1.1).

Sixteen of these features were putatively identified by combinations of accurate mass, retention time, MS/MS fragment ions and reference standards (details are given in Supplemental Information, Section 2). The sole non-lipid feature was identified as cotinine (*m/z* 177.10246), a metabolite of nicotine that was 262 times more abundant in smokers than in nonsmokers. The other tentatively identified features were all lipids that significantly discriminated for race. These race-discriminating lipids included 6 plasmalogens [PE(P-16:0/20:4) [*m/z* 724.53126 [M+H]+]; PE(P-18:1/20:4) [*m/z* 750.54676 [M+H]+, 751.55602 [M+H]+ (isotope)]; PE(P-18:0/20:4) [*m/z* 752.56303 [M+H]+, 750.54772 [M-H]-, 753.56640 [M+H]+ (isotope), 754.58059 [M+H]+ (isotope)]; PE(P-18:0/22:6) [*m/z* 776.56270 [M+H]+]; PE(P-18:0/22:5) [*m/z* 778.57958 [M+H]+, 776.56369 [M-H]-]; and PC(P-18:0/22:6) [*m/z* 818.61083 [M+H]+] that were present at approximately 2-fold greater concentrations in black subjects. Two other race-related features were tentatively identified from accurate masses, namely, DGLEA, an agonist of the endocannabinoid receptors (CB1 and CB2) [23–25] that was present at 4-fold greater

concentrations in white subjects and PC(16:0/18:1) [*m/z* 744.55806 [M-CH3]-, 745.56145 [M-CH3]- (isotope), *m/z* 746.56465 [M-CH3]- (isotope)] that was present at moderately higher concentrations in white subjects (1.23-fold difference).

Multivariable linear regression models were used to investigate whether levels of the race-related lipids were affected by gender, BMI or consumption of fat as recorded by 6-month dietary recall. According to the R^2 values of the regression models (Table 1.3), race accounted for 68.0% of the summed plasmalogen levels, gender for 6.1% and consumption of animal fat for 2.2% (vegetable fat was not associated with plasmalogen levels). The model for putative DGLEA showed that race accounted for 50% of the variation and that race, gender and their interaction jointly explained 62%, with white males having 7 times higher concentrations than black females. Race was the only significant predictor for putative PC(16:0/18:1) and explained 45% of the variance. With race in each model, BMI did not significantly contribute to explained variability.

1.5 Discussion

Using untargeted lipidomics with LC-FTMS, plasma lipid changes related to race, gender, and smoking status were detected in healthy young American adults. The fact that baseline concentrations of these lipids differ between racial groups could be relevant to interpretation of findings that chronic diseases are more prevalent in black Americans than white Americans [26–28].

Most of the race-discriminating lipids were plasmalogens that were present at 2-fold higher levels in black subjects. Plasmalogens are required for membrane integrity and messaging [29,30] and serve as free radical scavengers [6,31,32]. Thus, these lipids are generally regarded as health promoting and several plasmalogens were recently detected at significantly lower concentrations in subjects with pancreatic cancer than in control subjects [33]. On the other hand, some oxidation products of plasmalogens can be toxic [34–37]. Since animal fat is the major source of plasmalogens in Western diets [38], the observed differences could reflect higher dietary intake of animal fat in black and male subjects (Table 1.1). Indeed, dietary consumption of plasmalogens increased plasma levels of these lipids in rats [38]. Self-reported consumption of animal fat (but not vegetable fat) explained a small amount of the variability of plasmalogen concentrations in our subjects (Table 1.3) after adjusting for race and gender. The fact that race was a much stronger predictor of plasmalogen levels than animal fat in our study could point to imprecision in dietary assessment of fat consumption and from aggregation of subjects by race/gender pooling. Higher plasmalogen levels in black and male subjects could also point to differential plasmalogen biosynthesis, possibly related to peroxisome activity [29,30]. Although BMI was significantly greater in black subjects, it is noteworthy that BMI did not explain additional variability of identified features after adjustment for race and gender in multivariable models.

Putative DGLEA, which was found at 4-fold higher concentrations in white subjects, is an endocannabinoid that binds to receptors (CB1 or CB2) that are also the targets of tetrahydrocannabinol, the principal active component of marijuana [23]. Upon activation of at least one of these receptors, specific physiological short-range events can be triggered, including neurotransmitter release. Effects of these reactions include analgesia, increased appetite and neural tissue development [39]. Although the endocannabinoid system and its effects are not

well understood, disruption of this system has been implicated in metabolic syndrome and accumulation of excess visceral fat [40]. Since the corresponding acid (DGLA) has been shown to have minimal differences across racial groups [41], a differentiating event may occur in the pathway between DGLA and the transformation to an ethanolamide.

Our untargeted lipid profiling discovered 6 features in the mass range between 740 and 790 Da that were present at lower concentrations in smoking subjects. Since this mass range is characteristic of phospholipids or sphingomyelins, our results lend support to the hypothesis that smoking interferes with metabolism of these lipid classes as suggested by targeted profiling of serum samples from smokers and non-smokers by Wang-Sattler *et al*. [42]. Interestingly, we also found that the level of the PC plasmalogen, PC(P-18:0/22:6), which was associated with race in our study, was approximately 33% lower in smokers, compared with non-smokers, consistent with the targeted study [42]. The two features with *m/z* 567.38180 and 568.38515 (Table 1.2) were highly correlated with cotinine (Spearman r=0.928, 0.948), suggesting that they are metabolites or reaction products of tobacco.

Because we used archived plasma from a previous investigation [13], our study has several limitations. First, it was necessary to pool the specimens - and thereby anonymize subjects' identities - while retaining testable factors (race, gender and smoking status). Although pooling is generally undesirable for small studies and could have reduced our ability to detect significant differences in population characteristics, those features that differed between races and genders (DGLEA and plasmalogens) are unlikely to be false positives [43]. Second, the blood sampling protocol employed heparinized plasma, and differences in concentrations of numerous lipids have been observed across blood samples collected with different anticoagulants, including heparin [44]. Thus, the 6 plasmalogens and putative DGLEA and PC(16:0/18:1) should be interpreted as lipidomic features that differed significantly between races and genders in samples of serum obtained from heparinized whole blood after prolonged storage at -80ºC. Finally, as noted previously, archived plasma samples from the 158 individual subjects in the original investigation had been diluted with varying volumes of erythrocyte washes. This effectively precluded quantitation based on internal standards and motivated us to normalize individual features by the sum of all detected peaks. While this method of quantitation could also have reduced precision – and the ability to detect discriminating features - it should not have generated false positives.

1.6 Tables and Figures

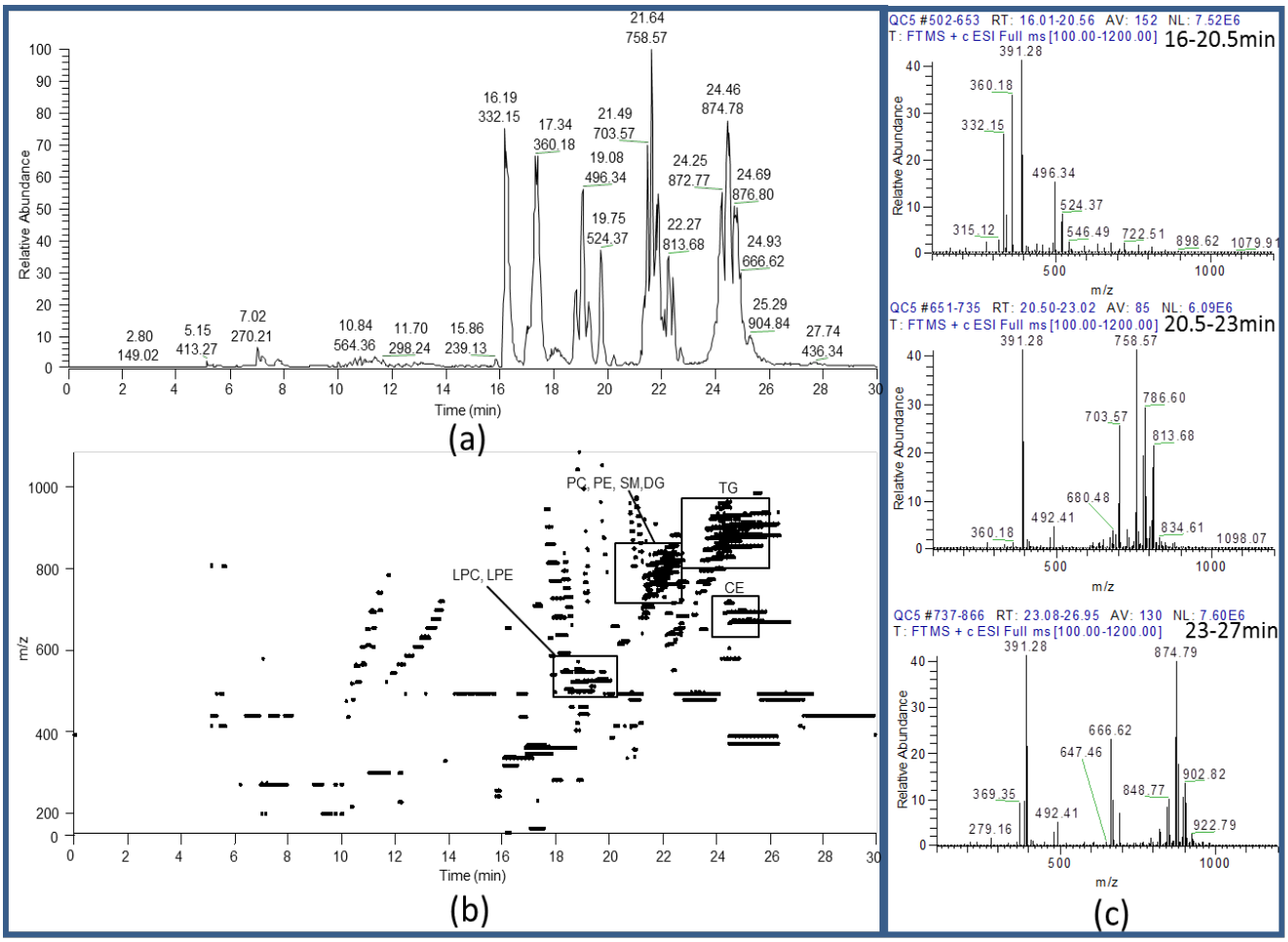

Figure 1.1: Base peak chromatogram of 20 µl of human plasma

(A) positizve ionization mode; (B) density map (m/z vs. retention time); (C) averaged mass spectra of three different time regions. LPC: monoglycerophosphocholines; LPE: monoglycerophosphoethanolamines; PE: glycerophosphoethanolamine; PC: glycerophosphocholine; SM: Sphingomyelin; DG: Diacylglycerol TG: Triacylglycerol; CE: Cholesteryl ester.

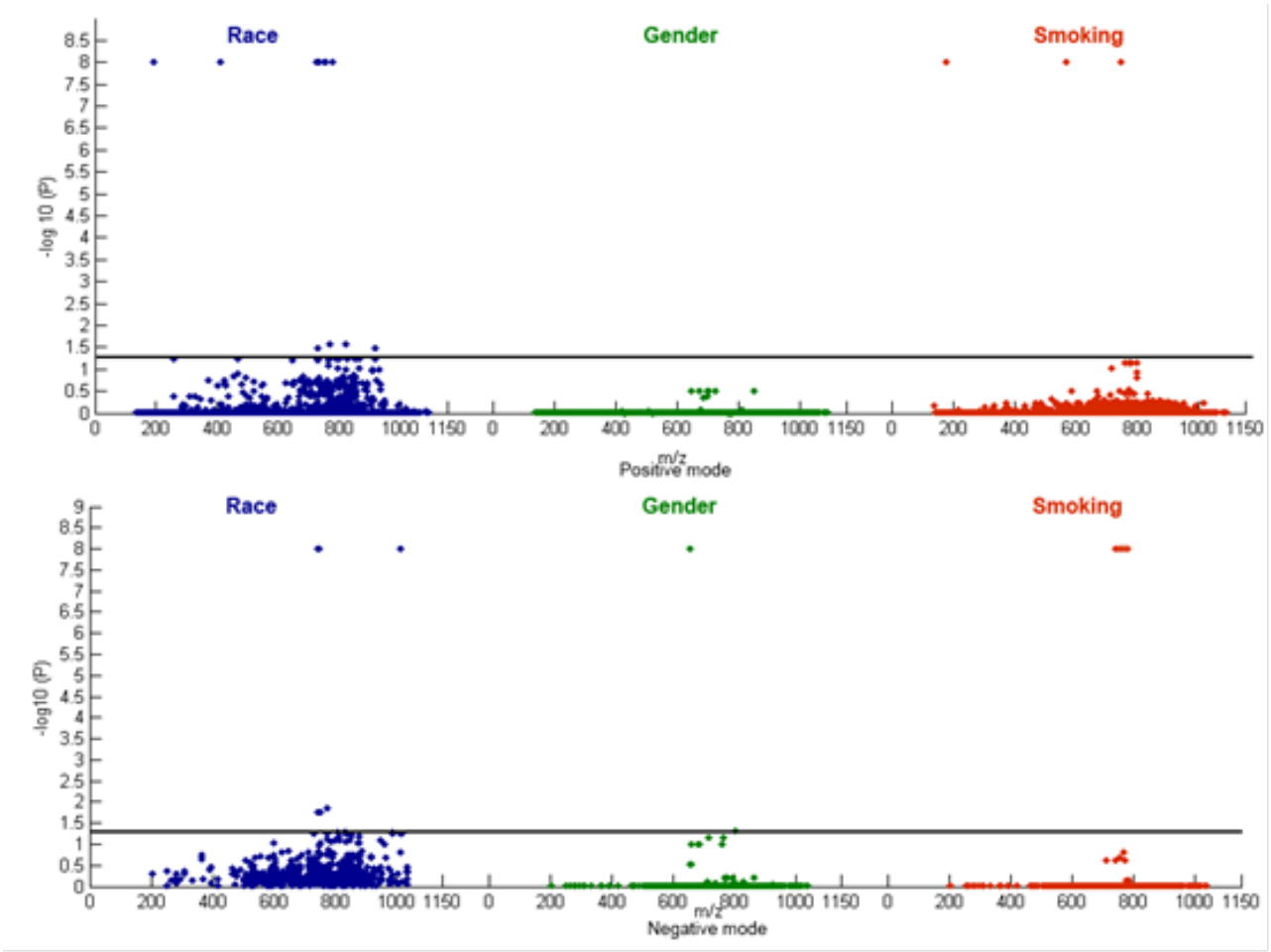

Figure 1.2: Scatter plots of P-values from ANOVA with BH correction

Based on data from (a) positive ionization mode and (b) negative ionization mode. Points above the bold line indicate a BH-adjusted *P*-value less than 0.05. (Points were truncated at a *P*-value of 10^{-8}).

Table 1.1: Characteristics and estimated intake of dietary fat across subjects represented by pooled plasma samples.

	Race			Gender			Smoking Status		
Characteristics[a]	**Black**	**White**	***P*-value[c]**	**Male**	**Female**	***P*-value[c]**	**Smokers**	**Non-smokers**	***P*-value[c]**
Number of samples	18 (9M, 9F; 9S, 9NS)	17 (9M, 8F; 10S, 7NS)		18 (9B, 9W; 10S, 8NS)	17 (9B, 8W; 9S, 8NS)		19 (10M, 9F; 9B, 10W)	16 (8M, 8F; 9B, 7W)	
Age (y)	26 ±2	25 ± 2	0.148	26 ±3	25 ± 2	0.627	26 ±3	25 ± 2	0.516
BMI (kg/m^2)	28.0±3.56	24.1±1.64	2.20E-04	25.7±2.21	26.6±4.35	0.410	26.9±3.94	25.2±2.42	0.143
Dietary Fat[b]									
Animal Fat (g)	59.7±21.82	39.2±9.84	1.38E-03	54.3±22.47	45.0±15.78	0.168	57.9±22.03	40.1±11.01	5.905E-03
Vegetable Fat (g)	39.2±14.5	27.3±5.67	3.77E-03	35.0±13.0	31.7±12.1	0.431	38.9±14.0	26.9±6.17	3.340E-03
Cholesterol (mg)	389±142	248±64.7	8.06E-04	356±155	283±88.2	0.097	371.2±146.33	261±77.8	0.010

[a] Between 4 and 6 individual plasma specimens were pooled by race, gender and smoking status.

[b] Dietary fats and fatty acids were compiled from standardized food frequency questionnaires applied to individual subjects and averaged for pooled-plasma specimens in the current investigation.

[c] Based on Student's *t*-test.

Table 1.2: Significant lipidomic features associated with race, gender and smoking status.

Factor	*m/z*	Ret. time (min)	Ionization mode	MSMS fragments	Compound	Species	HMDB ID[a]	Δppm	*P*-value[b]	Fold change[c]
Race	724.53126	21.861	ESI+	583.40, 364.22, 361.28	PE (P-16:0/20:4)	[M+H]+	08937	5.1	<1.00E-08	2.57
Race	750.54676	21.924	ESI+	609.42, 390.31, 361.23	PE (P-18:1/20:4)	[M+H]+	11419	4.7	<1.00E-08	1.77
Race	751.55602	21.945	ESI+	609.42, 390.27, 361.21	Isotope of PE (P-18:1/20:4)	[M+H]+	11419	6.4	<1.00E-08	1.98
Race	752.56303	22.071	ESI+	611.43, 392.32, 361.28	PE (P-18:0/20:4)*	[M+H]+	11386	5.6	<1.00E-08	1.81
Race	753.56640	22.062	ESI+	611.44, 392.28, 361.28	Isotope of PE (P-18:0/20:4)*	[M+H]+	11386	0.6	<1.00E-08	2.00
Race	754.58059	22.114	ESI+	612.46, 392.34, 362.30	Isotope of PE (P-18:0/20:4)*	[M+H]+	11386	7.6	<1.00E-08	2.31
Race	776.56270	22.044	ESI+	635.44, 392.33, 385.29	PE (P-18:0/22:6)	[M+H]+	11394	5.0	<1.00E-08	1.81
Race	778.57958	22.076	ESI+	637.45, 392.22, 387.39	PE (P-18:0/22:5)	[M+H]+	11394	6.5	<1.00E-08	1.49
Race	732.55708	21.525	ESI+	503.25, 311.18	Unknown				<1.00E-08	0.66
Race	733.56106	21.517	ESI+	504.33, 311.32	Unknown				<1.00E-08	0.68
Race	195.08792	10.185	ESI+	Undetected	Unknown				<1.00E-08	0.26
Race	410.36418	20.121	ESI+	392.29, 350.30, 326.28, 186.13, 168.11	8,11,14-Eicosatrienoylethanol-amide (DGLEA)	[M+Isopro+H]+	13625	1.9	<1.00E-08	0.23
Race	766.57813	21.844	ESI+	Undetected	Unknown				0.026	1.41
Race	818.61083	21.944	ESI+	635.58, 550.35, 508.59	PC(P-18:0/22:6)*	[M+H]+	11262	6.2	0.026	1.45
Race	728.58275	21.400	ESI+	Undetected	Unknown				0.033	1.34
Race	917.84378	24.052	ESI+	900.62, 618.43, 604.57	Unknown				0.033	0.62
Race	750.54772	21.853	ESI-	464.34, 303.26, 259.20	PE (P-18:0/20:4)*	[M-H]-	11386	4.6	<1.00E-08	1.70
Race	776.56369	21.839	ESI-	464.31, 329.24	PE (P-18:0/22:5)	[M-H]-	11393	4.7	0.014	1.56
Race	753.56665	21.715	ESI-	Undetected	Unknown				0.018	1.43

