VNIVERSITAT DE VALÈNCIA

Faculty of Mathematics

Department of Statistics and Operations Research

The role of blood DNA methylation in environment-related chronic disease: a biostatistical toolkit

Arce Domingo Relloso

PhD Thesis in Statistics and Optimization

Supervised by

José D. Bermúdez Edo

Maria Tellez-Plaza

February 2023

This thesis has been conducted in the National Center for Epidemiology (Carlos III Health Institutes, Madrid, Spain) and has been supported by a fellowship from “la Caixa” Foundation (ID 100010434) (fellowship code ”LCF/BQ/DR19/11740016”). Part of the research has been carried out during visits to Dr. Ana Navas Acien’s lab at Columbia University Mailman School of Public Health, NY, USA, and Dr. Belinda Phipson’s lab at the Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia.

Preface

The present dissertation is the result of the research performed in the last three years under the leadership of Arce Domingo Relloso, in collaboration with the Integrative Epidemiology Group, Department of Chronic Diseases Epidemiology from the National Center for Epidemiology - Carlos III Health Institute (Madrid, Spain), the Department of Statistics and Operations Research, University of Valencia (Spain), the Department of Environmental Health Sciences, Columbia University Mailman School of Public Health (New York, NY), and the Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research (Melbourne, Australia).

The data applications of this thesis use both simulated data and population-based data from the Strong Heart Study, a prospective cohort of American Indians from Arizona, Oklahoma, North Dakota and South Dakota. All research involving the Strong Heart Study is community based, thus, topics are proposed to the tribal communities and resulting science is directly communicated to them. All research conducted in this thesis involving the Strong Heart Study has been approved by the Strong Heart Study's steering committee.

This thesis has been conducted with the aim of contributing to the development of statistical methods for omics data research, which is on the spotlight of the scientific community due to its potential to contribute to the developement of treatments and early detection tools for disease. In a world in which multidisciplinary research is becoming essential to approach research problems, this thesis is an example of how several knowledge areas need to be put together in order to have a broader perspective in biomedical research. This is a multidisciplinary PhD which falls within the areas of biostatistics and epidemiology, while also including some bioinformatics applications. Part of the thesis is largely focused on the development of statistical methods. However, epidemiologic studies that contribute to the body of evidence in the field of DNA methylation, environmental factors and chronic disease are likewise an important part of this thesis.

The structure of this thesis is as follows. In chapter 1, we present the motivation and objectives and we provide a general introduction to DNA methylation and its association with environmental factors and chronic disease, as well as to transcriptomics, the direct biological consequence of DNA methylation. Chapter 2 describes the currently established and mainly used statistical methods for DNA methylation association analysis, including both one-marker-at-a-time and multiple-markers-at-a-time approaches, as well as Bayesian penalized methods and simple mediation analysis. Chapter 3 deepens into the problem of variable selection in the omics data setting. Section 3.1 describes the extension of the ISIS tool developed in this thesis. Section 3.2 includes two different applications of this statistical tool; the first application is a comparison between different penalization methods paired with ISIS, and the second is the application of ISIS to an epidemiologic problem: the association of arsenic exposure with DNA methylation and cardiovascular disease. Chapter 4 focuses on multiple uncausally correlated mediators in survival settings. Section 4.1 provides an introduction to multiple mediation analysis, whereas section 4.2 focuses on multiple mediation analysis with survival outcomes. Section 4.3 describes our contribution to the multimediate algorithm, which conducts multiple mediation analysis for uncausally correlated mediators in the context of survival analysis. Section 4.4 includes two applications of the multimediate algorithm: the first application is a simulation study to illustrate its utility, and the second is an application of the algorithm to an epidemiologic problem: the association of smoking with DNA methylation and smoking-related cancers. Chapter 5 depicts a future work line, in which we aim to extend our research to gene expression data. This chapter includes an evaluation of statistical methods to assess differences in variability in transcriptomics of single cells. Chapter 6 includes conclusions, limitations and final remarks. Last, chapter 7 summarizes the scientific production conducted during this doctoral thesis.

Acknowledgements

En palabras de la escritora y activista política Helen Keller, "la vida es, o bien una osada aventura, o bien nada en absoluto". Aquí llega, pues, el final de la mayor aventura en la que me he embarcado hasta la fecha. A caballo entre la pena de dejar atrás esta bonita etapa y la emoción de lo que está por venir, me gustaría dedicar estas líneas a aquellos que han hecho posible el desarrollo de esta tesis doctoral.

En primer lugar, quisiera agradecer a mis directores de tesis todo su esfuerzo y paciencia. A la **Doctora María Téllez Plaza**, la persona más influyente de mi vida profesional. Gracias por impulsarme hacia sendas que jamás creí poder transitar, a las que nunca hubiese llegado si no fuese por tu apoyo y perseverancia. También por ayudarme a relativizar, por tus sabios consejos y por saber levantar el ánimo en los momentos más complicados. Al **Doctor José Bermúdez**, quien a pesar de su gran carga de trabajo, se prestó a embarcarse conmigo en esta aventura. Gracias por depositar en este proyecto toda tu sabiduría y experiencia, y por enseñarme a buscar vías alternativas cuando un camino se cierra.

A la **Doctora Ana Navas Acién**, que podría considerarse pseudo-directora de esta tesis, y una de las personas a las que más admiro tanto en lo profesional como en lo personal. Gracias por apostar por mi siempre, por transmitirme tu gran pasión por la ciencia y por tu cercanía con todos los que te rodean.

También me gustaría dar las gracias a la **Fundación la Caixa** por creer en este proyecto y financiarlo, así como por la enorme motivación recibida en los trainings anuales. De igual manera, al **Centro Nacional de Epidemiología** y su gente, y a la **Universidad de Valencia**, donde me adentré por primera vez en el mundo de la bioestadística.

Likewise, I would like to thank **Doctor Belinda Phipson**, who welcomed me in Melbourne so warmly and opened new research horizons for me. I would also like to thank my colleagues at Columbia University (**Tiffany, Marisa, Ahlam, Annie, Fernando** and many oth-

ers) and Walter and Eliza Hall Institute of Medical Research (**Givanna, Melody, Nishika, Andy** and all Bioinformatics colleagues), from whom I have learned many valuable things and made these two visiting periods a huge scientific and social immersion.

I also owe an important part of this thesis to the **Strong Heart Study** investigators, staff and participants. I am so grateful to this project for having taught me the real importance of community-based research. I would like to thank my colleagues from Texas Biomedical Research Institute (**Shelley Cole, Karin Haack, Deborah Newman and Sandra Smith**) for their important contributions to this project and for letting me use their high performance computing cluster. I would also like to thank investigators from the **Framingham Heart Study**, the **Women's Health Initiative** and the **Multi-Ethnic Study of Atherosclerosis** for allowing replication of our results in independent cohorts, which is essential to validate science, and all our coauthors for providing such relevant feedback and recommendations.

In addition, I would like to give special thanks to **Doctors Yang Feng**, from New York University, and **Allan Jerolon**, from Paris Descartes University, for setting the methodological bases of our work and supporting our continuation of it.

No podría olvidarme de mis compañeras de doctorado, **María Grau, Zulema Rodríguez y Marta Gálvez**, ni de mis compañeros de **La Caixa (de papel)**, todos ellos fieles compañeros de viaje que han sido un gran apoyo durante estos años, que han compartido mis éxitos, fracasos e inquietudes, y han sido fundamentales para mantener siempre la motivación y la positividad ante este reto. Igualmente, a mis compañeros de **INCLIVA**, donde empezó mi aventura en la ciencia. En especial, a **Miguel Herreros**, que no solo nos enseñó cómo usar un servidor de alta computación, sino que lo ha mantenido con vida todos estos años. Sin él, este trabajo no hubiese sido posible.

A mi familia, y en especial a **mis padres**, que me inculcaron como pilar esencial el valor del esfuerzo y el trabajo. Gracias por vuestro incondicional apoyo, por creer en mi en todo momento, suscribir todas

mis decisiones y ser mi mayor ejemplo a seguir.

También a mis amigos, la familia que se elige, aquella que sostiene los cimientos de mi vida. En especial, a **Cristina y María José**, mis hermanas del alma. Gracias por ser el refugio de mi tormenta y, sobre todo, el arcoíris posterior. A mis amigos de Columbia, **Irene y Gonzalo**, que han vivido conmigo gran parte de este proceso, siendo siempre fieles consejeros y testigos de todas mis alegrías y penas profesionales, pero también personales. También a **Teresa, Bea y Elena**, por ser hogar durante todos estos años, y escucharme siempre cuando algo no salía bien. Igualmente, a toda esa gente que me alegra los días y el corazón y que por lo tanto, ha sido motor fundamental para mi durante este viaje. A mis **amigas de Vilviestre**, a mi **Crew de Madrid** y a mi **familia de Nueva York**. En todos ellos he encontrado siempre una mano tendida, y la calidez del hogar.

The role of blood DNA methylation in environment-related chronic disease: a biostatistical toolkit

UNIVERSITÀT DE VALENCIA

Faculty of Mathematics

Doctoral Program in Statistics and Optimization

General abstract

Epigenetic changes refer to modifications that alter gene expression without changing the genomic sequence. Environmental and behavioral factors are well-known epigenetic modifiers, leading to heritable changes that might disrupt essential biological processes and, in turn, influence the development of disease.

DNA methylation is the most widely studied epigenetic mark. Scientific evidence supports the association between environmental factors, such as smoking and metals, and DNA methylation dysregulations. In addition, the evidence supports the association between DNA methylation dysregulations and chronic disease, especially for cancer. However, it is unknown whether these associations are causal or happen due to DNA methylation being a biomarker of other disrupted biological processes.

In order to evaluate the potential role of genome-wide DNA methylation on the association between environmental factors and chronic disease, appropriate statistical methods for the analysis of ultra-high dimensional and highly correlated data are needed.

To begin with, we need to select which methylation sites in the genome are associated with our outcome of interest. Existing methods for variable selection and effect estimation lose predictive ability and are subject to bias in ultra-high dimensional settings. Additionally, they are not able to quantify statistical uncertainty.