Race	1013.78243	22.614	ESI-	835.69, 726.53, 702.55, 700.54, 329.28, 303.29	Unknown			<1.00E-08	1.52
Race	744.55806	21.645	ESI-	480.36, 281.21, 255.21	PC (16:0/18:1)	[M-CH3]-	5.1	<1.00E-08	0.81
Race	745.56145	21.648	ESI-	480.33, 281.20, 255.23	Isotope of PC (16:0/18:1)	[M-CH3]-	0.9	0.018	0.82
Race	746.56465	21.669	ESI-	481.34, 281.22, 255.19	Isotope of PC (16:0/18:1)	[M-CH3]-	7.1	0.018	0.81
Gender	657.49937	20.826	ESI-	Undetected	Unknown			<1.00E-08	1.79
Gender	806.61063	21.954	ESI-	Undetected	Unknown			0.048	1.35
Smoking	177.10246	7.481	ESI+	145.91, 120.01, 117.97, 97.85, 79.87	Cotinine	[M+H]+	1.3	<1.00E-08	262
Smoking	567.38180	7.475	ESI+	391.04, 177.15	Unknown			<1.00E-08	190
Smoking	568.38515	7.475	ESI+	392.29, 177.09	Unknown			<1.00E-08	598
Smoking	745.56131	21.475	ESI+	Undetected	Unknown			<1.00E-08	0.77
Smoking	743.61051	21.868	ESI-	711.10, 462.31, 279.26	Unknown			<1.00E-08	0.84
Smoking	757.62638	21.968	ESI-	739.04,717.26, 697.40, 667.07, 279.17, 271.23	Unknown			<1.00E-08	0.74
Smoking	769.62632	21.923	ESI-	751.03, 727.09, 709.24, 511.19, 279.24, 281.30	Unknown			<1.00E-08	0.85
Smoking	783.64208	22.036	ESI-	767.05, 710.95, 521.06	Unknown			<1.00E-08	0.82
Smoking	784.64551	22.036	ESI-	767.10, 711.98	Unknown			<1.00E-08	0.80

* Confirmed by reference standard

[a] Human Metabolome Database

[b] *P*-values are based on ANOVA with BH correction

[c]Fold change based upon ratios of mean values for black/white, male/female or smoker/nonsmoker

Table 1.3: Multivariable linear models of covariate effects on analytes representing significant lipidomic features. Analyte levels were modeled as ratios of individual peak intensities divided by the sum of all peak intensities

Analyte	**Parameter**	**Estimate**	***P*-value**	**R^2**	**$\Delta R^{2\,a}$**
Summed Plasmalogens	Intercept	3.786			
	Race (ref. = white)	3.336	<0.0001	0.6801	-
	Gender (ref. = female)	0.912	0.036	0.7414	0.0613
	Animal fat (g)	0.022	0.093	0.7642	0.0228
DGLEA[b]	Intercept	0.015			
	Race (ref. = white)	-0.063	0.016	0.504	-
	Gender (ref. = female)	0.074	0.0058	0.5343	0.0303
	Race×Gender	-0.091	0.014	0.6179	0.0836
PC(P-16:0/18:1)	Intercept	70.297			
	Race (ref. = white)	-13.07	<0.0001	0.4503	

[a] Change of R^2 value after adding the covariate to the model.

[b] Dihomo-γ-linolenoyl ethanolamide

Chapter 2: Evaluating ultra-long chain fatty acids as biomarkers of colorectal cancer risk

Perttula, Kelsi[a]; Edmands, William MB[a]; Grigoryan, Hasmik[a]; Cai, Xiaoming[a]; Iavarone, Anthony T[b]; Gunter, Marc J[c]; Naccarati, Alessio[d], Polidoro, Silvia[d], Hubbard, Alan[a]; Vineis, Paolo[d,e]; Rappaport, Stephen M[a]*

[a] School of Public Health, University of California, Berkeley, California, 94720, United States

[b] California Institute for Quantitative Biosciences, University of California, Berkeley, California 94720, United States

[c] International Agency for Research on Cancer, Lyon, France

[d] HuGeF Foundation, Torino, Italy

[e] MRC-PHE Centre for Environment and Health, Imperial College, Norfolk Place London W2 1PG, UK

A similar version of this manuscript has been published: Perttula K, Edmands WM, Grigoryan H, Cai X, Iavarone AT, Gunter MJ, Naccarati A, Polidoro S, Hubbard A, Vineis P, Rappaport SM. Evaluating Ultra-Long-Chain Fatty Acids as Biomarkers of Colorectal Cancer Risk. Cancer Epidemiology and Prevention Biomarkers. 2016 Aug 1;25(8):1216-23.

2.1 Abstract

Background: Cross-sectional studies reported a novel set of hydroxylated ultra-long-chain fatty acids (ULCFAs) that were present at significantly lower levels in colorectal cancer (CRC) cases than controls. Follow-up studies suggested that these molecules were potential biomarkers of protective exposure for CRC. To test the hypothesis that ULCFAs reflect causal pathways, we measured their levels in prediagnostic serum from incident CRC cases and controls.

Methods: Serum from 95 CRC patients and 95 matched controls was obtained from the Italian arm of the European Prospective Investigation into Cancer and Nutrition cohort and analyzed by liquid chromatography-high-resolution mass spectrometry. Levels of 8 ULCFAs were compared between cases and controls with paired *t*-tests and a linear model that used time to diagnosis (*ttd*) to determine whether case-control differences were influenced by disease progression.

Results: Although paired *t*-tests detected significantly lower levels of four ULCFAs in CRC cases, confirming earlier reports, the case-control differences diminished significantly with increasing *ttd* (7 d to 14 y).

Conclusion: Levels of several ULCFAs were lower in incident CRC cases than controls. However, because case-control differences decreased with increasing *ttd*, we conclude that these molecules were likely consumed by processes related to cancer progression rather than causal pathways.

Impact: ULCFA levels are unlikely to represent exposures that protect individuals from CRC. Future research should focus on the diagnostic potential and origins of these molecules. Our use of *ttd* as a covariate in a linear model provides an efficient method for distinguishing causal and reactive biomarkers in biospecimens from prospective cohorts.

2.2 Introduction

Colorectal cancer (CRC) accounts for one fourth of all cancer deaths worldwide and is the second leading cause of cancer mortality in the United States and Europe [45,46]. Since less than 15 percent of the variation in risk of CRC has been attributed to heritable genetic factors [47], exposures such as nutrients, microbial metabolites, toxins, and pathogens are likely to play a significant role in CRC development. Exposures that have been associated with increased risks of CRC include obesity, cigarette smoking, alcohol use, and consumption of n-6 polyunsaturated fatty acids, all of which contribute to oxidative stress and inflammation (reviewed in Stone, *et al*. [48]). On the other hand, regular consumption of aspirin – an antioxidant and anti-inflammatory drug - reduces CRC risk [48,49]. Aspirin inhibits both COX-1 and COX-2 enzymes, preventing the production of inflammatory prostaglandins and thromboxanes [50] and also acetylates COX-2 and thereby allows conversion of n-3 and n-6 fatty acids to inflammation-resolving compounds (lipoxins are derived from n-6 fatty acids and resolvins and protectins from n-3 and n-6 fatty acids) [51]. This combination of factors suggests that CRC may result from an imbalance in production and removal of reactive electrophiles and inflammatory products that can initiate and promote tumors [48,52,53].

Recently, Ritchie *et al*., used untargeted high-resolution mass spectrometry (HRMS) to detect a novel class of polyunsaturated, hydroxylated, ultra-long-chain fatty acids (ULCFAs, containing between 28 and 36 carbons) that was associated with reduced risks of CRC in three case-control studies [4]. Using accurate-mass signatures of a dozen representative ULFCAs, Ritchie *et al*. reported that concentrations of these molecules were not correlated with either the tumor stage or type of treatment in cases. Furthermore, ULCFA levels declined with increasing age (whereas risk of CRC increases with age) in cases and controls, indicating a possible protective effect of ULCFAs [54]. Moreover, a large follow-up study of colonoscopy patients by the same authors indicated that subjects under the age of 50 that were in the lowest decile of ULCFA-serum concentrations had a relative CRC risk of 10.1 (C.I.: 6.4 – 16.4) [5].

In attempting to elucidate a protective mechanism for these molecules, Ritchie *et al*. dosed human CRC (SW620) cells with 28-carbon ULFCAs that had been isolated from human serum, and reported reduced production of pro-inflammatory markers (NFκB, IκBα, and NOS2) [55]. Since, as noted above, inflammation has been a hallmark of CRC [48,52,56], the inverse correlation of ULCFA levels and CRC risk would be consistent with a cancer mechanism that favors a pro-inflammatory environment that increases with age. Furthermore, the purported anti-inflammatory or protective properties of ULCFAs could be similar to those of hydroxylated very-long chain fatty acids that are metabolized into inflammation-resolving compounds (*i.e*. lipoxins, resolvins, and protectins). These compounds are active in the pM – nM range [53] and have epimeric forms that are triggered by aspirin, which reduces risks of CRC and cancer generally [49,57].

Remarkably, the provocative findings of Ritchie *et al*. [4,5,33,54,55] implicating low serum levels of ULCFAs as potential causes of CRC have not been explored by other investigators. Since all of the reported associations between circulating levels of ULFCAs and CRC were derived from cross-sectional studies [4] it is particularly important to replicate Ritchie's findings with archived cohort samples that were collected prior to CRC diagnosis. This would reduce the likelihood that lower levels of ULFCAs in CRC cases resulted from tumor-induced dysregulation of homeostatic pathways (reverse causality). The purpose of this study is to test the hypothesis that ULFCAs are potentially protective against CRC with pre-diagnostic serum from 95 incident CRC cases and matched controls from the European Prospective Investigation of Cancer and Nutrition (EPIC).

2.3 Materials and Methods

Experimental Design

We adopted a simple regression model to determine whether ULCFAs represent biomarkers on the causal pathway to CRC or are reactive biomarkers related to progression of the disease. Since the EPIC serum had been obtained between 7 d and 14 y prior to CRC diagnosis, we used the (log-scale) difference in ULFCA concentrations (CRC case minus matched control) as the outcome variable in a linear model to simultaneously investigate effects of case status and time to diagnosis (*ttd*) on the risk of CRC. (Note that these log-scale case-control differences represent case:control ratios in natural scale). The model is shown as follows:

$$Y_i = \beta_0 + \beta_1(ttd)_i + \varepsilon_i, \qquad (1)$$

where Y_i represents the case-control difference of (log-transformed) ULCFA levels for the i^{th} case-control pair, β_0 is the intercept representing the case-control difference at recruitment, and β_1 is the coefficient for ttd (d). Evidence favoring a non-zero intercept (β_0) would indicate that a given ULCFA level differed on average between cases and controls. A negative intercept, illustrated with the hypothetical example in Figure 2.1A, would indicate higher ULCFA levels in controls (*i.e.* a protective effect) as suggested by Ritchie *et al.* [4]. Likewise, a significant coefficient for *ttd* (β_1), illustrated in Figure 2.1B, would indicate that the timing of blood collection relative to diagnosis affected the outcome and, therefore, that any case-control difference in the ULFCA level probably reflects progression of CRC. Thus, the combination of a negative β_0 and non-significant β_1 would point to a potentially causal biomarker of CRC while a significant β_1 would point to a reactive biomarker.

Study Population

EPIC is a large prospective cohort study with approximately 520,000 participants, aged 25–70 years at enrollment from 1992 through 2000, from 23 centers in 10 European countries [58]. All study participants provided written informed consent. Serum was collected at enrollment and dietary information was obtained with a food-frequency questionnaire [59,60]. The serum for this investigation consisted of 190 specimens (95 case-control pairs), collected between 1993 and 1997 from subjects in Turin, Italy. Controls were matched to incident cases by age, study enrollment year and season, and gender. Summary statistics for these subjects are listed in Table 2.1 including *ttd*, gender, body mass index (*bmi*), waist circumference, and self-reported consumption of fish and shellfish. These covariates were selected based on previous evidence that *bmi* and waist circumference are associated with CRC risk [61,62] and that diets rich in fish oil have reduced risks of inflammation-related diseases [63,64].

Chemicals

LC-MS grade (Fluka) isopropanol, methanol, water and ^{13}C- cholic acid (internal standard) were from Sigma-Aldrich (Milwaukee, WI, USA). LC-MS grade (Optima) acetic acid and chloroform were from Fisher Scientific (Santa Clara, CA, USA). All chemicals were of analytical grade and were used without purification.

Sample Processing

Shortly after collection, a 0.5-ml aliquot of each serum sample was placed in a cryostraw, sealed, and stored in liquid nitrogen (-196 ºC) at the International Agency for Research on Cancer in Lyon, France. Approximately one year prior to analysis, cryostraws were transported (with dry ice) to our laboratory in Berkeley, CA (USA), where they were maintained at -80 ºC. After opening each cryostraw, 20 µl of serum was mixed with 100 µl of a solvent mixture (isopropanol/methanol/water = 60:35:5) containing ^{13}C-cholic acid as an internal standard (3.0 µg/ml). After mixing samples for one minute with a vortex mixer, samples were allowed to stand at room temperature for 10 min. to precipitate proteins and were then centrifuged for 10 min at 10,000 g. The supernatant was removed and stored at 4 ºC prior to liquid chromatography (LC)-HRMS. Case control pairs were analyzed sequentially but in random order. A local quality-control sample, prepared by pooling aliquots from each serum sample, was analyzed as each tenth injection to provide technical replicates for estimating precision.

Liquid chromatography-HRMS was performed on two platforms. The first 132 samples were analyzed with an Agilent LC (1100 series) coupled to an Agilent HRMS (Model 6550 QTOF, Santa Clara, CA, USA). Due to a malfunction, this QTOF required repairs before analyses could be completed. In order to permit timely analysis, the remaining 58 samples were analyzed with an Agilent 1200 series LC (Santa Clara, CA, USA) coupled to an LTQ Orbitrap XL HRMS equipped with an Ion Max ESI source (Thermo Fisher Scientific, Waltham, MA, USA). On both platforms, 10 µl of each sample was injected from a full loop into a Luna C5 column (2.1 × 50 mm, 100 Å, 5 µm, Phenomenex, Los Angeles, CA) operated with gradient elution of mobile phase A (methanol/0.5 % acetic acid = 5:95) and mobile phase B (isopropanol/methanol/0.5 % acetic acid = 60:35:5) as follows: 100% A for 2 minutes at 0.05 ml/min; 0-83% B from 2-7 minutes at 0.3 ml/min; 83-100% B from 7-14 minutes at 0.3 ml/min; 100% B from 14-17 minutes; and 100% A from 17-22 minutes. The autosampler and column oven were maintained at 4 ºC and 40 ºC, respectively. The electrospray was operated in negative ionization mode. To monitor system stability, pooled quality control samples were injected every tenth sample. Tandem MS/MS spectra were obtained with the Orbitrap platform.

During processing, approximately one third of the serum samples was observed to have a gelled consistency that apparently resulted from a preservative(s) contained in the cryostraws [65,66]; gelled serum from EPIC cryostraws has been observed previously [67]. Pairs with at least one gelled sample were analyzed in a single batch (batch 1, $n = 96$) on the QTOF platform, and the remaining (non-gelled) pairs were analyzed in two batches on either the QTOF platform (batch 2, $n = 36$) or the Orbitrap platform (batch 3, $n = 58$).

Since previous reports had implicated consumption of seafood as being potentially protective of CRC [63,64], several fresh seafood samples were purchased from a local market in Berkeley, California and tested for the presence of ULCFAs. Four types of seafood were tested: raw white shrimp (Thailand), wild American sea scallops, and farmed American Littleneck clams and live mussels. Samples from these four species (50 µl) were extracted for lipids using the Bligh and Dyer chloroform extraction method [68,69]. These extracts were analyzed on the Orbitrap platform, with the same method as described above.

Data Processing

Raw data were converted to mzXML format for peak picking using ProteoWizard software (Spielberg Family Center for Applied Proteomics, Los Angeles, CA). Peak detection and retention time alignment were performed with the xcms package within the R statistical programming environment [70,71]. For the data collected on the QTOF, parameters include centwave feature detection, orbiwarp retention time correction, minimum fraction of samples in one group to be a valid group = 0.25, *P*-value thresholds for blank versus QC samples = 0.01, isotopic ppm error = 10, width of overlapping *m/z* slices (mzwid) = 0.015, bandwidth grouping (bw) = 2, minimum peak width = 2 s, maximum peak width=20 s. Parameters for the Orbitrap platform were the same except for: isotopic ppm error = 2.5, minimum peak width = 2 s, maximum peak width=70 s, bw = 5, prefilter peaks = 3, prefilter intensity = 5000, based on xcms parameters optimized for Orbitrap instruments [72]. The resulting peak tables of retention times, *m/z* values, and peak intensities were exported for further processing. Subsequent analyses were also performed with the R platform (version 3.2.1) [73].

Because reference standards for the ULCFAs are not available, mass spectra were interrogated for 13 accurate masses representing ULFCAs with between 28 and 36 carbons that had been reported by Ritchie *et al* [4,33]. These ULFCAs are listed in Table 2.2 along with their masses and elemental formulae. We targeted these 13 ions in our analyses and Table 2.2 shows the retention times and observed masses, along with the mass accuracy expressed as the mass deviation (ppm) between the theoretical and observed masses. Tandem MS analyses revealed fragment ions representing losses of CO_2 and one or two H_2O molecules for all 13 precursor ions. These losses are consistent with hydroxylated carboxylic acids and with fragment ions reported by Ritchie, *et al*. [4]. After extracting accurate masses for the 13 putative ULCFAs from total-ion chromatograms for all EPIC specimens, extracted-ion chromatograms were visually examined and five of the features were excluded because some peaks were not reproducibly detected above noise levels (ULFCAs 518, 574, 576, 578, and 592) (Table 2.2).

For quantitation of ULCFA levels, we followed the same approach as Ritchie *et al*. [54] and normalized analyte peak areas by the corresponding peak areas of an internal standard (^{13}C-cholic acid, final concentration = 3.0 µg/ml). These normalized ULCFA abundances are designated as 'peak-area ratios' (PARs). Preliminary statistical analyses indicated that use of PARs, rather than simply ULCFA peak areas, reduced nuisance variation from instrumental variability and matrix effects.

Statistical Analysis

Batch adjustment was performed with a linear model of the log-transformed PAR of each analyte, which included dummy variables for batch and gel status as independent variables. Residuals from these linear models were used as dependent variables in subsequent statistical analyses. These residuals represent log-transformed PAR values normalized to a mean of zero. Coefficients of variation (CVs) for the eight ULCFAs with acceptable peak morphology were estimated from the error variances (σ_e^2) of log-transformed PARs after batch and gel adjustment as $\sqrt{e^{\sigma_e^2} - 1}$ [74] (Table 2.2).

Analyte levels were compared between cases and controls using one-sided paired *t*-tests as well as the linear model (1) for evaluating both case-control differences and effects of *ttd* (Table 2.3). Additional linear models were constructed by adding waist circumference and self-reported consumption of fish and shellfish to model (1) as covariates (Table 4). Waist circumference had previously been associated with CRC [61,62] and consumption of fish and shellfish introduces n-3 fatty acids into the diet that purportedly reduce cancer risks [63,64] and are metabolized to anti-inflammatory lipoxins, resolvins, and protectins [55]. As noted above, some serum samples had a gelled consistency. When gel status was added to linear models, no significant main effect or interaction between case-control status and gel status was detected (results not shown).

2.4 Results

Approximately normal distributions of logged ULCFA PARs were verified for all three batches, and Kruskal–Wallis tests detected no significant differences across batches (P-value > 0.33). As indicated in Table 2.2, CVs ranged from 9.1 to 27.6% (mean 22%) for the 8 ULCFAs with acceptable peak morphology.