Once we get to select the set of epigenomic features associated with our outcome, mediation analysis is a valuable tool to quantify the potential intermediate effect of these methylation sites on the association between environmental factors and chronic disease. The most biologically plausible scenario is that several correlated DNA methylation marks (as opposed to a single one) are mediators between an exposure and an outcome. On the other hand, it is common to consider time-to-event outcomes in epidemiological settings, in order to incorporate the time in which the outcome happened into the statistical model. However, to date, no mediation analysis algorithms able to deal with multiple correlated mediators with survival outcomes have been developed.

Thus, this thesis has two main objectives, the first one related to variable selection in ultra-high dimensional settings, and the second one focused on multiple mediation analysis with survival outcomes.

Abstract of objective 1. The first objective of this thesis arises from the need to extend the Iterative Sure Independence Screening (ISIS) statistical tool, which conducts variable selection for ultra-high dimensional data, in order to improve its predictive accuracy, effect estimation and to incorporate statistical uncertainty. The objective was to pair the ISIS algorithm with two shrinkage methods: elastic-net and adaptive elastic-net (Aenet), and to include an algorithm for calculation of bootstrap-based confidence intervals. This extension of ISIS has been added to the *SIS* R package, which is available in the CRAN repository.

As part of this first objective, this dissertation shows two applications of the ISIS algorithm. For this purpose, we used data from the Strong Heart Study (SHS), the largest and longest prospective cohort of American Indians. The first application aimed to evaluate the improvements introduced by our extension of ISIS (Aenet, elastic-net, MSAenet) as compared to other shrinkage methods implemented in the original version. The ISIS algorithm paired with Aenet provides increased predictive ability as compared to the original ISIS version, especially for continuous and binary outcomes. Additionally, by pair-

ing ISIS with Aenet, a more consistent effect estimation is obtained because Aenet fulfills the oracle property. Our bioinformatics analysis reveals that it also leads to a more robust variable selection in terms of subsequent biological pathway enrichment.

The second application is an epidemiologic study in which we evaluate the potential intermediate role of single DNA methylation sites on the well-documented association between arsenic and cardiovascular disease (CVD). We used the ISIS algorithm paired with Aenet to select methylation sites associated with CVD, and we subsequently conducted a simple mediation analysis (one marker at a time) in the selected sites. We found statistically significant mediated effects for 21 and 15 differentially methylated positions (DMPs) for CVD incidence and mortality, respectively. In addition, six of the 21 DMPs showing statistically significant mediated effects for CVD incidence were replicated in three independent American cohorts (the Framingham Heart Study, Women's Health Initiative y Multi-Ethnic Study of Atherosclerosis) with the same direction in the association. The genes annotated to methylation sites with statistically significant mediated effects were also replicated in a mouse model. The biological plausibility of those genes in CVD provides additional robustness of the results.

Abstract of objective 2. The second objective of this thesis focuses on the extension of the multimediate algorithm, which conducts mediation analysis in the context of multiple correlated mediators, to survival outcomes. Jerolon and colleagues developed this algorithm for continuous and binary outcomes. Using the Lin-Ying additive models, we extended the multimediate algorithm as well as the theoretical results for identification of mediated effects to time-to-event data. In addition, we adapted the multimediate algorithm to incorporate potential exposure-mediator interactions. The extension of the algorithm to survival outcomes is available in the following Github repository: *https://github.com/AllanJe/multimediate*. The extension including exposure-mediator interactions will soon be posted in the same repository.

As part of this second objective, we also included two data appli-

cations of this algorithm. The first application is a simulation study in which we prove the better performance of the multimediate algorithm as compared to simple mediation analysis, even in settings of uncorrelated mediators.

The second data application is an epidemiologic study in which we investigate the potential intermediate role of multiple, potentially correlated, DNA methylation marks on the association between smoking and smoking-related cancers using data from the SHS. We first used the ISIS algorithm paired with elastic-net to select DNA methylation sites associated with cancer. Subsequently, we applied the multimediate algorithm to evaluate several methylation sites as potential mediators on the association between smoking and cancer. The algorithm identified a joint mediated effect of 81.3 % attributable to three DMPs for lung cancer, and of 64.4 % attributable to four DMPs for a combined endpoint including all smoking-related cancers available (lung, esophagus-stomach, colorectal, liver, pancreatic and kidney). The results of the mediation analysis were largely replicated in an independent population (the Framingham Heart Study), in which we also conducted functional validation using gene expression data. In general, we found inverse association between DNA methylation and gene expression for the methylation sites identified in our mediation analysis.

In addition to these two main objectives, this thesis presents a short section focused on gene expression, the biological process directly influenced by DNA methylation, which points to future research lines. Even if mediated effects of DNA methylation on the association between environmental factors and chronic disease are identified, this does not necessarily imply causality, as unmeasured confounders and other sources of bias might exist. Thus, investigating the biological processes influenced by DNA methylation might help as functional support of its role in chronic disease.

In particular, gene expression measured in single cells (scRNAseq) is at the forefront of omics data research, as it enables the characterization of cell heterogeneity. However, these data present statistical

challenges due to high proportions of zeros obtained in gene expression measurements for each individual gene and cell.

In addition to evaluating differences in means of gene expression across groups, differences in variability have shown to be biologically relevant. Several methods have been developed for the evaluation of differential variability in omics data. However, these methods are not specific for scRNAseq data. In this thesis, we have used simulations to evaluate the impact of high proportions of zero counts in statistical methods for the identification of differentially variable genes in scRNAseq data. We found that high proportions of zeros lead to inflated variances and p-values, as well as higher false discovery rates. The distinct algorithm, which uses permutation tests to identify differences in distributions across groups, shows the best performance in terms of compromise between false discovery and true positive rates.