As shown in Table 2.3, paired-*t* tests detected significantly lower PARs in cases compared to controls for four 28-carbon ULCFAs (446, 466, 468, and 494). Significant case-control differences of PARs were confirmed with a negative intercept from model (1) for the same 28-carbon ULCFAs and a fifth 30-carbon ULCFA (492). Interestingly, these five ULCFAs also showed statistically significant coefficients for time to diagnosis (*ttd*). Indeed, as shown in Figure 2.3, PAR differences between cases and controls increased with *ttd* for all 8 ULCFAs. Since case-control differences in levels of these ULCFAs appear to decline with increasing *ttd*, we conclude that these molecules are reactive biomarkers of CRC progression rather than biomarkers of protective exposure, as hypothesized by Ritchie, *et al*. [54].

Table 2.4 shows results from extensions of model (1) to include waist circumference, and self-reported consumption of fish and shellfish. No association was observed between the covariates and case-control differences in PAR values. No ULCFA peaks were distinguishable from background noise in the seafood samples.

Although our study confirms that levels of ULCFAs with 28–30 carbons are significantly lower in incident CRC cases than matched controls [4], the influence of *ttd* on case-control differences (Figure 2.2) suggests that these fatty acids are more likely to be markers of CRC progression rather than biomarkers of protective exposure.

Evidence that lower levels of ULFCAs may be linked to the progression of CRC points to tumor-induced metabolism as a likely contributor, but leaves open the question as to the origins of the molecules. Although Ritchie *et al*. readily observed ULCFAs in human serum, they failed to detect the same molecules in sera from rats, mice and cattle, in various plant tissues and grains, and in human cell lines from tumors and normal colonic tissue [4]. Aside from carbon-chain length, the proposed structures of ULCFAs [75] resemble those of the lipoxins, resolvins, and protectins (20-22 carbons); these are mono-, di-, and tri- hydroxylated products of long chain fatty acids such as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), that have been decarboxylated through metabolism [76–79]. Since EPA and DHA are present in oily tissues from marine species, we suspected that the ULCFAs might also be present in seafood. However, we did not detect ULCFAs in commercial samples of shrimp, scallops, clams or mussels.

While the origin of hydroxylated ULCFAs remains unknown, very long chain (VLC) PUFAs, ranging from 22-34 carbons, have been described [80,81] and detected in spermatozoa, retinas, and brain tissue [82,83]. PUFAs longer than 22 carbons are generated by elongase ELOVL-4, which is one of seven endoplasmic-reticulum-bound enzymes responsible for lengthening particular fatty acids [84]. While these VLC-PUFAs are not typically hydroxylated, it is plausible that they share common synthetic pathways with the hydroxylated ULCFAs described by Ritchie, *et al*. Alternatively, elongases ELOVL2 and ELOV5 extend typical-length PUFAs (18-22 carbon) but have not been investigated as possible progenitors of ULCFAs [85].

Our approach for simultaneously comparing paired case-control differences as a function of *ttd*, embodied in model (1), offers an efficient mechanism for differentiating biomarkers of exposure from those of disease progression and is sufficiently general for use with either targeted or untargeted analyses of biospecimens from prospective cohorts. Previous analyses that employed *ttd* in studies of disease etiology have been restricted to biomarker levels in cases only [60,86,87]

and have also been used to exclude cases diagnosed relatively soon after specimen collection (e.g. 2-5 years) [88–90].

For the CRC case-control samples evaluated in the current study, the 28-carbon ULCFAs were the class most highly associated with case status and *ttd* (Table 2.3). Ritchie, *et al*. reported that several 36-carbon compounds were also highly discriminating between cases and controls for both CRC [4,5] and pancreatic cancer [33,91]. However, the only 36-carbon ULCFA that we were able to quantify was 594, which was not significantly associated with either CRC case status or *ttd* (Table 2.3), although the plot in Figure 2.2 suggests a weak, but consistent, trend with *ttd*.

2.5 Discussion

Although our results tend to downplay the potential roles of ULCFAs as biomarkers of protective exposure, they may be worth evaluating as diagnostic biomarkers of CRC. Indeed, relationships shown in Table 2.3 point to significant reductions in three of the 28-carbon ULCFAs (446, 466, & 468) starting between about 1,500 - 3,000 d (3 – 7 y) prior to diagnosis.

Finally, we emphasize that our methods relied on accurate masses to pinpoint ULCFAs and employed quantitation relative to ^{13}C-cholic acid (internal standard). With availability of reference standards, it would be possible to detect and quantitate these molecules with greater precision and thus to reduce measurement errors and resulting attenuation biases that probably weakened associations observed with CRC status and *ttd*. However, improved standardization would be unlikely to remove the consistent effects of *ttd* that were observed in our samples of CRC cases and controls from the EPIC cohort (Figure 2.2).

We recognize that our study is small and has limited power to detect associations between ULCFAs and CRC. Nonetheless, these results offer important clues that the ULCFAs might be useful diagnostic markers. Validation with larger sample sets is now necessary.

In conclusion, these targeted analyses of 8 accurate masses, which are characteristic of ULCFAs reported by Ritchie *et al*. in case-control studies [4], confirmed that some ULCFAs were present at significantly lower levels in incident CRC cases than matched controls from the EPIC cohort. However, clear trends with *ttd* indicate that the observed case-control differences are unlikely to be due to the ULCFAs acting as protective exposures but rather reflect progression of the disease. Although ULCFAs are probably not involved with causal pathways leading to CRC, their correlations with *ttd* suggest that they may be useful diagnostic biomarkers. Future research regarding applications of these molecules in cancer research would benefit from synthesis of reference standards and knowledge of the dietary or metabolic origins of these novel molecules.

Our use of a linear model that employed *ttd* as a covariate [model (1)] provides an efficient method for distinguishing causal and reactive biomarkers in specimens of blood from prospective cohorts. The model is simple to apply and is sufficiently general for use with either targeted or untargeted analyses of biospecimens.

2.6 Tables and Figures

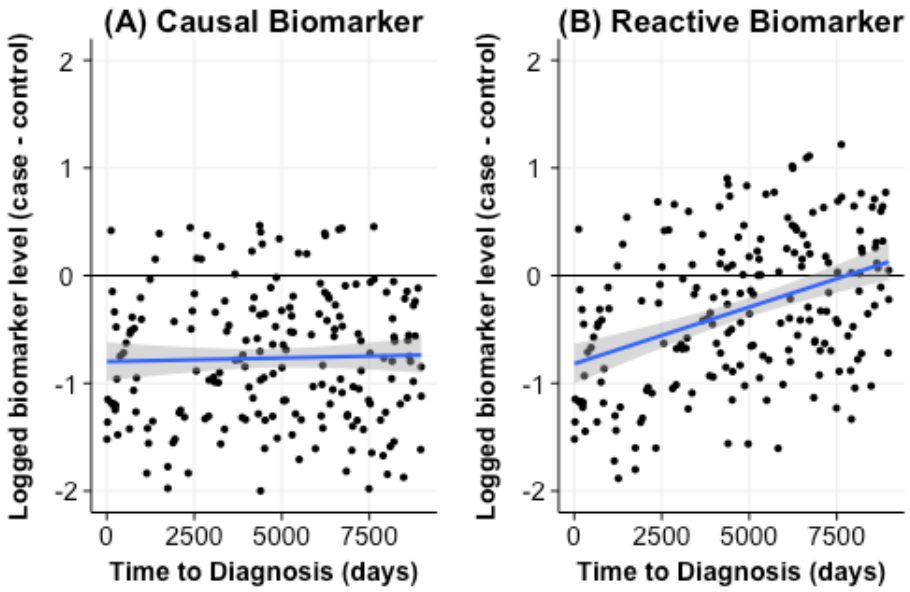

Figure 2.1: Use of a linear model (Model 1) to differentiate a causal biomarker from a disease-related biomarker

Hypothetical data representing levels of a biomarker were generated for case control pairs, transformed to natural logarithms, and the case-control differences plotted versus time to diagnosis (*ttd*). (A) Shows that case-control differences are consistently less than zero indicating that biomarker levels are greater in controls than in cases and are not affected by *ttd*. This would indicate a biomarker of protective effect. (B) Shows case-control differences that diminish with increasing *ttd*, consistent with a biomarker of disease progression.

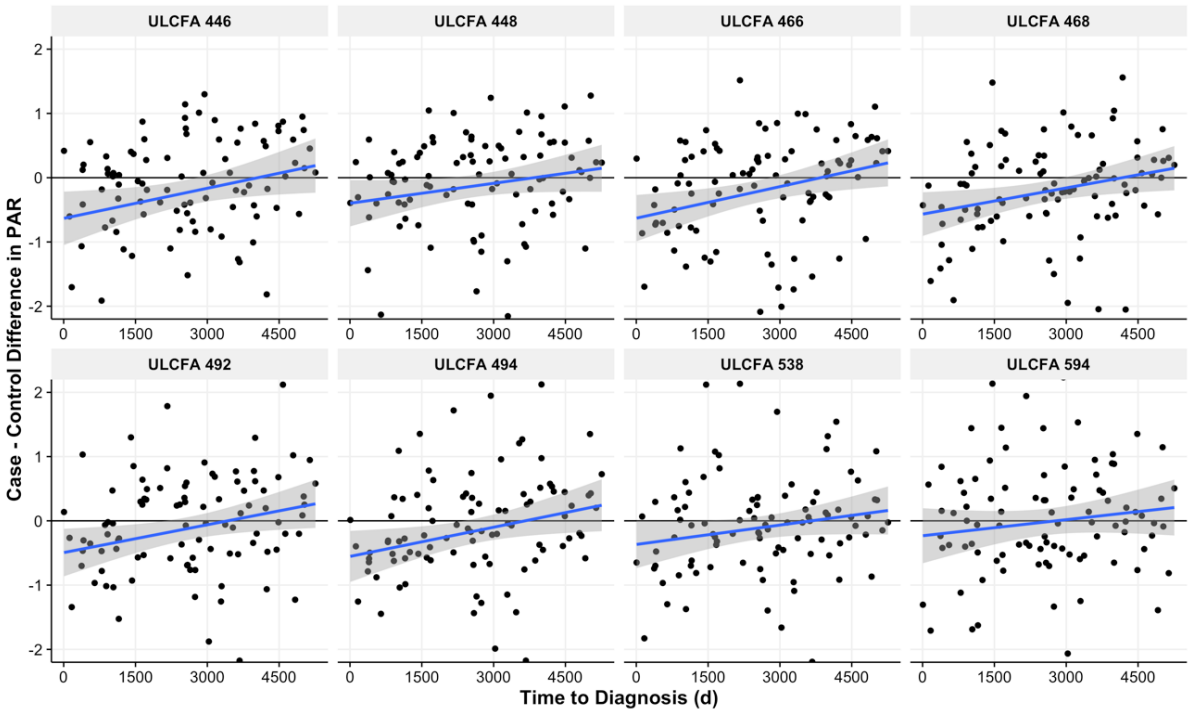

Figure 2.2: Linear-model plots.

Case-control differences for ULCFA levels versus time to diagnosis (*ttd*). Error bands represent 95% confidence intervals. ULCFA levels are reported as peak-area ratios (PARs) relative to ^{13}C-cholic acid (internal standard).

Table 2.1. Descriptive statistics of human subjects.

	Total *n*=190	**CRC cases *n*=95**	**Controls *n*=95**	***P*-value**
Gender	Male	68	68	
	Female	27	27	
Age at enrollment	median	57	57	
(y)	min	36	35	
	max	65	64	
Years to diagnosis	median	7.1	-	
(from enrollment)	min	0.1	-	
	max	14.4	-	
BMI	median	26.4	25.1	0.0090
	min	19.6	18.7	
	max	40.6	33.6	
Waist circumference	median	95	90	0.0005
(cm)	min	68	64	
	max	115	119	
Dietary fish	median	21	24	0.1660
(g/d)	min	1	0	
	max	77	83	
Dietary shellfish	median	4	3	0.4526
(g/d)	min	0	0	
	max	45	76	

Summary statistics of covariates from EPIC human study investigation, matched by age, study enrollment year and season, and gender.

Table 2.2. Ultra-long-chain fatty acids (ULCFAs) reported by Ritchie, *et al*.(8) and detected in the current investigation

ULCFA	Formula	Theoretical *m/z* [a]	Observed *m/z* [a]	Mass dev. (ppm)	Ret. time (sec)	Peak shape [b]	CV
446	$C_{28}H_{46}O_4$	445.3327	445.3324	0.70	610.94	pass	0.276
448	$C_{28}H_{48}O_4$	447.3483	447.3470	3.01	615.20	pass	0.262
466	$C_{28}H_{50}O_5$	465.3590	465.3586	0.88	583.05	pass	0.276
468	$C_{28}H_{52}O_5$	467.3742	467.3744	-0.38	605.56	pass	0.181
492	$C_{30}H_{52}O_5$	491.3741	491.3735	1.22	612.33	pass	0.185
494	$C_{30}H_{54}O_5$	493.3896	493.3906	-1.96	612.28	pass	0.236
518	$C_{32}H_{54}O_5$	517.3902	517.3883	3.59	616.13	fail	ND
538	$C_{32}H_{58}O_6$	537.4164	537.4155	1.58	604.36	pass	0.091
574	$C_{36}H_{62}O_5$	573.4527	573.4508	3.33	611.53	fail	ND
576	$C_{36}H_{64}O_5$	575.4683	575.4666	2.97	616.40	fail	ND
578	$C_{36}H_{66}O_5$	577.4837	577.4842	-0.79	629.90	fail	ND
592	$C_{36}H_{64}O_6$	591.4630	591.4637	-1.21	613.37	fail	ND
594	$C_{36}H_{66}O_6$	593.4786	593.4783	0.42	616.41	pass	0.252

ULCFA molecular formulae, the mass accuracy of the detected molecules, recorded retention time, whether the feature peak shape was sufficiently reproducible and above the noise level, and the coefficient of variation for each accepted peak.

m/z - mass-to-charge ratio; CV - coefficient of variance; ND - not determined.

[a] Theoretical and observed *m/z* values correspond to singly-charged negative ions.

[b] Based upon visual inspection of peak morphology for all selected-ion chromatograms.

Table 2.3: Statistical estimates for one-sided paired t-tests and time-to-diagnosis linear model

ULCFA	**Paired *t*-test**		**Linear model (1)**				
	Est.	***P*-value**	$\boldsymbol{\beta_0}$	***P*-value**	$\boldsymbol{\beta_1(\times 10^3)}$	***P*-value**	$\mathbf{R^2}$
446	-0.237	0.0116	-0.626	0.0037	0.150	0.0373	0.046
448	-0.139	0.0581	-0.390	0.0342	0.097	0.1186	0.026
466	-0.203	0.0139	-0.633	0.0008	0.166	0.0086	0.072
468	-0.215	0.0064	-0.567	0.0014	0.136	0.0219	0.055
492	-0.126	0.0873	-0.490	0.0104	0.140	0.0291	0.050
494	-0.183	0.0300	-0.536	0.0076	0.136	0.0430	0.043
538	-0.108	0.1193	-0.367	0.0527	0.100	0.1169	0.026
594	-0.008	0.4700	-0.238	0.2741	0.089	0.2281	0.016

Difference in means and *P*-values from one-sided *t*-tests of cases and controls; fitted coefficients and *P*-values from linear model (1) comparing case-control differences with time to diagnosis (*ttd*).

β_0- Intercept coefficient

β_1- *ttd* coefficient

Table 2.4: Results of multivariable models of covariates from the EPIC cohort

ULCFA	WC *P* -value	WC Δ R^2	Dietary fish *P*-value	Dietary fish Δ R^2	Dietary shellfish *P*-value	Dietary shellfish Δ R^2
446	0.1402	0.012	0.5714	-0.013	0.3225	-0.005
448	0.5706	-0.001	0.2390	0.017	0.7647	0.001
466	0.3259	-0.016	0.6431	0.012	0.1849	0.030
468	0.7061	-0.007	0.9843	0.016	0.3709	0.026
492	0.3488	0.016	0.6982	0.031	0.7683	0.030
494	0.5955	-0.012	0.5069	0.018	0.5631	0.017
538	0.2055	0.016	0.2654	0.030	0.9275	0.015
594	0.1947	0.019	0.1316	0.050	0.9861	0.023

Effect of adding the waist circumference, dietary fish consumption, or dietary shellfish consumption to linear model (1). The difference in these covariates between matched pairs were not significantly correlated with the difference between the feature levels for the eight ULCFAs.

WC- waist circumference

Chapter 3: Lipidomic features associated with colorectal cancer in a prospective cohort

Perttula, Kelsi[a]; Schiffman, Courtney[a]; Edmands, William MB[a]; Petrick, Lauren[a,g]; Grigoryan, Hasmik[a]; Cai, Xiaoming[a]; Gunter, Marc J[c]; Naccarati, Alessio[d], Polidoro, Silvia[d], Dudoit, Sandrine[a,b,f] ; Vineis, Paolo[d,e]; Rappaport, Stephen M[a]*

[a] School of Public Health, University of California, Berkeley, California, 94720, United States

[b] California Institute for Quantitative Biosciences, University of California, Berkeley, California 94720, United States

[c] International Agency for Research on Cancer, Lyon, France

[d] Italian Institute for Genomic Medicine (IIGM), Torino, Italy

[e] MRC-PHE Centre for Environment and Health, Imperial College, Norfolk Place London W2 1PG, UK

[f] Department of Statistics, University of California, Berkeley, CA, United States

[g] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY

Submitted for publication to BMC Cancer.

3.1 Abstract

Epidemiologists are beginning to employ metabolomics and lipidomics with archived blood from incident cases and controls to discover causes of disease. Although several such studies have focused on colorectal cancer (CRC), they all followed targeted or semi-targeted designs that limited their ability to find discriminating molecules and pathways related to the causes of CRC. Using an untargeted design, we measured lipophilic metabolites in prediagnostic serum from 66 CRC patients and 66 matched controls from the European Prospective Investigation into Cancer and Nutrition (Turin, Italy). Samples were analyzed by liquid chromatography-high-resolution mass spectrometry, resulting in 8,690 features for statistical analysis. Rather than the usual multiple-hypothesis-testing approach, we based variable selection on an ensemble of regression methods, which found nine features to be associated with case-control status. Of these nine features, four appear to be involved in CRC etiology and merit further investigation in prospective studies of CRC. Four other features appear to be related to progression of the disease (reverse causality), and may represent biomarkers of value for early detection of CRC.

3.2 Introduction

Colorectal cancer (CRC) accounts for over 25 percent of all cancer-related deaths with global incidence rates steadily rising [92–94]. Since less than 15 percent of CRC risk has been attributed to heritable genetics [47,95], non-shared exposures and their contributions to gut inflammation are believed to be important etiologic factors [96]. Increased CRC risks have been associated with cigarette smoking, alcohol use, lack of physical activity, obesity, abnormal glucose metabolism, and consumption of red meat and n-6 polyunsaturated fatty acids (PUFAs) [48,97–99]. Conversely, consumption of n-3 PUFAs, fruits, fish, vitamins D and E, and regular use of aspirin appear to reduce CRC risks [48,49,100]. There are also persistent suggestions that the interplay between dietary factors - particularly red meat, lipids, and fiber - and the gut microbiota are effect modifiers for CRC [52,96,101–103].

Many of the associations between exposures and CRC have been gleaned from epidemiological studies that employed self-reported dietary and lifestyle factors [97,98,103,104]. Given the inherent limitations of such data for discovering causal exposures, investigators have recently employed metabolomics to compare small-molecule features between CRC cases and controls. This strategy is based on the idea that small molecules in human blood reflect chemical exposures from both internal and external sources, including the diet, microbiota, psychosocial stress, and pollutants [105]. However, since molecules that discriminate cases from controls in cross-sectional studies can reflect both potential causes of CRC and dysregulation of metabolic processes that result from progression of the disease (reverse causality) [95,106], it is important that biospecimens be collected well before diagnosis to gain insights into causes and effects. Indeed, a class of ultra-long-chain fatty acids (ULCFAs) that discriminated for CRC in several cross-sectional studies [5,54] was essentially ruled out as a causal factor in a prospective cohort [107].

Metabolomic analyses of blood from prospective cohorts have found some associations between CRC incidence and small molecules, as summarized in Table 3.1, with periods of follow-up

ranging from 3.7 to 14.7 years [4,106–111]. Interestingly, all of these nested case-control studies followed targeted or semi-targeted designs where relatively few molecular features were tested between cases and controls. Two of the studies focused on metabolism of dietary choline and found that the mammalian metabolite, betaine, was moderately protective against CRC whereas trimethylamine-*N*-oxide (TMAO), a metabolite mediated via intestinal microbiota, was associated with increased risk [110,111]. A genetic link between TMAO and CRC risk has also been reported [112]. Intriguingly, red meat and other phosphatidylcholine-rich foods appear to contribute to dysbiotic microbiota that generate trimethylamine (the precursor of TMAO) [102,113], whereas fiber-rich foods appear to encourage symbiotic bacteria that are associated with decreased CRC risk [102,114].

Since untargeted metabolomics via liquid chromatography-mass spectrometry (LC-MS) can detect thousands of small-molecule features, traditional hypothesis-testing approaches, that adjust for multiple comparisons by controlling false positive error rates such as the false discovery rate (FDR) [22], can make it difficult to find features whose levels differ significantly between cases and controls. This may have motivated the semi-targeting strategy of Cross et al. (Table 3.1) [109], who limited hypothesis tests of the thousands of detected features to only 278 molecules that had been fully annotated. Such a strategy is likely to be biased towards well curated metabolites that participate in recognized human pathways [105], and thus can miss novel exposures of potential importance to initiation of cancer, including those experienced predominately by either cases or controls. Indeed, of the 278 small-molecules tested by Cross et al. [109], only glycochenodeoxycholate (a secondary bile salt) was associated with increased CRC risk in women (but not men) after using the conservative Bonferonni correction of the *p*-value.

Here, we report results of an untargeted metabolomics analysis of serum from 66 incident CRC cases and matched controls from the European Prospective Investigation of Cancer and Nutrition (EPIC). Given the involvement of lipids in inflammatory processes and CRC [96,115,116], the serum-extraction procedure favored lipophilic molecules. As an alternative to the traditional multiple-hypothesis-testing paradigm for selecting features of potential importance to CRC, we developed a variable-selection strategy that employs an ensemble of diverse prediction methods, including regularized linear regression and regression trees [117–119]. Such methods have recently been applied independently for analyzing metabolomic and other -omic data [117,118,120]. Our analyses point to a small set of features that were predictive of CRC-case status. However, as with all discovery studies, these potentially important features and the molecules they represent must be further validated with independent data sets.

3.3 Materials and Methods

Study Population

EPIC is a prospective cohort study with approximately 520,000 adult participants from across Europe that were enrolled from 1992 through 2000 [58]. The serum for this investigation consisted of 132 specimens (66 case-control pairs), collected between 1993 and 1997 from subjects in Turin, Italy. Controls were matched to incident cases by age, year and season of enrollment, and gender. Dietary data were collected with food frequency questionnaires [59,60]. Summary statistics for these subjects are listed in Table 3.2, including time to diagnosis (*ttd*), gender, body mass index (*bmi*), waist circumference, smoking status, diabetes status, physical activity, and alcohol and meat consumption. These covariates were selected based on previous evidence of associations with CRC risk [48,61,62]. Across our subjects, the only significant differences between CRC cases and controls were observed for *bmi* and waist circumference, both of which were higher in cases (p-values < 0.05, Table 3.2).

Chemicals

Isopropanol (LC-MS grade, Fluka), methanol, water and ^{13}C- cholic acid (internal standard) were from Sigma-Aldrich (Milwaukee, WI, USA). Acetic acid (LC-MS grade, Optima) and chloroform were from Fisher Scientific (Santa Clara, CA, USA). All chemicals were of analytical grade and were used without purification.

Sample Processing

Serum was stored after collection in 0.5-ml aliquots that were placed in cryostraws, sealed, and stored in liquid nitrogen (-196ºC) at the International Agency for Research on Cancer in Lyon, France. Approximately one year prior to analysis, cryostraws were transported (with dry ice) to our laboratory in Berkeley, CA (USA), where they were maintained at -80ºC. As previously reported [107], 20 µl of serum was mixed with 100 µl of a solvent mixture (isopropanol/methanol/water = 60:35:5) containing ^{13}C-cholic acid as an internal standard (final concentration of 3.0 µg/ml). After mixing samples for one minute with a vortex mixer, samples were left at room temperature for 10 min. to precipitate proteins, and were then centrifuged for 10 min at 10,000 g. The supernatant was retained and stored at 4ºC prior to LC-MS. Case-control pairs were analyzed sequentially but in random order. A local quality-control sample, prepared by pooling aliquots from all serum specimens of each batch, was analyzed after every ten samples to monitor system stability and estimate the precision of the analyses.

Mass Spectrometry

Analysis was performed with an Agilent LC (1100 series) coupled to an Agilent MS (Model 6550 QTOF, Santa Clara, CA, USA) as previously reported [107]. Briefly, 10 µl of extracts were slowly loaded on to a Luna C5 column (Phenomenex, Los Angeles, CA) with a 22-minute gradient elution of mobile phase A (methanol/0.5 % acetic acid = 5:95) and mobile phase B (isopropanol/methanol/0.5 % acetic acid = 60:35:5). The electrospray was operated in negative electrospray-ionization (ESI) mode. Tandem MS/MS spectra were obtained on the same platform in data-dependent mode (immediately after data collection) or targeted mode (analysis of the selected features). Full LC-MS acquisition parameters were previously published [107].

Approximately one third of the serum samples had a gelled consistency that resulted from an additive to the cryostraws [65,66,107]. Pairs with at least one gelled sample were analyzed in one (batch 1, $n = 96$), and the remaining (non-gelled) pairs were analyzed in a second batch (batch 2, $n = 36$).

Data Processing

Raw data were converted to mzXML format for peak picking using ProteoWizard software (Spielberg Family Center for Applied Proteomics, Los Angeles, CA). Peak detection and retention-time alignment were performed as described previously [107], using the XCMS package within the R statistical programming environment [70,71,73]. The CAMERA package was used to identify isotopes, ESI adducts, and in-source fragments with the custom rule set used from Stanstrup et al. [121,122]. Annotation of features was conducted using the compMS2Miner package [123], by comparing accurate masses and MS2 fragmentation patterns with the Human Metabolome Database (HMDB) and Metlin [124,125].

Over 24,300 features were initially detected in the negative ESI mode. Features were filtered by removing those with a mean fold-change in abundance less than 1.5 compared to the same peaks in reagent blanks (background noise) and those with coefficients of variation (CV) from QC samples greater than 30% [126,127]. This resulted in a final dataset of 8,690 features for statistical analysis. Feature intensities were (natural) log-transformed and adjusted for batch and gel-status effects using the following linear regression model, previously described in [107]:

$$logY_i = \beta_0 + \beta_1 X_{i,gel} + \beta_2 X_{i,batch} + \epsilon_i, \qquad [1]$$

where Y_i denotes the intensity of a given feature for the i^{th} subject and $X_{i,gel}$ and $X_{i,batch}$ are the corresponding categorical covariates for gel-status and batch. After fitting the linear model, normalized (logged) intensities were obtained by subtracting the estimated batch and gel effects from the original (logged) intensities.

Upper-quantile scaling was used to render the distributions of feature abundances more comparable across all subjects [128,129]. A correlation-network program (Cytoscape, [130]) and an R package clustering algorithm (RAMclust, [131]) were used to identify clustered ions and assist with annotations.

Statistical Methods: Variable Selection

In order to identify discriminating features between CRC cases and control, we shifted the paradigm from multiple hypothesis testing to variable selection based on a combination of three regression methods. First, we considered the following standard linear regression model for the raw intensity of a given feature Y in the i^{th} subject:

$$logY_i = \beta_0 + \beta_1 X_{i,caco} + \beta_2 X_{i,gel} + \beta_3 X_{i,batch} + \beta_4 X_{i,age} + \beta_5 X_{i,gender} + \epsilon_i, \qquad [2]$$

where *caco, gel,* and *gender* denote binary variables for case-control status, presence or absence of gelled serum, and the matched variables of *gender* and *age* (in years). Features were then ranked based on the nominal unadjusted p-value for the case-control coefficient ($_1$). Second, a regularized logistic regression (LASSO) was performed [117,132] with case-control status as the binary outcome variable regressed on the following covariates: normalized log intensities from

Equation [1] for all 8,690 metabolites, *age* and *gender*. Although *age* and *gender* were in the LASSO model, neither variable was selected by the regularized regression.

In order to stabilize feature selection with LASSO, 500 bootstrap samples were taken for each of a variety of penalty parameters [133]. Features chosen by LASSO in at least 10% of the bootstrap samples, across a wide range of penalty parameters, were retained. A data-driven cutoff of 10% was chosen based on plots of the percentage of time that each metabolite was selected during the bootstrap iterations across various penalty parameters, sorted in decreasing order. There was an obvious gap between metabolites selected more than and less than 10% of the time, which lead to choosing this as a natural cutoff. Third, the random forest algorithm [118,134] was used to build a predictor of case-control status, using the same covariates as for the LASSO regression. No obvious jump in variable importance could be seen for the sorted random forest variable importance or the sorted linear regression *p*-values. Therefore, a cutoff of 1% was selected for both of these criteria because this cutoff is relatively stringent, yet still included a reasonable number of variables for consideration. In summary, to select a final set of variables, we included only features that were selected by the bootstrap LASSO and were also among the top 1% of features ranked by linear regression *p*-values and random forest variable importance.

When a set of features was selected that satisfied all criteria, the (EICs) were visually inspected and those with poor peak morphology (ill-defined Gaussian shape) or integration were removed. Then, the three variable selection methods were repeated as needed to arrive at a final set of selected features of good quality peak morphology and integration.

Initially, only covariates on which the samples were matched (i.e. age and gender) were included in the models used for variable selection. We did not include other dietary or health related covariates because we did not want to obscure possible associations between the metabolites and case-control status. However, we did subsequently test for associations between the nine selected metabolites and the following covariates *weight, bmi, smoking status, and consumption of beef, pork, and alcohol* by adding each into the LASSO and random forest models (SI 2). None of these covariates was associated with the CRC outcome.

3.4 Results and Discussion

Using untargeted metabolomics in serum samples from 66 pairs of CRC cases and controls from the EPIC cohort, we sought evidence linking lipophilic molecules with the etiology of CRC. The LC-MS data collected from these samples included over 24,000 features. After filtering for noise (mean fold-change above blank samples), reproducibility (CVs and EIC peak morphology), and likely artifacts (CAMERA), 8,690 features were available for evaluating potential associations with CRC case status.

It has been standard practice in metabolomics to identify features that discriminate for case-control status using a multiple-testing approach, e.g., based on a cutoff for *p*-values that have been adjusted to control for a false positive error rate such as the FDR [135]. Since untargeted metabolomics can detect thousands of features, FDR correction is severe [135] and can drastically reduce the number of selected metabolites, thereby resulting in false negatives. Thus, we shifted our paradigm to a variable-selection approach, based on an ensemble of diverse regression methods, in order to uncover a reliable set of features for further investigation. After

applying the three variable-selection methods described above, two of which prioritize predictive ability (LASSO and random forest), nine features were selected. The volcano plot in Figure 3.1 relates case-control fold changes and $-\log_{10} p$-values (for the model in Equation [2]) for all features and highlights the nine selected metabolites (shown in Table 3.3). Case-control fold-changes ranged from approximately 0.2 to 3.0 overall and between 0.40 and 1.40 for the nine selected features. Due to the nature of our variable selection method, the *p*-values of the selected features were not necessarily the smallest, nor were their fold-changes necessarily the largest. Nevertheless, the nine selected metabolites resulted in a 79% correct classification rate when they were used to fit a logistic regression model on the learning set to predict case-control status. Although this correct classification is likely optimistic because the same data were used to perform the variable selection and to build and test the predictor, the selected features are worthy of validation in independent samples of CRC cases and controls from prospective cohorts.

Potentially Causal and Reactive Biomarkers

The nine selected features were evaluated to determine their associations with time to diagnosis (*ttd)* as a means of discerning whether they represent potentially causal exposures or reactive effects of disease progression [107]. If the log fold-change for a given feature was constant across the whole range of *ttd* in a linear model (*p*-value > 0.05; see supplemental information (SI) 1), the feature was classified as potentially causal (C) and if the case-control difference decreased with increasing *ttd*, the feature was classified as potentially reactive (R). These (C) and (R) classifications are listed in Table 3.3 for the nine selected features and the plots for the *ttd* linear models are shown in Figure 3.2. This process resulted in four potentially causal features (no apparent effect of *ttd* for IDs: 5080, 3207, 6054 & 839), four potentially reactive features (case-control differences diminish with *ttd* for IDs: 235, 4250, 4294 & 14963), and one that could not be classified as either (C) or (R) (case control differences increased with *ttd*, ID 5749). While the four potentially causal features may be linked to exposures that contribute to CRC, the four reactive features may be useful pre-diagnostic biomarkers.

Possible Annotations

Potential annotations of the nine selected features were based on comparisons of MS2 spectra with human metabolome database (HMDB) entries as summarized in Table 3.4. Focusing first on the four potentially causal features shown in Table 3.3, MS2 were only obtained for IDs 3207 & 6054, which were positively correlated (Pearson correlation coefficient of 0.64) and were both present at lower levels in cases than in controls, indicating possibly protective effects. The two MS2 fragment ions detected for ID 3207 could not be identified. However, ID 6054 had fragments characteristic of $[C_{16}H_{29}O_2]$- and losses of two H_2O molecules, consistent with the loss of two hydroxyl groups and a hexadecenoic acetate fragment. These fragments are suggestive of ceramide lipids [136]. Although IDs 3207 & 6054 and the two other potentially causal features (IDs 5080 & 839) could not be fully annotated, the identified characteristics of accurate mass, retention time and MS2 fragments can be used for validation in future studies.

Turning now to the likely reactive features, ID 4294 was putatively identified as ULCFA 468, which had been evaluated separately in our targeted study of 8 ULFCAs [107] and had been first reported by Ritchie et al. [4]. This feature had neutral losses of H_2O and CO_2, characteristic of a hydroxylated fatty acid and a likely molecular formula for $[M-H]^-$ of $C_{28}H_{52}O_5$, within 1.56 ppm of the exact mass. Another reactive feature (ID 4250) also displayed these characteristic neutral

losses of H_2O and CO_2 and was highly correlated with ID 4294 with a correlation coefficient of 0.85. This suggests that ID 4250 is a previously uncharacterized ULCFA with molecular formula for $[M-H]^-$ of $C_{27}H_{50}O_5$, within -2.76 ppm. While odd-numbered fatty acids are less common in humans, microbial single carbon metabolism in very long chain fatty acids has been reported [137]. As a class, ULCFAs tend to be present at higher levels among controls compared to paired cases, but this difference diminishes with *ttd*, suggesting that they result from disease progression [107]. Nonetheless, the fact that these two ULCFAs were selected from approximately 9000 features that survived filtering of the untargeted metabolomics data offers partial validation to our variable-selection strategy. Based on correlation maps (data not shown) both of these features clustered with five other ULCFAs that have been described by Ritchie, *et al*. [4] (ULCFAs 465, 466, 492, 518, and 538; exact masses within 10 ppm of calculated *m/z*), and were also analyzed in our targeted study [107]. Another selected feature (ID 235) exhibited similar reactive (R) behavior to the ULCFAs (Figure 3.2), and the presence of a neutral loss of CO_2, indicating that ID 235 may be a fatty or bile acid. Deoxycholic acid (3α, 12α-dihydroxy-5β-cholanic acid) $[M-H]^-$, chenodeoxycholic acid (3α, 7α-dihydroxy-5β-cholanic acid) $[M-H]^-$, and adrenic acid $[M+HAc-H]^-$ were eliminated as possible annotations of feature 235 by comparison of retention times between the experimental data and analytical standards. However, these two tested molecules are just two isomers of a large class of bile acids, some of which are positively correlated with CRC [109,138]; in our study, feature 235 was negatively correlated with CRC.

Limitations of this study include the small sample size, which reduced the power to detect differences between case-control pairs, and lack of information regarding aspirin consumption and a family history of CRC, two covariates that have been associated with CRC incidence [92,139]. The gelling of some samples from the cryostraw-storage process was a source of variation that could not be completely removed by adjustment in the linear model (Equation 1) and probably reduced our ability to detect differences between cases and controls. The storage of biological specimens for decades is challenging because preservation of cells, proteins, DNA, small molecules, and other biological molecules of interest must be considered. At the time our specimens were collected, in the early 1990's, state-of-the-art methods and materials for such storage were selected without knowing the types of analyses that would be performed in the future. However, decades later, shortcomings of then-contemporary technology (such as gelling of serum) may be revealed and their reporting may improve the design of future investigations.

3.5 Conclusion

In summary, of the nearly 9,000 filtered features subjected to statistical analysis, four appear to be potentially causal features that are worthy of following up in an independent set of prospective CRC cases and controls. When these four features alone were used to build a logistic regression predictor of case/control status on the learning set, they resulted in a correct classification rate of 72%. Again, this is likely an optimistic correct classification rate, but given that only four features were used for prediction, it is quite promising. Four other selected features, notably some ULCFAs and related fatty acids, appear to be products of disease progression and, therefore, could be useful diagnostic biomarkers for early detection of CRC. Since ULCFAs had previously been shown to discriminate CRC cases from controls in several cross-sectional investigations, it is reassuring that two putative ULCFAs (IDs 4294 and 4250) were selected as predictive features in this untargeted analysis. While the relatively modest

number of samples limited the power to detect stronger associations, the nine features selected in our study correctly predicted case-control status in 79% of the samples. The stability of these features across three disparate feature-selection methods is promising. Furthermore, based on *m/z* and annotation information, these nine features appear to be different than those reported in the prospective CRC study by Cross *et al* [109], warranting further identification and validation.

3.6 Tables and Figures

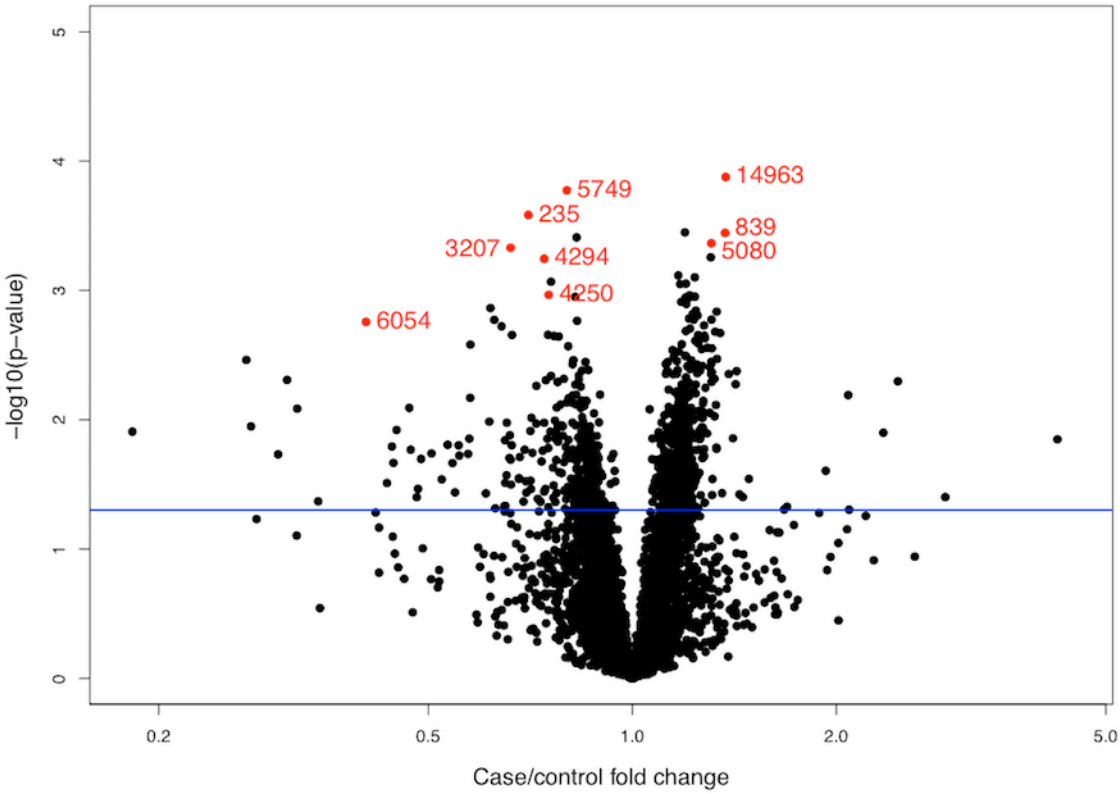

Figure 3.1: Volcano plot of analyzed features

The nine selected features are highlighted in red with the ID labels from Table 3.3. An arbitrary p-value = 0.05 threshold line is present for reference. p-values and fold-changes are calculated based on the regression model in Equation [2].