In summary, this thesis has contributed to the field of omics data research, both by providing novel statistical methods for DNA methylation data analysis, which can also be used for other omics, and by contributing to the body of epidemiological evidence that supports a role of environmental epigenetics in chronic disease.

Herramientas bioestadísticas para la evaluación del papel de la metilación del ADN en enfermedades crónicas relacionadas con factores ambientales

UNIVERSITÀT DE VALENCIA

Facultad de Ciencias Matemáticas

Programa de Doctorado en Estadística y Optimización

Resumen general

La epigenética se refiere al estudio de las marcas químicas que alteran la expresión génica sin cambiar la secuencia genética. Los factores ambientales y conductuales son conocidos modificadores de la epigenética, resultando así en cambios heredables que pueden dar lugar a alteraciones en procesos biológicos esenciales y, por consiguiente, al desarrollo de enfermedades.

La metilación del ADN es la marca epigenética más estudiada. Sucede cuando un grupo metilo se adhiere a la molécula del ADN. Existe amplia evidencia científica de la asociación entre factores ambientales tales como tabaco y metales, y desregulaciones en la metilación del ADN. Asimismo, existe amplia evidencia de la asociación entre desregulaciones en metilación del ADN y enfermedades crónicas, en especial para el cáncer. Sin embargo, aún está por descifrar si estas asociaciones son causales o suceden debido a que la metilación del ADN es un biomarcador de otros procesos biológicos alterados, siendo estos procesos los que influyen en las enfermedades de forma causal.

Para evaluar el papel de la metilación del ADN en la asociación entre los factores ambientales y las enfermedades crónicas, se requieren métodos estadísticos apropiados para el análisis de datos de muy altas dimensiones y altamente correlacionados. Tradicionalmente, cada posición de metilación se evaluaba en modelos de regresión separados.

Sin embargo, esta metodología no es la más adecuada, ya que no incluye todas las posiciones de metilación en un mismo modelo y, por lo tanto, no es capaz de tener en cuenta las correlaciones entre las mismas. Además, esperaríamos que las posiciones de metilación influyan de manera conjunta en la biología, y no por separado. Por ello, en los últimos años, se ha optado por incluir todas las posiciones de metilación en el mismo modelo como método preferente. Sin embargo, y puesto que en el contexto de datos ómicos nos enfrentamos a cientos de miles, e incluso millones de variables, los métodos de reducción de la dimensionalidad son esenciales para el apropiado análisis de estos datos.

En primer lugar, debemos ser capaces de seleccionar qué posiciones genómicas de metilación están asociadas con nuestra variable respuesta de interés. Los métodos de penalización, que constituyen el mecanismo más utilizado para la selección de variables, pierden capacidad predictiva y presentan sesgos en contextos de dimensiones muy altas, ya que tienden a introducir sesgo para disminuir la varianza. Además, no cuantifican la incertidumbre estadística. El algoritmo Least Absolute Shrinkage and Selection Operator (LASSO), en concreto, ha sido el más utilizado para selección de variables. Sin embargo, se ha demostrado que presenta sesgos no ignorables en la estimación de los coeficientes y, además, no es adecuado para contextos en los que existe multicolinealidad, puesto que no es capaz de seleccionar más de un predictor de un conjunto de predictores correlacionados. Métodos como elastic-net, una combinación entre LASSO y la regresión Ridge, y adaptive elastic-net (Aenet), una modificación del elastic-net que introduce pesos adaptivos en la norma L_1, han sido presentados como mejoras del algoritmo LASSO en cuanto a capacidad predictiva y reducción del sesgo en estimación de coeficientes. El algoritmo Aenet, además, cumple la propiedad de oracle, que garantiza la consistencia en estimación de coeficientes. La propiedad de oracle establece que la estimación de los coeficientes se realiza de manera igual de precisa que si se conociese con anterioridad el conjunto de variables seleccionadas. Sin embargo, estos métodos también presentan un pérdida de rendimiento cuando se aplican en dimensiones muy

altas.

Una vez seleccionado el conjunto relevante de posiciones de metilación asociadas con nuestra variable respuesta, el análisis de mediación es una herramienta útil para cuantificar el potencial efecto intermedio de estas posiciones de metilación en la asociación entre factores ambientales y enfermedades crónicas. El contexto más probable es que varias marcas de metilación (y no una única marca) sean intermediarias entre estos dos procesos, estando además posiblemente correlacionadas. Por otro lado, es habitual que las variables respuesta analizadas en contextos epidemiológicos sean de supervivencia, con el fin de incorporar al modelo el tiempo hasta el evento de salud. Sin embargo, hasta la fecha, no se han desarrollado algoritmos de mediación que incorporen múltiples mediadores correlacionados en el contexto de análisis de supervivencia. Además, los métodos tradicionales de mediación tales como el metodo de la diferencia entre coeficientes y el método del producto entre coeficientes no son capaces de incorporar las interacciones entre la exposición y el mediador.