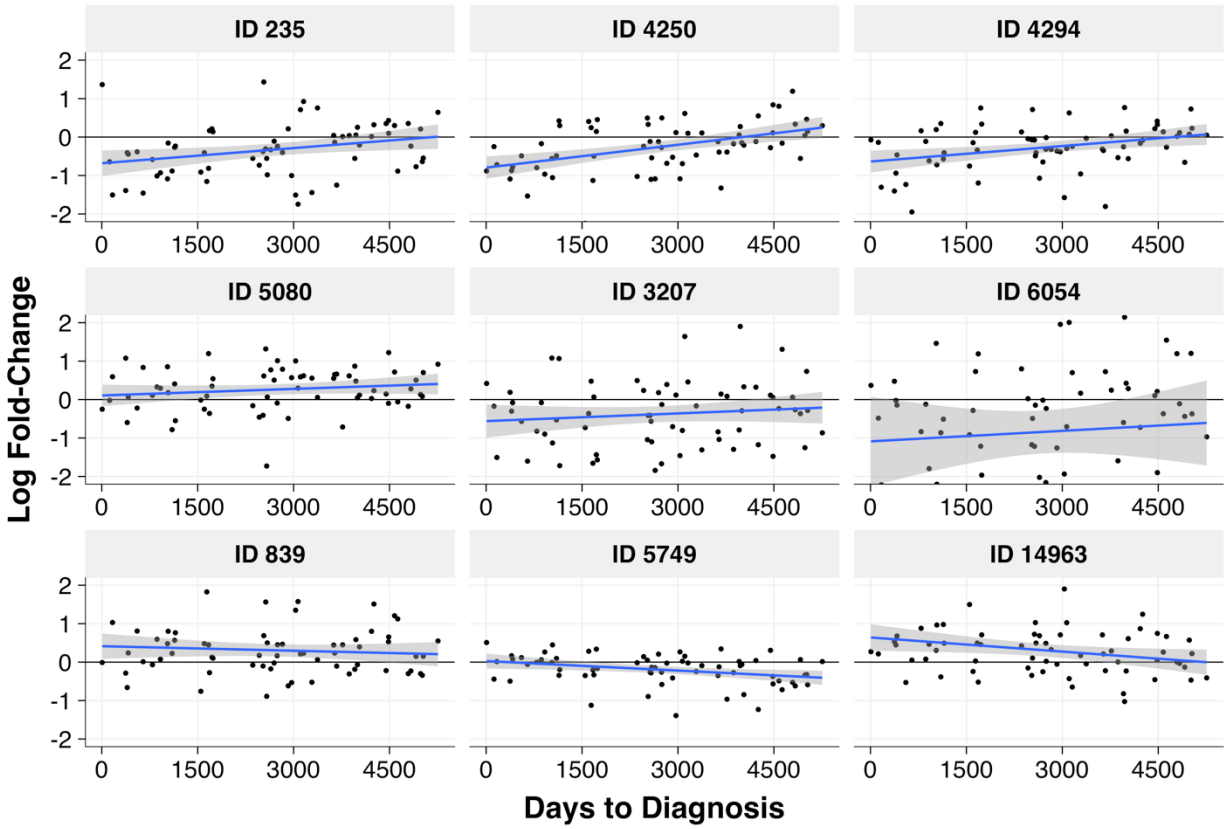

Figure 3.2: Scatterplots of case-control log fold-change vs. time to diagnosis (*ttd*) of the selected features

The blue line is the linear regression fit and the gray band represents a 95% confidence intervals, calculated with the 'lm' method of the R function 'geom_smooth' in the package 'ggplot2'.

Table 3.1: Studies that investigated associations of colorectal cancer with small molecules in plasma or serum from prospective cohorts

Cohort	Cases/ Controls	Follow-up[a] (y)	Analytical method	Design	Exposure variable	Likely associations	Ref.
WHI-OS	835/835	5.2	LC-MS	Targeted	Choline and its metabolites	TMAO (+); betaine/choline ratio (-)	[110]
EPIC [1]	1367/2323	3.7	LC-MS	Targeted	Methionine and choline metabolites	Methionine, choline, and betaine (-)	[111]
EPIC [2]	95/95	14.7	LC-MS	Targeted	8 Ultra-long-chain hydroxylated fatty acids	All associations (-) diminished with time to diagnosis (reverse causality)	[107]
EPIC [3]	1238/1238	3.8	Colorimetry & turbidimetry	Targeted	Triglycerides, cholesterol, and lipoproteins	HDL (-)	[106]
PLCO	254/254	7.8	LC-MS & GC-MS	Semi-targeted	278 Annotated metabolites detected in >80% of specimens	Glycochenodeoxycholate (+) in women but not men	[109]

Legend: WHI-OS, Women's Health Initiative Observational Study; EPIC, European Prospective Investigation into Cancer; GC-MS, gas chromatography-mass spectrometry; HDL, high-density lipoprotein cholesterol; LC-MS, liquid chromatography-mass spectrometry; PLCO, Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial; TMAO, trimethylamine-*N*-oxide; WHI-OS, Women's Health Initiative-Observational Study, (+), positively associated with CRC; (-), negatively associated with CRC.

[a] Mean period of follow-up

Table 3.2: Descriptive statistics of human

	Total *n*=132	**CRC cases *n*=66**	**Controls *n*=66**	***p*-value***
Gender	Male	51	51	
	Female	15	15	
Age	median	56	56	
at enrollment	min	35	35	
(y)	max	65	65	
Years to	median	7.52	-	
diagnosis	min	0.10	-	
	max	14.40	-	
BMI	median	26.9	25.3	0.0322
	min	19.7	18.7	
	max	36.7	33.6	
Waist	median	97	90	0.0005
circumference	min	68	66	
(cm)	max	115	119	
Diabetes	yes	2	2	
	no	64	64	
Smoking Status	current	15	16	
	former	27	23	
	never	21	22	
	NA	3	5	
Alcohol	median	23.0	22.6	0.4617
consumption	min	0.0	0.1	
(ml/day)	max	79.8	113	
Physical Activity	high	13	13	
(min/day)	medium	15	20	
	low	25	18	
	none	10	10	
	NA	3	5	
Total meat	median	75.6	67.6	0.4488
consumption	min	5.9	8.8	
(g/day)	max	248.3	201.3	

NA – not available

* Nominal *p*-values calculated from a two-sided *t*-test.

Subjects matched by age, study enrollment year and season, and gender, and selected covariates

Table 3.3: Untargeted features selected as predictors of case-control status

Feature ID	Observed *m/z* [a, b]	Ret. time (sec)	Fold change	*p*-value	Feature type [c]
235	391.2832	596.4	0.702	0.000261	R
4250	453.3592	605.5	0.753	0.001082	R
4294	467.3744	605.6	0.741	0.000569	R
5080	519.1965	595.9	1.308	0.000432	C
3207	531.1558	563.9	0.661	0.000468	C
6054	551.1781	563.9	0.404	0.001750	C
839	577.2698	620.6	1.370	0.000359	C
5749	882.6393	718.2	0.880	0.000168	I
14963	907.4806	617.6	1.373	0.000133	R

Legend: *m/z* – mass-to-charge ratio; *p*-value from the regression model (Equation [2]); C- potentially causal feature; R -potentially reactive feature; I, indeterminate.

[a] Observed *m/z* values correspond to singly-charged negative ions.

[b] Feature selected by bootstrap LASSO and by being in the top 1% of features ranked by both the *p*-values from the case-control regression (Equation [2]) and the random forest variable importance measure.

[c] Based on regression of case-control difference on time to diagnosis (*ttd*, Figure 3.2).

Table 3.4: Results of tandem MS/MS analyses of features associated with case-control status

Feature ID	Observed *m/z*	Ret. time (s)	Prominent MS2 fragments with possible fragment IDs	Putative ID	Species	Molecular formula	Δ ppm
235	391.2832	596.4	347.2961(loss of CO_2), 197.0725 ([$C_{10}H_{13}O_4$]-)	possible fatty acid	[M-H]-	$C_{24}H_{40}O_4$	4.22
4250	453.3592	605.5	59.0131, 409.3687 (loss of CO_2), 391.3568 (loss of CO_2 and H_2O), 279.2336 ([$C_{18}H_{31}O_2$]-), 435.3462 (loss of H_2O)	possible ULCFA	[M-H]-	$C_{27}H_{50}O_5$	-2.76
4294	467.3744	605.6	449.3639 (loss of H_2O), 263.2368 ([$C_{18}H_{31}O$]-), 423. (loss of CO_2), 162.8392, 405.3724 (loss of CO_2 and	ULCFA 468	[M-H]-	$C_{28}H_{52}O_5$	-1.56
5080	519.1965	595.9	No MS2 spectra	unknown			
3207	531.1558	563.9	481.3110, 256.2357	unknown			
6054	551.1781	612.3	478.2903, 515.1326 (loss of 2 H_2O;), 253.2165 ([$C_{16}H_{29}O_2$]-)	unknown			
839	577.2698	596.8	No MS2 spectra	unknown			
5749	882.6393	718.2	124.0075, 822.6453 (loss of acetate)	unknown	[M+HAc-H]-		
14963	907.4806	617.6	No MS2 spectra	unknown			

Chapter 4: Lipid and Cys 34 Adduct Multi-Omic Correlation of Smoking and Non-Smoking Subjects

4.1 Introduction

The use of data from combinations of -omics technologies (genomics, transcriptomics, proteomics, metabolomics, etc.) provides opportunities to relate population differences to biological pathways [140]. For example, the interleafing of data from genome-wide association studies (GWAS) and metabolomic analyses has provided mechanistic insights into genetic influences on disease processes [141–143]. However, because heritable genetics contribute only modestly to the incidence of chronic diseases [95], multi-omics analyses may be more fruitfully applied to factors related to health-impairing exposures received by populations from both endogenous and exogenous sources. Information about such exposures can be captured by untargeted analyses of circulating small molecules and reactive intermediates that are generated from inflammatory and metabolic processes. Whereas metabolomic and lipidomic platforms can readily be applied to characterize small stable molecules, parallel untargeted analyses of reactive intermediates can focus on modifications to nucleophilic loci of proteins and DNA using approaches called 'adductomics' [144–146]. Our laboratory has been exploiting an adductomic pipeline that characterizes all modifications to Cys34 of human serum albumin (HSA), which is an important scavenger of small reactive molecules in the body [144].

Here, we combined adductomic and lipidomic data from plasma samples representing 158 healthy volunteer subjects that had been pooled by smoking status, race and gender [144,147]. From a set of 33 pooled samples common to both studies, nearly 3,000 lipidomic features had been measured along with 43 Cys34 (and related) adducts. In the original studies, several significant associations had been detected between particular lipids or adducts and covariates, namely, smoking status, race, gender, body mass index (bmi) and consumption of animal and vegetable fats [144,147]. In this combined analysis, we mapped correlations of lipids and adducts to gain mechanistic insights into potential pathways involving both types of molecules and their connections to smoking and other covariates. Using regularized regression (LASSO), we also analyzed the data for robust lipid-adduct associations. Since some Cys34 adducts – notably sulfoxidation products and mixed disulfides - reflect the redox biology of the Cys proteome [148–150], particular attention was paid to connections between these adducts and lipid products of oxidative damage and inflammation.

4.2 Methods

Plasma specimens

Both lipidomic and adductomic data were collected for a set of pooled plasma samples prepared from 158 young healthy subjects [144,147]. The samples had been collected approximately 13 years earlier, with informed consent, from non-fasting subjects in a previous study [13], and stored at –80 °C until the two sets of analyses were performed.

Demographic characteristics, including race, age, height and weight were obtained with a standardized questionnaire at the time of phlebotomy. Smoking status was based upon current smoking (yes/no) or the number of cigarettes smoked per day (for correlation analysis). A semi-quantitative food-frequency questionnaire containing 131 items was used to evaluate average daily consumption of fat (animal, vegetable and cholesterol) over the past six months for each individual [14,15]. All dietary-intake values were compiled at the Channing Laboratory, Harvard Medical School [16,17].

Instrumental analysis of lipids and adducts.

Samples of plasma were pooled by combining aliquots from four to six randomly selected subjects stratified by smoking status (smoker/non-smoker), race (black/white), and gender. Pooling of these specimens was required to ensure anonymity of subjects.

Lipids

For the lipid analysis, serum was extracted as described previously [151]. Briefly, lipophilic molecules were extracted from 10 µL of each of the 34 pooled plasma sample with 2:1:1 chloroform:methanol:phosphate buffer. The organic layer was removed, dried under N_2, and re-suspended in chloroform for LC-MS analysis. Liquid chromatography mass spectrometry (LC-MS) analysis was performed on the lipid extracts with a Surveyor LC and an LTQ-FTMS with a heated electrospray ionization source (ESI) (Thermo Fisher Scientific, Waltham, MA). The MS was operated in both positive and negative modes with untargeted data collected from 100 to 1200 *m/z*. For LC separation, a Luna C5 column (4.6×50 mm, 100 Å, 5 μm, Phenomenex, Los Angeles, CA) eluted all potential lipids. Injection volumes were 20 μl and 25 μl for ESI+ ionization and ESI- ionization, respectively. Mobile phases contained 0.1% formic acid for ESI+ ionization and 0.1% ammonium hydroxide for ESI- ionization. The column was eluted with a gradient of mobile phase A (methanol:50 mM ammonium formate 5:95) and mobile phase B (isopropanol:methanol:50 mM ammonium formate 60:35:5), with gradient and mass spectrometer details published elsewhere [147].

HSA Adducts

To isolate human serum albumin (HSA) for adduct analysis, 5 µL of each pooled plasma specimen was added to 60 µL of 50% methanol. Precipitates were removed and the samples were diluted with four volumes of digestion buffer [50 mM TEAB buffer containing 1 mM EDTA (pH 8.0)]. Purified HSA in digestion buffer containing 10% methanol was transferred to a digestion vessel (MT-96, Pressure Bio- sciences Inc., South Easton, MA) and 1 µL of 10 mg/mL trypsin was added. Proteolytic digestion was performed at 37 °C using a pressurized system (NEP2320, Pressure Biosciences Inc., South Easton, MA) that cycled between ambient pressure (15 s) and 138 mPa (45 s) for 30 min. After digestion, 3 µL of 10% formic acid was added and samples were immediately centrifuged to precipitate trypsin and protein aggregates. A 40-µL aliquot of each digest was diluted to a final volume of 120 µL with an aqueous solution containing 2% acetonitrile and 0.2% formic acid.

Technical replicates (2) of the samples were analyzed with an isotopically labeled internal standard. LC–HRMS Analysis. Freshly digested samples were analyzed with a Dionex Ultimate 3000 nanoflow LC system via a Flex Ion nanoelectrospray-ionization source (Thermo Fisher Scientific) and an LTQ Orbitrap XL HRMS. The peptides were separated on a Dionex PepSwift monolithic nanoflow column (100-µm i.d. × 25 cm) (Thermo Scientific, Sunnyvale, CA), operated at room temperature with a flow rate of 750 nL/min. Mobile phase A was water/0.1% formic acid and mobile phase B was acetonitrile/0.1% formic acid. Complete method details are available elsewhere [144].

Previous results

The lipidomic experiments resulted in over 3000 features (2862 positive, 717 negative mode features) [147]. These features were sum-intensity normalized, evaluated with the covariates (linear model), and adjusted for the false discover rate (Benjamini-Hochberg) [22,147]. These analyses resulted in 34 features of interest, that were significantly associated with gender, race, and/or smoking status [147].

As previously noted, the adduct data was collected in duplicate injections and adjusted for batch effects with a mixed-effects model [144]. Using this model, adjusted values for 43 adducts were calculated. The adducts that passed initial quality control checks, along with dietary and demographic covariates, were subjected to Wilcoxon rank sum exact tests and multivariate analyses. Several of these adducts were associated with smoking status, while a smaller number were associated with race, genders, bmi, or dietary vegetable and animal fats [144].

The normalized and filtered –OMIC variables were combined for the 33 samples common to both the adduct and lipid data collections. Other continuous variables, including food frequency questionnaire data of specific lipids and lipid classes, along with bmi, age, and cigarettes per day, were also included for analysis. Statistical analyses reveal new associations between the lipids and adducts, which may better explain possible functions or pathways of these variables.

4.3 Results

Correlation map.

A network visualization program (Cytoscape, [130]) was used to depict relationships between analyzed variables. Specifically, the analyzed variables are nodes that were linked to other nodes with an edge for all pairs with an absolute value of the Pearson correlation coefficient greater than or equal to 0.65. The nodes were defined by the four classes of variables: adducts, lipidomic features, dietary lipids (animal fats, vegetable fats, cholesterol), and demographic features (including age, bmi, and cigarettes per day). Generally, nodes were included in the map if they met the correlation threshold with at least one other variable. However, due to the large number of variables within the dietary and LCMS lipids, these nodes had an additional criterion to only highlight features relevant to this integrative analysis. The dietary lipids were only included if they met the correlation threshold with at least one variable from a different class (demographic, adduct, or LCMS lipid). Likewise, LCMS lipids were only included if they met the correlation threshold for a variable from one of the other classes. These relationships are illustrated in Figure 4.1.

Regularized Regression

For consistent normalization, the data distribution levels of the adducts were re-examined. Of the 43 adducts, 34 were determined to have satisfactory distributions for analysis of correlation with levels of the lipidomic features. To select lipids that are potentially related to each adduct, a regularized regression (LASSO) was performed [117,132], with the adduct level as the outcome variable and all of the normalized log intensities of the lipids as the independent variables (~9000 variables).

It has been shown that in data sets with large numbers of highly correlated independent variables, LASSO can potentially be unstable as a variable selection method [132,152]. Therefore, to stabilize the variable selection, LASSO regression was performed on 500 bootstrap samples of the data, for each of a variety of penalty parameters. The percentage of times each lipid is selected in the 500 bootstrapped samples is a measure of its variable importance. Lipids were selected as important if they were selected by the LASSO regression for at least 60% of the bootstrap samples across a variety of penalty parameters. This cutoff is somewhat flexible, and the final results are robust to cutoffs between 40%-60% since there was always a clear jump in the percentage of times the top features were selected as opposed to the other lipids in the model.

For 13 of the adducts, between one and three lipids were selected by the bootstrap LASSO method, while the other adducts had no lipids selected. Next, for each of these 13 adducts, a linear regression was fit, with the adduct abundance as the dependent variable regressed on its corresponding lipids selected from the LASSO regression. The *p*-values from the coefficients of the lipids in each model provide some measure of significance for the association (Table 4.1). However, these *p*-values are likely to be overly optimistic because the same data were used to perform the variable selection and to calculate the significance of the coefficients of the lipid variables. Finally, the Pearson correlations are shown in Table 4.1 for lipids that had relatively high correlation with their corresponding adducts.

The LCMS lipidomic data from the original study [147] were reanalyzed for the current application. For this analysis, putative identification relied upon matching accurate masses from FTMS (with a mass tolerance of 10 ppm) with entries in the Human Metabolome Database (HMDB) (http://www.hmdb.ca/), the Structure Database of Lipid Maps (LMSD) (http://www.lipidmaps. org) and the Metabolite and Tandem MS Database (METLIN) (http:// metlin.scripps.edu/). Most of the putative annotations are based on the accurate masses which matched with species from different lipid classes: monoacylglycerols (MAG), diacylglycerols (DG), triacylglycerols (TG), glycerophosphocholines (PC), glycerophosphoethanolamines, (PE), monoglycerophosphocholines (LysoPC), and sphingomyelins (SM). The putative species, formulae, and ppm differences are listed in Table 4.2. However, further experiments and comparison with reference standards would be required for conclusive identifications.

4.4 Discussion

Lipids and adduct associations are summarized in Figure 4.1 and Tables 4.1 and 4.2. The results from the correlation map (Figure 1) and bootstrap LASSO complement and supplement each other. With a correlation threshold of 0.65 for the correlation map, some of the lipids selected by LASSO were excluded. Conversely, some features with higher correlations are included in the correlation map but were not selected by LASSO. This indicates that a high Pearson correlation coefficient is not necessarily indicative of a significant association as measured by the regularized regression and vice versa. Nonetheless, relationships found in both the correlation map and regularized regression can reinforce one another and are worthy of investigation.

The correlation map also compared other continuous variables, such as dietary lipids, cigarettes per day, age, and bmi. Adduct-lipid relationships that met the 0.65 threshold and were also selected by LASSO are represented with a dashed edge in Figure 4.1. However only cigarettes per day and some dietary fats were associated with lipids with this threshold. Here, we examine potential biological explanations for several of the observed associations.

Dietary lipids

Some dietary lipids were correlated particular lipids, as shown in the correlation map. Summed omega-3, eicosapentaenoic, and docosohexanoic acids were positively correlated with a triglyceride (TG (56:11)). Triglycerides have been observed in subjects who consumed fish oil, a well-known source of omega-3 fatty acids [153,154]. Various dietary lipids, including animal fat, cholesterol, and palmitic, oleic, and arachidonic acids were all correlated to a putative phophatidylethanolamine (PE (36:4)) based on accurate mass. Although dietary lipids were not highly correlated with any of the analyzed Cys34 adducts, dietary animal and vegetable fats were associated with several adducts in the previous multivariable analyses [144]. Unlike the earlier analyses of these two data sets, the current investigation focused on correlation and regression analyses with continuous variables and did not adjust for binary covariates of gender, race, and smoking status as in the previous analyses.

Smoking, oxidation products, and sphingomyelins

The Cys34 sulfoxidation products (A9, A12, and A15) were highly correlated with each other and were previously reported to be present at lower levels in smokers [144,155]. The lipidomic feature n_783.64208_22.037, which was also previously reported to be lower in smokers [147], has an accurate mass that matches sphingomyelin (SM) (40:2). This putative sphingomyelin is associated with the sulfinic acid Cys34 sulfoxidation product (A12), with a Pearson correlation coefficient of 0.659. This lipid-adduct relationship was confirmed via the robust selection process of regularized regression. In fact, the LASSO procedure selected this same presumptive sphingomyelin as the second variable for the three sulfoxidation product adducts (A9, A12, and A15).

Another Cys34 sulfoxidation product, the cysteine-glycine crosslink (A1), was correlated with A9, A12, and another possible disulfide adduct (A34). This adduct was positively correlated with the lipid feature n_785.65799_22.189, which has an accurate mass matching another sphingomyelin (SM (40:1)). Like the above SM-adduct relationship, SM (40:1) was selected as variable for adduct A1 by LASSO, and the two lipids were highly correlated (0.694).

It is understood that cigarette smoking introduces an abundance of reactive oxygen species (ROS), and is associated with diseases that involve oxidative mechanisms (colorectal cancer, atherosclerosis, chronic obstructive pulmonary disease (COPD)) [156–159]. The ROS from cigarette smoke can also lead to the peroxidation of unsaturated lipids, like unsaturated sphingomyelins, and this could potentially diminish their abundance among smokers. In addition, sphingomyelinases (SMases) hydrolyze sphingomyelins (SMs) into ceramides via an oxidative method [160] and ceramides are linked to stress-induced apoptosis [160,161]. Oxidative stress, particularly from cigarette smoke, activates SMases causing increased hydrolysis of SMs into apoptotic ceramides [156,158,160]. Apoptosis from this mechanism may explain the epithelial lung cell death observed in smoking-related diseases such as COPD and lung cancer [158]. In fact, SM depletion has been observed in apoptotic cells [162]. This SMase pathway, the observed decrease of presumptive SM (40:2) with the increase of number of cigarettes smoked, and positive relationships of these SMs with sulfoxidation adducts (A1, A9, A12, and A15) that were also lower in smokers, present an intriguing pattern.

Since A12, the sulfinic acid oxidation product, is positively correlated with the selected SM (40:2) and inversely correlated to cigarettes per day; and A9, the sulfonamide oxidation product, is positively correlated to other Cys34 oxidation products (A1, A12, and A15), the Cys34 sulfoxidation products are again associated with lower levels of smoking. This relationship is counterintuitive because ROS are introduced with cigarette smoke and it would be reasonable to expect that Cys34 oxidation products would also increase. Yet, the opposite was observed in multiple regression models from the previous adductomic study of these samples [144]. Theories as to why the products were lower in smokers include smoking-associated hypoxia or perturbations to the redox proteome [144,163,164]. Another possibility is that sulfoxidation precursors are being diverted and, like the sphingomyelins, are converted into other products as a result of cigarette smoke exposure.

While no single specific adduct was strongly and negatively correlated to the Cys34 sulfoxidation products, there could be several reactions consuming the transient sulfenic acid intermediate in the presence of cigarette smoke. These may result in disulfides or, less likely, other ROS-produced adducts that were higher in smokers [144]. Although these sulfoxidation products were lower in smokers, they were recently observed to be higher in non-smoking workers exposed to benzene [Grigoryan, et. al, 2017, submitted]. A smoking-specific diversion of sulfoxidation precursors could be a partial explanation of the inverse relationship between the sulfoxidation products and smoking, compared to the proportional association seen among benzene exposed workers.

Diacylglycerols and smoking-related adducts

Also interesting is the high correlation between a putative diacylglycerol (DG) (DG (41:2)) and the sulfoxidation product A9, which was corroborated by LASSO feature selection. Another putative DG, (DG (36:2)), was inversely correlated (Pearson's coefficient = -0.658) to an unknown adduct (A31) that was higher in smokers [144]. Dietary DGs, found in vegetable oils such as rapeseed and soy oils, are highly correlated with DGs in serum [165]. Furthermore, DGs are protective against obesity and diabetes as they are correlated with lower levels of serum triglycerides after meals [166]. Since obesity and diabetes are chronic diseases that involve oxidative stress, it is interesting that these particular DGs would both be protective for diabetes and be positively correlated with oxidation product adducts related to smoking. From the secondary relationships of the DGs to smoking via the oxidation products, it would seem that DGs are lower in smokers, and higher in non-smokers. Perhaps the oxidative stress from smoking decreases baseline levels of these compounds, possibly through the oxidative processes connected to Cys34 adducts. In any case, further work is needed to elucidate these relationships.

Other associations of Adduct A31 with PCs and PEs

As previously noted, A31 is an unknown Cys34 adduct, likely a disulfide, that was more abundant in smokers compared to non-smokers [144]. Besides putative DG (36:2), two other features were negatively correlated with A31. These had accurate masses within 8 ppm of putative phosphatidylcholines or phophatidylethanolamines. Since the mass accuracy was technically within 10 ppm, further analysis would be required before positing possible identities of these features. One of the features was highly correlated and selected by LASSO, so this could be an interesting relationship to elucidate.

LycoPCs inversely correlated with unmodified T3

Unmodified T3 is the peptide representing the unmodified form of Cys34 (mercaptoalbumin), and is the precursor of Cys34 adducts. Theoretically, higher levels of unmodified T3 could be an inverse marker for oxidative stress, as it may indicate lower levels of ROS. Unmodified T3 (A7) was inversely correlated with three lipid features that had accurate masses of a lysophosphatidylcholine (LysoPC), the LysoPC precursor phosphatidylcholine (PCs), and/or a formic acid adduct of these lipids. Dysregulation of LysoPCs have been used as biomarkers of chronic diseases linked to oxidative stress [167–169]. Several LysoPCs, including LysoPC (18:2), the formic acid adduct detected as n_564.33124_18.570, were lower in lung cancer cases compared to healthy controls [168], whereas other LysoPCs were elevated in ovarian cancer patients, relative to controls [169]. Given the complicated association of LysoPCs with diseases linked to oxidative stress and the ambiguous meaning of unmodified T3 (A7), the negative correlation of these to unmodified T3 is interesting. Furthermore, the correlation with A7 and a methylated non-Cys34 adduct (A10) in light these lower amounts of LysoPCs may have implications for the role of this methylated adduct which was also elevated in smokers.

Limitations

Older samples which have been frozen and thawed several times, as these samples were, may have led to increases in LysoPCs from the oxidation of other lipids [170]. As LysoPCs are of biological interest, this may have had an effect on the detected amounts of some of the LysoPCs. However, since all of these samples were treated uniformly, the bias from this oxidation should be minimal. Our power was also limited due to the pooling of these samples.

This analysis focused on the relationship between levels of Cys34 adducts and lipids as well as to a few continuous covariates, such as cigarettes/day. Some of the analyses from the previous two studies referred to analysis of dichotomous covariates that were used to adjust for race and gender [144,147], and these variables were not included in the current analyses.

4.5 Tables and Figures

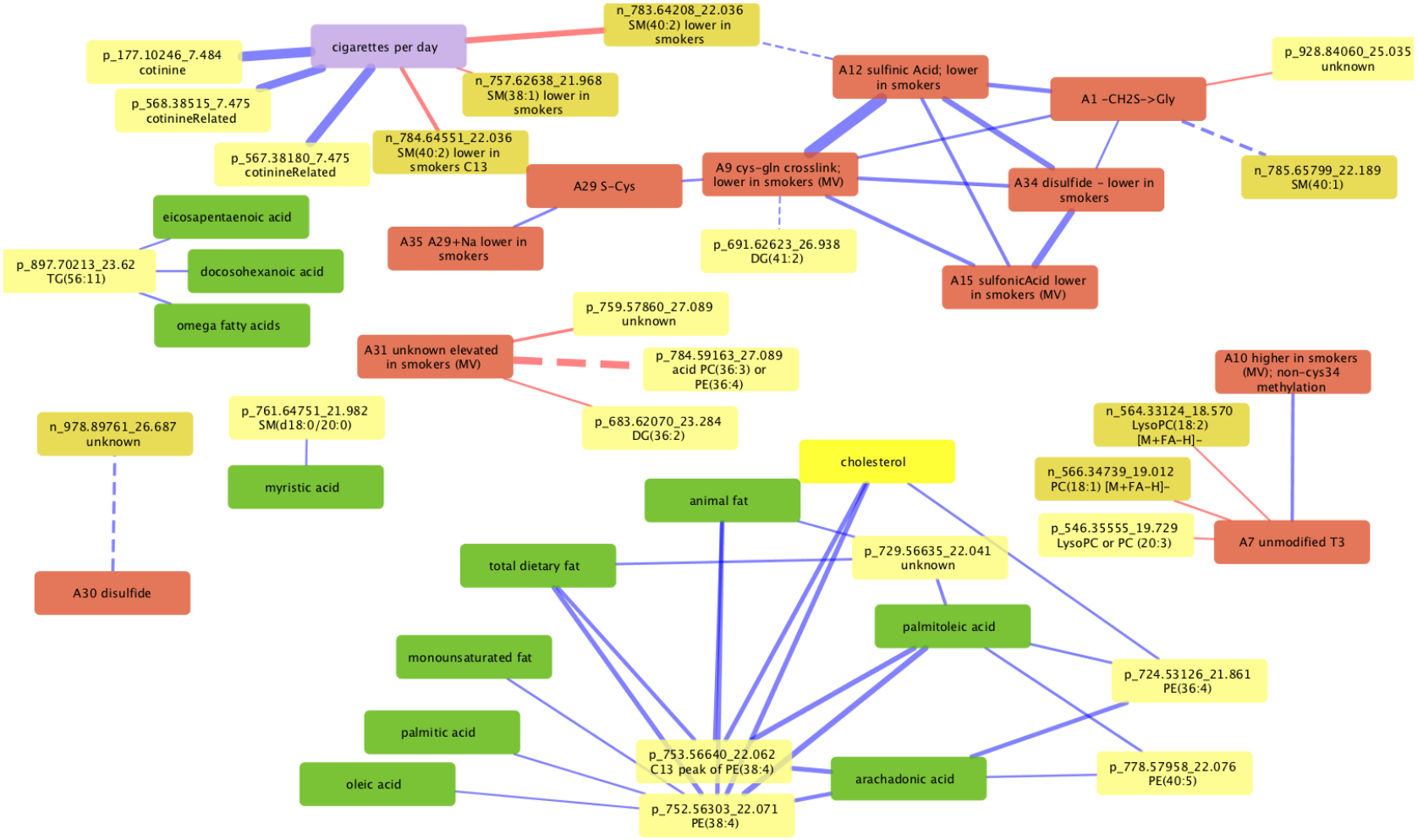

Figure 4.1: A correlation network made with Cytoscape.

Correlations (Pearson's) of adducts (orange), dietary lipids (green), lipidomic features (yellow), and cigarette per day (purple) +/- 0.65 or higher. Negative (red) and positive (blue) correlations are edges. Strength of correlation proportional with edge width. Correlations also selected by LASSO are dashed lines.

Table 4.1: List of each selected LASSO-MS Lipid relationship, grouped by adduct.

Adduct	Adduct annotation	MS Lipid[a]	R	*p*-value[b]	Possible lipid annotation and MS species[c]
A1	-CH2S; Cys34->Gly	n785.65799_22.2	0.694	0.000042	[SM(40:1)-H]-
A1	-CH2S; Cys34->Gly	p379.28305_20.1	-0.563	0.003250	[MAG(20:4)+H]+
A5	not Cys34 adduct	p664.60662_24.5	0.632	0.000081	[CE(18:3)+NH4]+
A7	unmodified T3	p426.35934_19.1	-0.648	0.001680	a carnitine
A7	unmodified T3	p429.24182_14.3	0.593	0.004150	Formula only: $[C_{22}H_{37}O_6P+H]+$
A7	unmodified T3	p898.74192_24.3	-0.573	0.008140	unknown
A9	-H2, +O; Cys-34-Gln Xlink	n783.64208_22.0	0.642	0.005652	SM(40:2)-H]-
A9	-H2, +O; Cys-34-Gln Xlink	n856.59458_21.9	-0.519	0.000947	[PC(42:8)-H]-
A9	-H2, +O; Cys-34-Gln Xlink	p691.62623_26.9	0.650	0.004159	[DG(41:2)+H]+
A10	+CH3; methylation; not Cys34	n556.31840_19.0	-0.650	0.012300	unknown
A10	+CH3; methylation; not Cys34	n618.48099_18.0	-0.640	0.019500	unknown
A12	+ HO2; Cys34 sulfinicAcid	n783.64208_22.0	0.659	0.003880	SM(40:2)-H]-
A12	+ HO2; Cys34 sulfinicAcid	p691.62623_26.9	0.610	0.022410	[DG(41:2)+H]+
A15	+ HO3; Cys34 sulfonicAcid	n783.64208_22.0	0.638	0.011000	SM(40:2)-H]-
A15	+ HO3; Cys34 sulfonicAcid	p798.64391_22.1	0.582	0.070000	[PC 38:2-H]- 8.6 ppm
A23	+ C2H3O2S; *S*-mercaptoacetic acid	p729.56635_22.0	0.640	0.003910	unknown
A28	+ C4H6NOS; *S*-hCys (-H2O)	n858.79774_7.6	-0.630	0.000094	unknown
A29	+ C3H6NOS; *S*-hCys	n783.64208_22.0	0.573	0.000500	SM(40:2)-H]-
A30	+C3H5O3S; *S*-Cys (NH2 ->OH)	n978.89761_26.7	0.686	0.000011	unknown
A31	unknown	p784.59163_27.089	-0.790	0.000033	unknown
A34	+ C4H9O3S; unknown	p379.28305_20.1	-0.600	0.000248	[MAG(20:4)+H]+
A35	A29 + Na; Na adduct of A29	p572.39763_17.3	-0.600	0.000276	unknown

[a] The MS lipid feature begins with n or p, for a feature detected in negative mode or positive mode, respectively. The next 8 digits are the observed *m/z* to the fifth decimal position, followed by "_" and 3-5 digits, the retention time in minutes.

[b] Values from the final linear model after LASSO selection which ignores the variability added with the model selection

[c] See Table 4.2 for more complete descriptions of the possible lipid annotation

Table 4.2: List of each MS Lipid relationship, LASSO or correlation map and possible annotations, grouped by lipid.