Así pues, esta tesis consta de dos objetivos principales, el primero relacionado con la selección de variables en muy altas dimensiones, y el segundo relacionado con el análisis de mediación múltiple para datos de supervivencia. Para llevar a cabo estos objetivos, utilizamos tanto datos simulados como datos del Strong Heart Study, la cohorte prospectiva de indios americanos con más participantes y de mayor duración en Estados Unidos. Esta cohorte incluye 12 tribus de indios americanos de Arizona, Oklahoma, Dakota del Sur y Dakota del Norte, y consta de casi 30 años de seguimiento en enfermedad cardiovascular y cáncer. Asimismo, la metilación en ADN en sangre fue medida en la primera visita en 2351 participantes, junto a los metales en orina y las variables clínicas de interés. Los datos de metilación fueron procesados siguiendo los procedimientos estándar en el área, incluyendo control de calidad, normalización y corrección por variabilidad técnica no deseada. La metilación en sangre se midió usando el microarray Illumina MethylationEPIC Beadchip, obteniendo así datos de proporciones de metilación (valores beta) para cada una de las casi 800,000 posiciones genómicas incluidas en el array. Estas proporciones se transforman

usando la transformación logística en base 2 para obtener los valores M, más utilizados en el análisis estadístico debido a su naturaleza más homocedástica que la de las proporciones de metilación.

Resumen del objetivo 1. El primer objetivo de esta tesis surge de la necesidad de extender la herramienta estadística Iterative Sure Independence Screening (ISIS), que realiza selección de variables en contextos de muy altas dimensiones, para mejorar su capacidad predictiva, su estimación de efectos y para incorporar la incertidumbre estadística. El algoritmo ISIS se basa en aprendizaje de correlaciones, de manera que evalúa cada variable por separado en relación a la variable respuesta, aplicando después un método de regularización y evaluando la contribución adicional de los predictores que no han sido seleccionados de manera iterativa. Este último paso se realiza para lidiar con la situación en que existen altas correlaciones entre los predictores, situación muy habitual en datos ómicos. También para tener en cuenta las variables que no están individualmente asociadas con la variable respuesta, pero sí lo están en presencia de otras variables.

Así pues, nuestro objetivo consiste en combinar el algoritmo ISIS con los métodos de regularización elastic-net, Aenet y multi-step adaptive elastic-net (MSAenet), una versión modificada del Aenet, que aplica pesos proporcionales a la magnitud de los coeficientes tanto a la norma L_1 como a la norma L_2, mientras que el Aenet los aplica solo a la normal L_1. Estos tres métodos de regularización presentan mejoras en cuanto a reducción de sesgo y mayor capacidad predictiva con respecto a los métodos pareados con el algoritmo ISIS hasta la fecha (LASSO, Smoothly Clipped Absolute Deviation [SCAD] y Minimax Concave Penalty [MCP]). Además, tienen la capacidad de lidiar con la multicolinealidad, pudiendo seleccionar más de una variable de un conjunto de variables correlacionadas. Estos métodos de regularización se incluyeron en la función tune.fit del algoritmo SIS.

Por otro lado, nuestro objetivo incluye la implementación de un algoritmo para el cálculo de intervalos de confianza basados en bootstrap, en el que incluimos un mecanismo de control basado en los errores estándar de los intervalos de confianza para asegurar que la

variabilidad de los estimadores bootstrap no es demasiado grande. Esta herramienta se desarrolla en la función boot.ci del algoritmo SIS, y es opcional, pudiendo el usuario decidir si desea computar intervalos de confianza o no. La extensión del algoritmo SIS que hemos desarrollado ha sido incluida en el paquete SIS de R, que está disponible en el repositorio público CRAN.

En la línea de este objetivo, esta tesis incluye dos aplicaciones prácticas del algoritmo ISIS. Para ello, hemos usado datos del Strong Heart Study. La primera aplicación es metodológica y evalúa las mejoras introducidas por nuestra extensión del paquete (la inclusión de Aenet, elastic-net y MSAenet para combinar con el algoritmo ISIS) en comparación a los métodos de regularización incluidos en la versión original (LASSO, SCAD y MCP). Evaluamos, como principal variable respuesta, el índice de masa corporal como variable continua. Como variables respuesta secundarias, evaluamos la incidencia de cáncer de pulmón como variable de supervivencia y la incidencia de diabetes como variable dicotómica. La razón de considerar estas variables respuesta como secundarias es que, puesto que las variables respuesta de supervivencia y dicotómicas son menos informativas que las variables respuesta continuas, que representan un valor real, se necesita un tamaño muestral mayor para llegar a seleccionar el mismo número de variables. Por ello, SIS automáticamente fija el número máximo de variables seleccionadas por defecto a una cantidad más baja para las variables respuesta de supervivencia y dicotómicas, en comparación a las continuas. El número máximo de variables seleccionadas, sin embargo, puede ser modificado por el usuario. Utilizamos el error cuadrático medio en el conjunto de aprendizaje y en el conjunto de prueba como medida de capacidad predictiva para la variable respuesta continua. Para las variables respuesta de supervivencia y binarias, utilizamos el índice de concordancia y el área bajo la curva ROC, respectivamente. También evaluamos el número de variables seleccionadas y el coste computacional. Además, realizamos un análisis bioinformático (enriquecimiento de rutas biológicas KEGG) para evaluar la plausibilidad biológica de nuestros resultados. El algoritmo ISIS pareado con Aenet presentó una mejora en capacidad predictiva

con respecto a la versión original de ISIS, en especial para variables respuesta continuas y binarias. Además, al parear ISIS con Aenet, se obtiene una estimación de efectos más consistente debido al cumplimiento de la propiedad de oracle. Nuestro análisis bioinformático reveló que también da lugar a una selección más robusta de variables desde el punto de vista biológico. Es importante destacar que el algoritmo MSAenet no presentó ninguna mejora en capacidad predictiva con respecto a Aenet ni a elastic-net, lo cual sugiere que introducir pesos en las normas L_1 y L_2 no ofrece ninguna mejora con respecto a introducirlos solo en la norma L_1, tal como se hace en el algoritmo Aenet. Además de ello, MSAenet, SCAD y MCP podrían dar lugar a la selección de conjuntos de variables excesivamente pequeños, lo cual es una limitación en el contexto de datos ómicos, puesto que si la finalidad de la selección de variables es el descubrimiento biológico y no la predicción, el análisis estadístico debería estar centrado en no dejar de seleccionar variables importantes.

La segunda aplicación del algoritmo ISIS incluida en esta tesis es un estudio epidemiológico que evalúa el potencial rol intermedio de los cambios en metilación del ADN en la ampliamente documentada asociación entre el arsénico y la enfermedad cardiovascular. El arsénico es un metaloide tóxico presente en el agua, en el aire, en los alimentos y en la tierra de ciertos terrenos. El arsénico ha sido asociado con enfermedad cardiovascular, incluso en dosis bajas o moderadas, pero los mecanismos biológicos no han sido esclarecidos. Una de las potenciales rutas biológicas mediante las que el arsénico podría estar asociado a la enfermedad cardiovascular es a través de la epigenética. Para analizar esta cuestión, empleamos el algoritmo ISIS pareado con Aenet para seleccionar las posiciones de metilación asociadas con la enfermedad cardiovascular, y posteriormente realizamos un análisis de mediación simple en esas posiciones. Encontramos efectos mediados estadísticamente significativos en 21 y 15 posiciones diferencialmente metiladas (DMPs) para incidencia cardiovascular y mortalidad cardiovascular, respectivamente. Además, de las 21 DMPs con efectos mediados significativos para enfermedad cardiovascular, seis fueron replicadas en tres cohortes americanas independientes (Fram-

ingham Heart Study, Women's Health Initiative y Multi-Ethnic Study of Atherosclerosis) con la misma dirección de asociación. Esto indica que la metilación en esas seis posiciones genómicas se asocia de manera robusta con la enfermedad cardiovascular. Los genes asociados a las posiciones de metilación significativas en nuestro análisis de mediación también fueron replicados en un estudio animal con ratones. Las funciones biológicas de estos genes, ampliamente relacionadas con la enfermedad cardiovascular, proporcionan evidencia de la robustez de los resultados. En conclusión, encontramos que parte de la asociación del arsénico con la enfermedad cardiovascular estaría potencialmente mediada por cambios en posiciones de metilación con funciones biológicas asociadas con enfermedad cardiovascular. Sin embargo, este análisis también revela la necesidad de realizar análisis de mediación múltiple, puesto que los efectos mediados individuales podrían estar inflados debido a interrelaciones entre las rutas biológicas y a correlaciones entre las posiciones de metilación.

Resumen del objetivo 2. El segundo objetivo de la tesis se centra en la extensión del algoritmo multimediate, que realiza análisis de mediación múltiple para mediadores correlacionados, a datos de supervivencia. Jerolon y colaboradores desarrollaron este algoritmo para variables respuesta continuas y binarias. En esta tesis, extendimos ese algoritmo a variables respuesta de supervivencia. Los modelos aditivos son más apropiados que los multiplicativos para realizar análisis de mediación en el contexto de supervivencia, debido a la no colapsibilidad del hazard ratio y del odds ratio. El modelo de Lin-Ying es un modelo aditivo de Aalen que tiene coeficientes y covariables invariantes en el tiempo, de modo que el único término que varía en el tiempo en este modelo es el riesgo base. Utilizando los modelos aditivos de Lin-Ying, hemos extendido los resultados teóricos para la identificación de efectos mediados, directos y totales, así como el propio algoritmo, al contexto de supervivencia. Usando el marco contrafactual, hemos demostrado teóricamente que estos efectos pueden ser calculados usando la tasa, cuando el modelo de la variable respuesta en mediación está definido como un modelo aditivo de Lin-Ying. Nótese que para que estos efectos puedan ser identificados, deben cumplirse las asunciones

de ignorabilidad secuencial para múltiples mediadores. Estas asunciones consisten en que no haya confusores no medidos en las asociaciones entre la exposición y el mediador, la exposición y la variable respuesta y el mediador y la variable respuesta (condiciones de intercambiabilidad), así como el hecho de que no haya múltiples versiones del tratamiento o la exposición (consistencia), y que haya individuos expuestos y no expuestos en cada uno de los estratos de los confusores (positividad).