Lipidomic feature	Correlates with	Map or LASSO	Name	Formula	Observed *m/z*	Theoretic *m/z*	ppm
n_556.31840_19.045	A10 (-0.65)	L	unknown	unknown	556.3184	566.3458	n/a
n_564.33124_18.570	A7 (-0.654)	M	LysoPC(18:2) [M+FA-H]-	$C_{26}H_{50}NO_7P$	564.3312	564.3307	0.95
n_566.34739_19.012	A7 (-0.659)	M	PC(18:1) [M+FA-H]-	$C_{26}H_{52}NO_7P$	566.3474	566.3458	2.81
n_618.48099_18.089	A10 (-0.64)	L	unknown	unknown	618.4801	unknown	n/a
n_757.62638_21.968	cigsPerDay	M	SM(38:1)	$C_{43}H_{87}N_2O_6P$	757.6264	757.6230	4.49
n_783.64208_22.036	A12 (0.659) & cigsPerDay (-0.762)	M & L	SM(40:2)	$C_{45}H_{89}N_2O_6P$	783.6421	783.6386	4.44
n_783.64208_22.036	A15 (0.638)	L	SM(40:2)	$C_{45}H_{89}N_2O_6P$	783.6421	783.6386	4.44
n_783.64208_22.036	A9 (0.642)	L	SM(40:2)	$C_{45}H_{89}N_2O_6P$	783.6421	783.6386	4.44
n_784.64551_22.036	cigsPerDay	M	SM(40:2) isotope peak	$C_{45}H_{90}N_2O_6P$	784.6455	784.6465	1.20
n_785.65799_22.189	A1 (0.694)	M & L	SM(40:1)	$C_{45}H_{91}N_2O_6P$	785.6580	785.6543	4.70
n_856.59458_21.852	A9 (-0.519)	L	PC(42:8)	$C_{50}H_{84}NO_8P$	856.5946	856.5863	9.67
n_858.79774_7.613	A28 (-0.63)	L	unknown	unknown	unknown	unknown	n/a
n_978.89761_26.687	A30 (0.686)	M & L	unknown	unknown	unknown	unknown	n/a
p_379.28305_20.086	A1 (-.563)	L	MG(20:4)	$C_{23}H_{38}O_4$	379.2831	379.2842	3.06
p_379.28305_20.086	A34 (-0.6)	L	MG(20:4)	$C_{23}H_{38}O_4$	379.2831	379.2842	3.06
p_426.35934_19.096	A7 (-0.648)	L	a carnitine	$C_{25}H_{47}NO_4$	426.35934	426.3577	3.83
p_429.24182_14.314	A7 (0.593)	L	unknown- possible formula	$C_{22}H_{37}O_6P$	429.24182	429.2401	4.01
p_546.35555_19.729	A7 (-0.654)	M	LysoPC(203)	$C_{28}H_{52}NO_7P$	546.3556	546.3553	0.39
p_546.35555_19.729	A7 (-0.654)	M	PC(20:3)	$C_{28}H_{52}NO_7P$	546.3556	546.3553	0.39
p_572.39763_17.32	A35 (-0.6)	L	unknown	unknown	unknown	unknown	n/a
p_664.60662_24.485	A5 (0.632)	L	unknown	unknown	unknown	unknown	n/a
p_683.62070_23.284	A31 (0.-658)	M	DG(36:1) [M+IsoProp+H]	$C_{39}H_{74}O_5$	683.6207	683.619	2.49
p_691.62623_26.938	A12 (0.61)	L	DG(41:2)	$C_{44}H_{82}O_5$	691.6262	691.6234	4.09
p_691.62623_26.938	A9 (0.65)	M & L	DG(41:2)	$C_{44}H_{82}O_5$	691.6262	691.6234	4.09

p_729.56635_22.041	various lipids; animal fat	M	unknown	unknown	unknown	unknown	n/a
p_729.56635_22.041	A23 (0.64)	L	unknown	unknown	unknown	unknown	n/a
p_752.56303_22.071	saturated/animal fats	M	PE(38:4)	$C_{43}H_{78}NO_7P$	752.5630	752.5589	5.49
p_759.57860_27.089	A31 (-0.681)	M	unknown	unknown	759.5786	unknown	n/a
p_784.59163_27.089	A31 (-0.79)	M & L	PC(36:3) or PE(36:4) [M-IPA+H]+	$C_{44}H_{82}NO_8P$	784.5916	784.5851	8.33
p_798.64391_22.103	A15 (0.582)	L	PC(38:2)	unknown	798.6439	798.6370	8.60
p_897.70213_23.627	omega fatty acids	M	TG(56:11)	$C_{59}H_{92}O_6$	897.7021	897.6966	6.17
p_898.74192_24.301	A7 (-0.573)	L	unknown	unknown	unknown	unknown	n/a
p_928.84060_25.035	A1 (-0.659)	M	unknown	unknown	unknown	unknown	n/a

Gray-shaded areas represent lipids with multiple relationships or possible annotations

Chapter 5: Conclusions

Smoking and air quality severely impact public health [97,171] by contributing to chronic illnesses, including cancer and cardiovascular disease [45,46]. Biomarker discovery using lipidomics and other high-dimensional, biological methods (e.g., genomics, proteomics, adductomics, etc.) offer a unique opportunity to detect changes in small molecules, genes, and proteins that may be related to environmental exposures and disease endpoints. Applying these methods to prospective samples affords the opportunity to reveal biological differences that exist in subjects prior to the onset of disease symptoms.

While these methods offer great potential, there are still several limitations. False positives may plague high dimensional –OMIC analyses due to multiple hypothesis testing with relatively small sample sizes. Adjusting for false discovery rates may reduce the number of false positive findings but also potentially exclude meaningful associations that may not otherwise be detected due to limited statistical power. Here, we attempted to identify meaningful serum biomarkers related to colorectal cancer and smoking by using newer approaches to reduce dimensionality and reveal stable variables with a combination of techniques. These initial findings are suggestive of associations between colorectal cancer and lipophilic features representing potential environmental exposures, including cigarette smoking. Replication of these associations in follow-up studies involving comparable sample populations would strengthen arguments regarding potentially causal exposures.

In Chapter 1 the lipidomic method was first applied to pooled samples in our laboratory. A key purpose of this work was to refine the LCMS method for detecting a broad range of lipids, including the ULCFAS targeted in Chapter 2, these experiments revealed lipids associated with various covariates. While the pooling of samples likely limited the power to detect population characteristics, the lipids related to race and smoking status are intriguing. Many of the identified features, particularly the plasmalogens and DGLEA, seem unlikely to be false positives. Large effect changes and *p*-values that survived FDR correction combined with plausible biological mechanisms distinguished several of these features [43]. Finally, these data were integrated with adductomic data for the same samples, allowing a rare opportunity to combine –OMIC data which revealed additional information about previously observed associations.

This integrative –OMIC analysis in Chapter 4 focused on the relationship between HSA adducts and lipid features and several continuous covariates. As these analyses differed from binary covariate analysis from the original analyses of the same samples [144,147], some of the associations differed. Yet, several of the same adduct and lipid features that were associated with smokers and non-smokers were also directly or indirectly correlated with the number of cigarettes per day. Plausible biological pathways consistent with published epidemiologic or mechanistic data were observed. These results resulted in a more nuanced view of the previously detected associations, as well as a procedure to analyze data from these two types of -OMIC datasets.

As noted above, colorectal cancer is a leading cause of cancer death worldwide and a small portion of CRC risk is attributable to genetics alone. While some non-genetic CRC risk factors are well known, the bulk of CRC risk is unknown [105]. In Chapters 2 and 3 we sought out exposures, in the form of lipophilic molecules, detectable in serum of 190 nested case-control subjects from the EPIC prospective cohort. The resulting molecules were assessed as potentially contributing to CRC risk. Most intriguing, the prospective samples and the varying time periods from serum collection to case diagnosis offered an opportunity to differentiate between CRC biomarkers causally-related to the disease versus those that are result from disease progression.

The testing of ULCFAs, which were previously reported as likely protective of CRC, revealed that the depletion of these fatty acids in CRC cases was probably related to disease-related metabolic dysregulation. The ULCFAs, as well as nearly 9,000 quality-control-filtered features, were subjected to untargeted variable selection methods, followed by a regression model that revealed associations with case control status and time-to-diagnosis. Case-control variable selection methods were chosen based on a priority to select stable features most worthy of further analyses. These methods, including regularize regression (LASSO) and random forest, were less focused on traditional "*p*-value hunting" approaches that have proven troublesome, especially in –OMIC analyses. Evaluation of the nine selected features for associations with time to diagnosis (*ttd*) yielded interesting patterns. Four features appeared to be potentially causal features. Four other selected features, notably some of the previously mentioned ULCFAs and related fatty acids, appeared to be products of disease progression and, therefore, could be useful diagnostic biomarkers for early detection of CRC. These eight features are worthy of targeted analyses in an independent set of prospective CRC cases and controls.

The identities of these untargeted features proved more elusive and only partial annotation was achieved based on accurate masses and on-line databases. With the exception of the ULCFAs, these features differed from molecules that had previously been reported in prospective metabolomic studies of CRC cases and controls [109–111].

Instead of hypothesis-driven analyses of only known compounds, data-driven analyses of reliably detected OMIC features from untargeted analyses can generate hypotheses of possible disease-causing exposures. A requirement of these types of studies is the need to demonstrate reproducibility among validation sets from independent populations, regardless of *p*-value or effect size. Therefore, it is advisable that resource-intensive efforts toward annotation be preceded by replication of putative associations with independent samples. Upon replication, interesting features can be identified and targeted for follow-up to confirm causality and seek mechanistic understanding.

References:

1. Rainville PD, Stumpf CL, Shockcor JP, Plumb RS, Nicholson JK. Novel application of reversed-phase UPLC-oaTOF-MS for lipid analysis in complex biological mixtures: A new tool for lipidomics. J. Proteome Res. 2007;6:552–8.

2. Brown HA, Murphy RC. Working towards an exegesis for lipids in biology. Nat. Chem. Biol. Nature Publishing Group; 2009;5:602–6.

3. Ma X, Yang J. Lipidomics in Cancer Biomarker Discovery. Chapter B. Omi. Technol. Cancer Biomark. Discov. 2009;Cancer Gen.

4. Ritchie SA, Ahiahonu PWK, Jayasinghe D, Heath D, Liu J, Lu Y, et al. Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. BMC Med. 2010;8.

5. Ritchie SA, Tonita J, Alvi R, Lehotay D, Elshoni H, Su-Myat, et al. Low-serum GTA-446 anti-inflammatory fatty acid levels as a new risk factor for colon cancer. Int. J. Cancer. 2013;132:355–62.

6. Smith RE, Lespi P, Di Luca M, Bustos C, Marra FA, De Alaniz MJT, et al. A reliable biomarker derived from plasmalogens to evaluate malignancy and metastatic capacity of human cancers. Lipids. 2008;43:79–89.

7. Gross RW, Han X. Lipidomics in diabetes and the metabolic syndrome. Methods Enzymol. 2007;433:73–90.

8. Han X, Rozen S, Boyle SH, Hellegers C, Cheng H, Burke JR, et al. Metabolomics in early Alzheimer's disease: identification of altered plasma sphingolipidome using shotgun lipidomics. PLoS One. 2011;6:e21643.

9. Chan RB, Oliveira TG, Cortes EP, Honig LS, Duff KE, Small SA, et al. Comparative lipidomic analysis of mouse and human brain with Alzheimer disease. J. Biol. Chem. 2012;287:2678–88.

10. Graessler J, Schwudke D, Schwarz PEH, Herzog R, Shevchenko A, Bornstein SR. Top-down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients. PLoS One. 2009;4:e6261.

11. Cole LK, Dolinsky VW, Dyck JRB, Vance DE. Impaired phosphatidylcholine biosynthesis reduces atherosclerosis and prevents lipotoxic cardiac dysfunction in ApoE-/- Mice. Circ. Res. 2011;108:686–94.

12. Quehenberger O, Armando AM, Brown AH, Milne SB, Myers DS, Merrill AH, et al. Lipidomics reveals a remarkable diversity of lipids in human plasma. J. Lipid Res. 2010;51:3299–305.

13. Lin YS, McKelvey W, Waidyanatha S, Rappaport SM. Variability of albumin adducts of 1,4-benzoquinone, a toxic metabolite of benzene, in human volunteers. Biomarkers. 2006;11:14–27.

14. Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. Am. J. Epidemiol. 1992;135:1114–26.

15. Hu FB, Rimm E, Smith-Warner SA, Feskanich D, Stampfer MJ, Ascherio A, et al. Reproducibility and validity of dietary patterns assessed with a food frequency questionnaire. Am. J. Clin. Nutr. 1999;69:243–9.

16. Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am. J. Epidemiol. 1985;122:51–65.

17. Salvini S, Hunter DJ, Sampson L, Stampfer MJ, Colditz GA, Rosner B, et al. Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption. Int. J. Epidemiol. 1989;18:858–67.

18. Nomura DK, Long JZ, Niessen S, Hoover HS, Ng S-W, Cravatt BF. Monoacylglycerol lipase regulates a fatty acid network that promotes cancer pathogenesis. Cell. Elsevier Ltd; 2010;140:49–61.

19. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, et al. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal. Chem. 2003;75:4818–26.

20. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M. Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics. 2007;8:93.

21. Anderson MJ, Ter Braak CJF. Permutation tests for multi-factorial analysis of variance. J. Stat. Comput. Simul. 2003;73:85–113.

22. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. 1995;57:289–300.

23. Hanus L, Gopher A, Almog S, Mechoulam R. Two new unsaturated fatty acid ethanolamides in brain that bind to the cannabinoid receptor. J. Med. Chem. 1993;36:3032–4.

24. Joyce E, Ma L, Mitchell L, Felder C. Comparison the Human of the Pharmacology and Signal Transduction Cannabinoid CB1 and CB2 Receptors of. Mol. Pharmacol. 1995;48:443–50.

25. Balvers MGJ, Verhoeckx KCM, Bijlsma S, Rubingh CM, Meijerink J, Wortelboer HM, et al. Fish oil and inflammatory status alter the n-3 to n-6 balance of the endocannabinoid and oxylipin metabolomes in mouse plasma and tissues. Metabolomics. 2012;8:1130–47.

26. Mokdad AH, Ford ES, Bowman BA, Nelson DE, Engelgau MM, Vinicor F, et al. Diabetes trends in the US: 1990-1998. Diabetes Care. 2000;23:1278–83.

27. Hayes DK, Greenlund KJ, Denny CH, Croft JB, Keenan NL. Racial/ethnic and socioeconomic disparities in multiple risk factors for heart disease and stroke - United States, 2003 (Reprinted from MMWR, vol 54, pg 113-117, 2005). Jama-Journal Am. Med. Assoc. 2005;293:1441–3.

28. Rodriguez CJ, Diez-Roux A V, Moran A, Jin Z, Kronmal RA, Lima J, et al. Left Ventricular Mass and Ventricular Remodeling Among Hispanic Subgroups Compared With Non-Hispanic Blacks and Whites MESA (Multi-Ethnic Study of Atherosclerosis). J. Am. Coll. Cardiol. 2010;55:234–42.

29. Wallner S, Schmitz G. Plasmalogens the neglected regulatory and scavenging lipid species. Chem. Phys. Lipids. 2011;164:573–89.

30. Braverman NE, Moser AB. Functions of plasmalogen lipids in health and disease. Biochim. Biophys. Acta. Elsevier B.V.; 2012;1822:1442–52.

31. Colas R, Pruneta-Deloche V, Guichardant M, Luquain-Costaz C, Cugnet-Anceau C, Moret M, et al. Increased Lipid Peroxidation in LDL from Type-2 Diabetic Patients. Lipids. 2010;45:723–31.

32. Igarashi M, Ma K, Gao F, Kim H-W, Rapoport SI, Rao JS. Disturbed Choline Plasmalogen and Phospholipid Fatty Acid Concentrations in Alzheimer's Disease Prefrontal Cortex. J. Alzheimers Dis. 2011;24:507–17.

33. Ritchie SA, Akita H, Takemasa I, Eguchi H, Pastural E, Nagano H, et al. Metabolic system alterations in pancreatic cancer patient serum: potential for early detection. BMC Cancer. 2013;13.

34. Felde R, Spiteller G. Plasmalogen oxidation in human serum lipoproteins. Chem. Phys. Lipids. 1995;76:259–67.

35. Loidlstahlhofen A, Hannemann K, Felde R, Spiteller G. Epoxidation plasmalogens: source for long-chain alpha-hydroxyaldehydes in subcellular fractions of bovine liver. Biochem. J. 1995;309:807–12.

36. Hecht SS, Seow A, Wang M, Wang R, Meng L, Koh W-P, et al. Elevated levels of volatile organic carcinogen and toxicant biomarkers in Chinese women who regularly cook at home. Cancer Epidemiol. Biomarkers Prev. 2010;19:1185–92.

37. Wang G, Wang T. The Role of Plasmalogen in the Oxidative Stability of Neutral Lipids and Phospholipids. J. Agric. Food Chem. 2010;58:2554–61.

38. Nishimukai M, Wakisaka T, Hara H. Ingestion of plasmalogen markedly increased plasmalogen levels of blood plasma in rats. Lipids. 2003;38:1227–35.

39. Nomura DK, Morrison BE, Blankman JL, Long JZ, Kinsey SG, Marcondes MCG, et al. Endocannabinoid hydrolysis generates brain prostaglandins that promote neuroinflammation. Science. 2011;334:809–13.

40. Kim J, Li Y, Watkins BA. Fat to treat fat: emerging relationship between dietary PUFA, endocannabinoids, and obesity. Prostaglandins Other Lipid Mediat. Elsevier Inc.; 2013;104–105:32–41.

41. Steffen BT, Steffen LM, Tracy R, Siscovick D, Jacobs D, Liu K, et al. Ethnicity, plasma phospholipid fatty acid composition and inflammatory/endothelial activation biomarkers in the Multi-Ethnic Study of Atherosclerosis (MESA). Eur. J. Clin. Nutr. 2012;66:600–5.

42. Wang-Sattler R, Yu Y, Mittelstrass K, Lattka E, Altmaier E, Gieger C, et al. Metabolic profiling reveals distinct variations linked to nicotine consumption in humans--first results from the KORA study. PLoS One. 2008;3:e3863.

43. Sadiq ST, Agranoff D. Pooling serum samples may lead to loss of potential biomarkers in SELDI-ToF MS proteomic profiling. Proteome Sci. 2008;6:16.

44. Gonzalez-Covarrubias V, Dane A, Hankemeier T, Vreeken RJ. The influence of citrate, EDTA, and heparin anticoagulants to human plasma LC–MS lipidomic profiling. Metabolomics. 2012;9:337–48.

45. Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. CA. Cancer J. Clin. 2014;64:104–17.

46. Howlader N, Noone A, Krapcho M, Garshell J, Miller D, Altekruse S, et al. SEER Cancer Statistics Review, 1975-2011 [Internet]. Natl. Cancer Inst. 2014. p. based on November 2013 SEER data submission, poste. Available from: http://seer.cancer.gov/csr/1975_2011/

47. Hemminki K, Czene K. Attributable Risks of Familial Cancer from the Family-Cancer Database Attributable Risks of Familial Cancer from the Family-Cancer Database 1. 2002;1638–44.

48. Stone WL, Krishnan K, Campbell SE, Palau VE. The role of antioxidants and pro-oxidants in colon cancer. World J. Gastrointest. Oncol. 2014;6:55–66.

49. Rothwell PM, Fowkes FGR, Belch JFF, Ogawa H, Warlow CP, Meade TW. Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. Lancet. Elsevier Ltd; 2010;377:31–41.

50. Awtry E, Loscalzo J. Aspirin. Circulation. 2000;101:1206–18.

51. Chiang N, Arita M, Serhan CN. Anti-inflammatory circuitry: lipoxin, aspirin-triggered lipoxins and their receptor ALX. Prostaglandins. Leukot. Essent. Fatty Acids. 2005;73:163–77.

52. Terzić J, Grivennikov S, Karin E, Karin M. Inflammation and Colon Cancer. Gastroenterology. 2010;138:2101–2114.e5.

53. Masoodi M, Mir AA, Petasis NA, Serhan CN. Europe PMC Funders Group Simultaneous lipidomic analysis of three families of bioactive lipid mediators leukotrienes , resolvins , protectins and related hydroxy-fatty acids by liquid chromatography / electrospray tandem mass

spectrometry. 2008;22:75–83.

54. Ritchie SA, Heath D, Yamazaki Y, Grimmalt B, Kavianpour A, Krenitsky K, et al. Reduction of novel circulating long-chain fatty acids in colorectal cancer patients is independent of tumor burden and correlates with age. BMC Gastroenterol. BioMed Central Ltd; 2010;10.

55. Ritchie SA, Jayasinghe D, Davies GF, Ahiahonu P, Ma H, Goodenowe DB. Human serum-derived hydroxy long-chain fatty acids exhibit anti-inflammatory and anti-proliferative activity. J. Exp. Clin. Cancer Res. BioMed Central Ltd; 2011;30:59.

56. Babbs CF. Free Radicals and the Etiology of Colon Cancer. Free Radic. Biol. Med. 1990;8:191–200.

57. Rothwell PM, Wilson M, Elwin CE, Norrving B, Algra A, Warlow CP, et al. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. Lancet. Elsevier Ltd; 2010;376:1741–50.

58. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002;5:1113–24.

59. Williams MD, Reeves R, Resar LS, Hill HH. Metabolomics of colorectal cancer: Past and current analytical platforms. Anal. Bioanal. Chem. 2013;405:5013–30.

60. Nolen BM, Brand RE, Prosser D, Velikokhatnaya L, Allen PJ, Zeh HJ, et al. Prediagnostic serum biomarkers as early detection tools for pancreatic cancer in a large prospective cohort study. PLoS One. 2014;9:e94928.

61. Moore LL, Bradlee ML, Singer MR, Splansky GL, Proctor MH, Ellison RC, et al. BMI and waist circumference as predictors of lifetime colon cancer risk in Framingham Study adults. Int. J. Obes. Relat. Metab. Disord. 2004;28:559–67.