Asimismo, hemos adaptado el algoritmo multimediate para la incorporación de potenciales interacciones entre la exposición y el mediador. Estas interacciones no se pueden tener en cuenta de forma directa en los métodos tradicionales para el análisis de mediación, que incluyen el método de la diferencia entre coeficientes y el método del producto entre coeficientes. La extensión de este algoritmo a datos de supervivencia está disponible en el siguiente repositorio de Github: *https://github.com/AllanJe/multimediate*. La extensión que incluye la posibilidad de considerar interacciones entre la exposición y el mediador se incluirá próximamente en el mismo repositorio.

En este segundo objetivo, también se incluyeron dos aplicaciones a datos de este algoritmo. La primera es un estudio de simulación en el que se evaluó el rendimiento del algoritmo multimediate en comparación a los algoritmos de mediación simple. Para ello, simulamos tres mediadores, con una variable de exposición dicotómica que tomaba valores 0 o 1, para simplificar el escenario. Además, consideramos tres escenarios de correlaciones (correlaciones negativas, correlaciones positivas o mediadores incorrelados), y tres escenarios distintos para el riesgo base. El primero de ellos contempla un riesgo base constante, el segundo, un riesgo base dependiente del tiempo, y el último, un riesgo base no monotónico. Estos riesgos fueron utilizados posteriormente para simular los tiempos de supervivencia. Se calcularon errores cuadráticos medios, sesgo y varianza de los efectos estimados, así como el porcentaje de cobertura de los intervalos de confianza al 95 %. Los resultados de nuestra simulación muestran la superioridad del algoritmo multimediate con respecto a la mediación simple, incluso en el caso de mediadores no correlacionados. El algoritmo mul-

timediate presenta un menor error cuadrático medio, sobre todo para el efecto indirecto en contextos de mediadores correlacionados. Se percibe una especial mejora del algoritmo multimediate con respecto a mediación simple en la cobertura de los intervalos de confianza, que baja notablemente en el caso de mediación simple cuando los mediadores están correlacionados. El algoritmo multimediate mantiene una excelente cobertura de los intervalos de confianza en cualquier escenario. La definición del riesgo base no influyó en el rendimiento del algoritmo, ofreciendo resultados similares en los tres escenarios. Este algoritmo asume que las correlaciones entre los mediadores no dependen del tratamiento. Futuras líneas de trabajo deberían incluir evaluaciones de potenciales violaciones de esa asunción y posibles formas de relajar la misma.

La segunda aplicación relacionada con el algoritmo multimediate es un estudio epidemiológico en el que estudiamos el papel intermedio de múltiples marcadores de metilación, potencialmente correlacionados, en la asociación entre el tabaco y el cáncer usando datos del Strong Heart Study. El tabaco es la exposición ambiental con asociación con la metilación del ADN más ampliamente documentada. Esta asociación ha sido reportada en poblaciones de todo el mundo. Los genes más conocidos en relación con el tabaco y la metilación son *AHRR* y *F2RL3*, que tienden a ser hipometilados por el tabaco. Utilizamos el algoritmo ISIS pareado con elastic-net para seleccionar posiciones de metilación asociadas con cáncer y posteriormente evaluamos estas posiciones en un análisis de mediación simple. Se obtuvieron 29 posiciones de metilación con efectos mediados significativos para cáncer de pulmón, y 37 para una variable respuesta combinada de todos los cánceres asociados con el tabaco de los que disponíamos datos (pulmón, esófago-estómago, colorrectal, hígado, páncreas y riñón). Posteriormente, introdujimos las posiciones de metilación que resultaron significativas en el análisis de mediación simple en el algoritmo multimediate, para evaluar varias posiciones de metilación como potenciales mediadores conjuntos en la asociación entre el tabaco y el cáncer. El algoritmo multimediate detectó un efecto mediado conjunto del 81.3 % atribuible a tres posiciones de metilación para el

cáncer de pulmón (incluyendo el gen *AHRR*), y del 64.4 % atribuible a cuatro posiciones de metilación para la variable respuesta combinada de cánceres asociados con el tabaco. Así, aunque el análisis de mediación simple detectó que muchas posiciones de metilación presentaban efectos mediados, al evaluar el efecto conjunto mediado por todas las posiciones de metilación solo tres y cuatro posiciones de metilación fueron relevantes. Esto ilustra el hecho de que muchas posiciones de metilación presentaban efectos mediados por el mero hecho de estar correlacionadas con otras posiciones de metilación. Asimismo, los resultados del análisis de mediación fueron ampliamente replicados en una población independiente (Framingham Heart Study), en la que también llevamos a cabo validación funcional con datos de expresión génica. En general, encontramos una asociación inversa entre metilación del ADN y expresión génica en las posiciones de metilación identificadas en nuestro análisis de mediación. También realizamos un análisis bioinformático mediante análisis de enriquecimiento de rutas biológicas KEGG, en el que encontramos que muchas de las rutas enriquecidas en nuestros resultados estaban asociadas con cáncer. Estos resultados contribuyen a la identificación de potenciales mecanismos biológicos relacionados con la asociación entre el tabaco y el cáncer. Son necesarios estudios experimentales para evaluar la potencial asociación causal, ya que no se puede descartar la presencia de confusores no medidos en estudios observacionales.