62. Aleksandrova K, Boeing H, Jenab M, Bueno-de-Mesquita HB, Jansen E, Van Duijnhoven FJB, et al. Metabolic syndrome and risks of colon and rectal cancer: The european prospective investigation into cancer and nutrition study. Cancer Prev. Res. 2011;4:1873–83.

63. Larsson SC, Kumlin M, Ingelman-sundberg M, Wolk A. Dietary long-chain n ⧖ 3 fatty acids for the prevention of cancer : a review of potential mechanisms 1 – 3. 2004;

64. Song M, Chan AT, Fuchs CS, Ogino S, Hu FB, Mozaffarian D, et al. Dietary intake of fish, ω-3 and ω-6 fatty acids and risk of colorectal cancer: A prospective study in U.S. men and women. Int. J. cancer J. Int. du cancer. 2014;

65. Talwar P. Manual of Assisted Reproductive Technologies and Clinical Embryology: Jaypee Brothers Medical Publisher Pvt. Limited; 2014.

66. Saint-Ramon J-G, Beau C, Ehrsam A. Tubes for conservation biological particle; for use as tool in biological sampling. Google Patents; 2001.

67. Fages A, Ferrari P, Monni S, Dossus L, Floegel A, Mode N, et al. Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. Metabolomics. 2014;10:1074–83.

68. Bligh EG, Dyer WJ. A rapid method of total lipid extraction and purification. Can. J. Biochem. Physiol. 1959;37:911–7.

69. Schlechtriem C, Focken U, Becker K. Effect of different lipid extraction methods on delta13C of lipid and lipid-free fractions of fish and different fish feeds. Isotopes Environ. Health Stud. 2003;39:135–40.

70. Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan and GS. LC / MS Preprocessing and Analysis with xcms. Anal. Chem. 2006;78:779–787.

71. Benton HP, Wong DM, Trauger S a., Siuzdak G. XCMS 2 : Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization. Anal. Chem. 2008;80:6382–9.

72. Patti GJ, Tautenhahn R, Siuzdak G. Meta-analysis of untargeted metabolomic data from multiple profiling experiments. Nat. Protoc. Nature Publishing Group; 2012;7:508–16.

73. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. R Found. Stat. Comput. Vienna, Austria. 2013.

74. Aitchison J, Brown JA. The Lognormal Distribution. Lognormal Distrib. London, England: Cambridge University Press; 1957.

75. Ritchie S, Goodenowe D, Khan MA, Ahiahonu PWK. Hydroxy fatty acid compounds and uses thereof for disease treatment and diagnosis. Google Patents; 2012.

76. Serhan CN. Novel eicosanoid and docosanoid mediators: resolvins, docosatrienes, and neuroprotectins. Curr. Opin. Clin. Nutr. Metab. Care. 2005;8:115–21.

77. Schwab JM, Serhan CN. Lipoxins and new lipid mediators in the resolution of inflammation. Curr. Opin. Pharmacol. 2006;6:414–20.

78. Serhan CN, Chiang N, Van Dyke TE. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. Nat. Rev. Immunol. 2008;8:349–61.

79. Serhan CN. Pro-resolving lipid mediators are leads for resolution physiology. Nature. 2014;510:92–101.

80. Aveldaño MI, Robinson BS, Johnson DW, Poulos A. Long and very long chain polyunsaturated fatty acids of the n-6 series in rat seminiferous tubules. J. Biol. Chem. 1993;268:11663–9.

81. Robinson BS, Johnson DW, Poulos A. Novel molecular species of sphingomyelin containing 2-hydroxylated polyenoic very-long-chain fatty acids in mammalian testes and spermatozoa. J.

Biol. Chem. 1992;267:1746–51.

82. Poulos A. Very long chain fatty acids in higher animals--a review. Lipids. 1995;30:1–14.

83. Agbaga M-P, Mandal MN a, Anderson RE. Retinal very long-chain PUFAs: new insights from studies on ELOVL4 protein. J. Lipid Res. 2010;51:1624–42.

84. Leonard AE, Pereira SL, Sprecher H, Huang YS. Elongation of long-chain fatty acids. Prog. Lipid Res. 2004;43:36–54.

85. Jakobsson A, Westerberg R, Jacobsson A. Fatty acid elongases in mammals: Their regulation and roles in metabolism. Prog. Lipid Res. 2006;45:237–49.

86. Vickers AJ, Ulmert D, Serio AM, Björk T, Scardino PT, Eastham J a, et al. The predictive value of prostate cancer biomarkers depends on age and time to diagnosis: towards a biologically-based screening strategy. Int. J. Cancer. 2007;121:2212–7.

87. Erlinger TP, Platz E a, Rifai N, Helzlsouer KJ. C-reactive protein and the risk of incident colorectal cancer. JAMA. 2004;291:585–90.

88. Dorgan JF, Longcope C, Stephenson HE, Falk RT, Miller R, Franz C, et al. Relation of prediagnostic serum estrogen and androgen levels to breast cancer risk. Cancer Epidemiol. Biomarkers Prev. 1996;5:533–9.

89. McSorley M a, Alberg AJ, Allen DS, Allen NE, Brinton L a, Dorgan JF, et al. C-reactive protein concentrations and subsequent ovarian cancer risk. Obstet. Gynecol. 2007;109:933–41.

90. Cust AE, Kaaks R, Friedenreich C, Bonnet F, Laville M, Lukanova A, et al. Plasma adiponectin levels and endometrial cancer risk in pre- and postmenopausal women. J. Clin. Endocrinol. Metab. 2007;92:255–63.

91. Ritchie SA, Chitou B, Zheng Q, Jayasinghe D, Jin W, Mochizuki A, et al. Pancreatic cancer serum biomarker PC-594 : Diagnostic performance and comparison to CA19-9. 2015;21:6604–12.

92. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2016;0:1–9.

93. Haggar FA, Boushey RP. Colorectal Cancer Epidemiology : Incidence, Mortality, Survival, and Risk Factors. Clin. Colon Rectal Surg. 2009;22:191–7.

94. American Cancer Society. Cancer Facts & Figures 2015. Cancer Facts Fig. 2015. 2015;1–9.

95. Rappaport SM. Genetic Factors Are Not the Major Causes of Chronic Diseases. PLoS One. 2016;11:e0154387.

96. Gassler N, Klaus C, Kaemmerer E, Reinartz A, Gassler N, Klaus C, et al. Modifier-concept of colorectal carcinogenesis : Lipidomics as a technical tool in pathway analysis. 2010;16:1820–7.

97. Leufkens AM, Van Duijnhoven FJB, Siersema PD, Boshuizen HC, Vrieling A, Agudo A, et al. Cigarette smoking and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition study. Clin. Gastroenterol. Hepatol. 2011;9:137–44.

98. Platz EA, Willett WC, Colditz GA, Rimm EB, Spiegelman D, Giovannucci E. Proportion of colon cancer risk that might be preventable in a cohort of middle-aged US men. Cancer Causes Control. 2000;11:579–88.

99. Oostindjer M, Alexander J, Amdam G V, Andersen G, Bryan NS, Chen D, et al. The role of red and processed meat in colorectal cancer development : a perspective. MESC. Elsevier B.V.; 2014;97:583–96.

100. Ma Y, Zhang P, Wang F, Yang J, Liu Z, Qin H. Association Between Vitamin D and Risk of Colorectal Cancer : A Systematic Review of Prospective Studies. J. Clin. Oncol. 2017;29.

101. Keefe S. Diet, microorganisms and their metabolites, and colon cancer. Nat. Publ. Gr. Nature Publishing Group; 2016;13:691–706.

102. Vipperla K, O'Keefe S. Diet, microbiota, and dysbiosis: a "recipe" for colorectal cancer. Food Funct. Royal Society of Chemistry; 2016;1731–40.

103. Norat T, Bingham S, Ferrari P, Slimani N, Jenab M, Mazuir M, et al. Meat, fish, and colorectal cancer risk : the European Prospective Investigation into cancer and nutrition. J. Natl. Cancer Inst. 2005;97:906–16.

104. Chao A, Connell CJ, Mccullough ML, Jacobs EJ, Flanders WD, Rodriguez C, et al. Meat Consumption and Risk of Colorectal Cancer. J. Am. Med. Assoc. 2005;293.

105. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The blood exposome and its role in discovering causes of disease. Environ. Health Perspect. 2014;122:769–74.

106. Van Duijnhoven JB, Bueno-de-mesquita HB, Calligaro M, Jansen HJM, Frohlich J, Ayyobi A, et al. Blood lipid and lipoprotein concentrations and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition. Gut. 2011;1–10.

107. Perttula K, Edmands WMB, Grigoryan H, Cai X, Iavarone AT, Gunter MJ, et al. Evaluating Ultra-long-Chain Fatty Acids as Biomarkers of Colorectal Cancer Risk. Cancer Epidemiol. Biomarkers Prev. 2016;25:1216–24.

108. Nishiumi S, Kobayashi T, Ikeda A, Yoshie T, Kibi M, Izumi Y, et al. A novel serum metabolomics-based diagnostic approach for colorectal cancer. PLoS One. 2012;7:1–10.

109. Cross AJ, Moore SC, Boca S, Huang W-Y, Xiong X, Stolzenberg-Solomon R, et al. A prospective study of serum metabolites and colorectal cancer risk. Cancer. 2014;1–9.

110. Bae S, Ulrich CM, Neuhouser ML, Malysheva O, Bailey LB, Xiao L, et al. Plasma choline metabolites and colorectal cancer risk in the women's health initiative observational study. Cancer Res. 2014;74:7442–52.

111. Nitter M, Norgård B, Vogel S De, Eussen SJPM, Meyer K, Ulvik A, et al. Plasma methionine, choline, betaine, and dimethylglycine in relation to colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann. Oncol. 2014;1609–15.

112. Xu R, Wang Q, Li L. A genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine N-oxide (TMAO), a gut microbial metabolite of dietary meat and fat. BMC Genomics. 2015;16:1–9.

113. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. Nature Publishing Group; 2011;472:57–63.

114. Zackular JP, Rogers MAM, Ruf MT, Schloss PD. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. Cancer Prev. Res. 2014;7:1112–22.

115. Santos CR, Schulze A. Lipid metabolism in cancer. FEBS J. 2012;279:2610–23.

116. Lydic TA, Townsend S, Adda CG, Collins C, Mathivanan S, Reid GE. Rapid and comprehensive “shotgun” lipidome profiling of colorectal cancer cell derived exosomes. Methods. Elsevier Inc.; 2015;87:83–95.

117. Friedman J, Hastie T, Tibshirani RJ. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 2010;33.

118. Liaw A, Wiener M. Classification and Regression by randomForest. R news. 2002;2:18–22.

119. Huang BFF, Boutros PC. The parameter sensitivity of random forests. BMC Bioinformatics. BMC Bioinformatics; 2016;17:1–13.

120. Saeys Y, Inza I, Larranaga P. Gene expression A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507–17.

121. Kuhl C, Tautenhahn R, Christoph B, Larson TR, Neumann S. CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. Anal. Chem. 2011;84:283–9.

122. Stanstrup J, Gerlich M, Dragsted LO, Neumann S. Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Anal. Bioanal. Chem. 2013;405:5037–48.

123. Edmands WMB, Petrick LM, Barupal DK, Scalbert A, Wilson M, Wickliffe J, et al. compMS2Miner : an automatable metabolite identification, visualization and data-sharing R package for high-resolution LC-MS datasets. Anal. Chem. 2017;

124. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: The human metabolome database. Nucleic Acids Res. 2007;35.

125. Smith CA, O’Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN A

Metabolite Mass Spectral Database. Proc. 9Th Int. Congr. Ther. Drug Monit. Clin. Toxicol. 2005;27:747–51.

126. Veselkov KA, Vingara LK, Masson P, Robinette SL, Want E, Li J V., et al. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. Anal. Chem. 2011;83:5864–72.

127. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat. Protoc. 2011;6:1060–83.

128. Bolstad BM, Irizarry RA. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. 2003;19:185–93.

129. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

130. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

131. Broeckling CD, Afsar FA, Neumann S, Prenni JE. RAMClust: A Novel Feature Clustering Method Enables Spectral- Matching-Based Annotation for Metabolomics Data. 2014;

132. Tibshirani R. Regression Selection and Shrinkage via the Lasso. J. R. Stat. Soc. B. 1996. p. 267–88.

133. Bach FR, Project-team IW. Bolasso : Model Consistent Lasso Estimation through the Bootstrap. 2008;

134. Breiman L. Random forests. Mach. Learn. 2001;45:5–32.

135. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006;2:171–96.

136. Byrdwell WC. Dual parallel liquid chromatography with dual mass spectrometry (LC2/MS2) for a total lipid analysi. Front. Biosci. 2008;13:100–20.

137. Rezanka T, Sigler K. Odd-numbered very-long-chain fatty acids from the microbial, animal and plant kingdoms. Prog. Lipid Res. 2009;48:206–38.

138. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. PLoS One. 2013;8.

139. Agle SC, Philips P, Martin RCG. Environmental Exposures, Tumor Heterogeneity, and Colorectal Cancer Outcomes. Curr. Colorectal Cancer Rep. 2014;10:189–94.

140. Ge H, Walhout AJM, Vidal M. Integrating “omic” information: A bridge between genomics and systems biology. Trends Genet. 2003;19:551–60.

141. Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T, et al. Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. PLoS Genet. 2008;4.

142. Rueedi R, Ledda M, Nicholls AW, Salek RM, Marques-Vidal P, Morya E, et al. Genome-Wide Association Study of Metabolic Traits Reveals Novel Gene-Metabolite-Disease Links. PLoS Genet. 2014;10.

143. Hartiala J, Bennett BJ, Tang WHW, Wang Z, Stewart AFR, Roberts R, et al. Comparative genome-wide association studies in mice and humans for trimethylamine N-Oxide, a proatherogenic metabolite of choline and L-carnitine. Arterioscler. Thromb. Vasc. Biol. 2014;34:1307–13.

144. Grigoryan H, Edmands W, Lu SS, Yano Y, Regazzoni L, Iavarone AT, et al. Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin. Anal. Chem. 2016;88:10504–12.

145. Carlsson H, Stedingk H Von, Nilsson U, To M. LC – MS/MS Screening Strategy for Unknown Adducts to N ‑ Terminal Valine in Hemoglobin Applied to Smokers and Nonsmokers. Chem. Res. Toxicol. 2014;27:2062–70.

146. Balbo S, Turesky RJ, Villalta PW. DNA Adductomics. Chem. Res. Toxicol. 2014;27:356–66.

147. Cai X, Perttula K, Pajouh S, Hubbard A, Nomura DK, Rappaport SM. Untargeted Lipidomic Profiling of Human Plasma Reveals Differences due to Race, Gender, and Smoking Status. Metabolomics Open Access. 2014;4:1–8.

148. Salzano AM, Renzone G, Scaloni A, Torreggiani A, Ferreri C, Chatgilialoglu C. Human serum albumin modifications associated with reductive radical stress. Mol. Biosyst. 2010;7:889–98.

149. Szkudlarek A, Chudzik M. Spectrochimica Acta Part A : Molecular and Biomolecular Spectroscopy Alteration of human serum albumin binding properties induced by modi fi cations : A review. Spectrochim. Acta. 2018;188:675–83.

150. Rappaport SM, Li H, Grigoryan H, Funk WE, Williams ER. Adductomics: Characterizing exposures to reactive electrophiles. Toxicol. Lett. 2012;213:83–90.

151. Nomura DK, Long JZ, Niessen S, Hoover HS, Ng SW, Cravatt BF. Monoacylglycerol Lipase Regulates a Fatty Acid Network that Promotes Cancer Pathogenesis. Cell. Elsevier Ltd; 2010;140:49–61.

152. Bach F. Bolasso: model consistent Lasso estimation through the bootstrap. Proc. 25th Int. Conf. Mach. Learn. 2008;33–40.

153. Ottestad I, Hassani S, Borge GI, Kohler A, Vogt G, Hyötyläinen T, et al. Fish oil supplementation alters the plasma lipidomic profile and increases long-chain PUFAs of phospholipids and triglycerides in healthy subjects. PLoS One. 2012;7:e42550.

154. Hyötyläinen T, Bondia-Pons I, Orešič M. Lipidomics in nutrition and food research. Mol. Nutr. Food Res. 2013;57:1306–18.

155. Grigoryan H, Li H, Iavarone AT, Williams ER, Rappaport SM. Cys34 adducts of reactive oxygen species in human serum albumin. Chem. Res. Toxicol. 2012;25:1633–42.

156. Valavanidis A, Vlachogianni T, Fiotakis K. Tobacco Smoke : Involvement of Reactive Oxygen Species and Stable Free Radicals in Mechanisms of Oxidative Damage , Carcinogenesis and Synergistic Effects with Other Respirable Particles. Int. J. Environ. Res. Public Health. 2009;6:445–62.

157. Ambrose JA, Barua RS. The Pathophysiology of Cigarette Smoking and Cardiovascular Disease. J. Am. Coll. Cardiol. Elsevier Masson SAS; 2004;43.

158. Chung S, Vu S, Filosto S, Goldkorn T. Src regulates cigarette smoke-induced ceramide generation via neutral sphingomyelinase 2 in the airway epithelium. Am. J. Respir. Cell Mol. Biol. 2015;52:738–48.

159. Filosto S, Castillo S, Danielson A, Franzi L, Khan E, Kenyon N, et al. Neutral sphingomyelinase 2: A novel target in cigarette smoke-induced apoptosis and lung injury. Am. J. Respir. Cell Mol. Biol. 2011;44:350–60.

160. Levy M, Khan E, Careaga M, Goldkorn T. Neutral sphingomyelinase 2 is activated by cigarette smoke to augment ceramide-induced apoptosis in lung cell death. Am. J. Physiol. Lung Cell. Mol. Physiol. 2009;297:L125–33.

161. Clement AB, Gamerdinger M, Tamboli IY, Lütjohann D, Walter J, Greeve I, et al. Adaptation of neuronal cells to chronic oxidative stress is associated with altered cholesterol and sphingolipid homeostasis and lysosomal function. J. Neurochem. 2009;111:669–82.

162. Tepper AD, Ruurs P, Wiedmer T, Sims PJ, Borst J, Van Blitterswijk WJ. Sphingomyelin hydrolysis to ceramide during the execution phase of apoptosis results from phospholipid scrambling and alters cell-surface morphology. J. Cell Biol. 2000;150:155–64.

163. Sagone A, Lawrence T, Balcerzak S. Effect of Smoking on Tissue Oxygen Supply. Blood. 1973;41.

164. Go YM, Jones DP. The redox proteome. J. Biol. Chem. 2013;288:26512–20.

165. Saito S, Tomonobu K, Hase T, Tokimitsu I. Effects of diacylglycerol on postprandial energy expenditure and respiratory quotient in healthy subjects. Nutrition. 2006;22:30–5.

166. Yasunaga T, Yasukawa K. Nutritional functions of dietary diacylglycerols. J. Oleo Sci. 2001;50.

167. Ha CY, Kim JY, Paik JK, Kim OY, Paik Y-H, Lee EJ, et al. The association of specific metabolites of lipid metabolism with markers of oxidative stress, inflammation and arterial stiffness in men with newly diagnosed type 2 diabetes. Clin. Endocrinol. (Oxf). 2012;76:674–82.

168. Dong J, Cai X, Zhao L, Xue X, Zou L, Zhang X, et al. Lysophosphatidylcholine profiling of plasma: Discrimination of isomers and discovery of lung cancer biomarkers. Metabolomics. 2010;6:478–88.

169. Okita M, Gaudette DC, Mills GB, Holub BJ. Elevated levels and altered fatty acid composition of plasma lysophosphatidylcholine(lysoPC) in ovarian cancer patients. Int J Cancer. 1997;71:31–4.

170. Hyötyläinen T, Orešič M. Optimizing the lipidomics workflow for clinical studies-practical considerations. Anal. Bioanal. Chem. 2015;