Además de estos dos objetivos principales, esta tesis presenta un breve apartado relacionado con la expresión génica, el proceso directamente influenciado por la metilación del ADN. Incluso obteniendo efectos mediados significativos de la metilación del ADN en la asociación entre exposiciones ambientales y enfermedades crónicas, desconocemos si este efecto es causal o no, debido, entre otros tipos de sesgos, a que podrían existir confusores no medidos. Así pues, estudiar los procesos biológicos que son influenciados por la metilación del ADN podría contribuir a evaluar su papel en las enfermedades crónicas.

El proceso biológico directamente influenciado por la metilacion

del ADN es la expresión génica, que posteriormente daría lugar a la creación de proteínas o RNAs sin codificar. La secuenciación del RNA es el método más popular para medir expresión génica. Mediante este método, se obtiene una matriz de conteos que representa la expresión génica en cada gen y muestra biológica. Sin embargo, agrupar grandes cantidades de células en muestras da lugar a pérdida de información e incapacita la cuantificación de la heterogeneidad calular. La expresión génica medida en forma de secuenciación de células individuales (scRNAseq) se sitúa a la vanguardia de la investigación de los datos ómicos, debido a su capacidad para capturar y evaluar la heterogeneidad celular. Esta tecnología es capaz de medir la expresión génica en células individuales. Sin embargo, estos datos presentan retos estadísticos para su análisis, debido a las grandes proporciones de ceros que se obtienen en las mediciones de la expresión génica para cada gen y célula en la matriz de conteos.

Además de evaluar diferencias en medias de expresión entre grupos, las diferencias en variabilidad de expresión han demostrado ser biológicamente relevantes, por ejemplo para el envejecimiento o el cáncer. Varios métodos han sido desarrollados para la identificación de variabilidad diferencial en datos ómicos, aunque no para datos de scRNAseq. En esta tesis hemos evaluado, usando datos simulados, cómo influye la presencia de grandes proporciones de ceros en los métodos estadísticos utilizados para la identificación de genes diferencialmente variables en datos de scRNAseq. Para ello, hemos simulado datos de scRNAseq usando la librería muscat de R, y hemos aplicado diversas técnicas estadísticas que podrían favorecer la evaluación de diferencias en variabilidad de expresión génica entre grupos. Estas técnicas incluyen los algoritmos diffVar, SuperCell, SAVER, distinct y scDD. Hemos concluido que la presencia de altas proporciones de ceros da lugar a varianzas y p-valores inflados, así como a subidas en las tasas de descubrimientos falsos. Las tasas de verdaderos descubrimientos, por el contrario, no se ven afectadas por la introducción de grandes proporciones de ceros. La agrupación de células con perfiles de expresión génica parecidos, realizada por el algoritmo SuperCell, no mejoró las tasas de falsos positivos y falsos negativos obtenidas. Tam-

poco lo hizo la imputación de los ceros llevada a cabo por el algoritmo SAVER. El algoritmo distinct, que utiliza tests de permutaciones para identificar diferencias en distribuciones entre grupos, es el que mejores resultados presenta en cuanto a equilibrio entre tasa de verdaderos descubrimientos y de falsos descubrimientos. Sin embargo, es necesario el desarrollo de algoritmos que lleven a cabo la identificación de variabilidad en expresión génica para datos de scRNAseq, puesto que la herramienta distinct no es específica para la cuantificación de diferencias en variabilidad, sino de diferencias en distribución.

En resumen, esta tesis ha contribuido al área científica de los datos ómicos, tanto mediante el desarrollo de métodos estadísticos innovadores para el análisis de datos de metilación del ADN, como realizando contribuciones a la evidencia epidemiológica relacionada con metilación del ADN en asociación con exposiciones ambientales y enfermedades crónicas. Hemos mejorado la herramienta SIS para facilitar la selección de variables en muy altas dimensiones mejorando la capacidad predictiva, la estimación de coeficientes e incorporando incertidumbre estadística. Por otro lado, hemos implementado el algoritmo multimediate para la evaluación de múltiples mediadores correlacionados en el contexto de analisis de supervivencia. Hemos utilizado nuestras novedosas herramientas estadísticas para identificar efectos mediados de las diferencias en metilación del ADN en la asociación entre el arsénico y la enfermedad cardiovascular, y en la asociación entre el tabaco y los cánceres asociados al tabaco. También hemos mostrado la plausibilidad biológica de nuestros resultados realizando análisis bioinformáticos.

Futuras potenciales líneas de trabajo deberían incluir la optimización del algoritmo SIS para bajar el coste computacional. Otra de las potenciales futuras líneas de investigación podría ser la implementación del algoritmo multimediate para casos en los que existen correlaciones causales entre los mediadores, puesto que la versión actualmente implementada solo contempla los casos en los que los mediadores son, o bien independientes, o bien no causalmente correlacionados. Además, sería importante adaptar los métodos para análisis de sensibilidad existentes para análisis de mediación al algoritmo multimediate, puesto

que las asunciones de confusores no medidos en las asociaciones entre la exposición y el mediador, el mediador y la variable respuesta y la exposición y la variable respuesta son imposibles de verificar en la práctica para estudios observacionales. Por ello, cuantificar el potencial sesgo que estos confusores podrían introducir en nuestros efectos mediados constituiría un trabajo futuro de interés.

Contents

List of Algorithms

List of Tables

List of Figures

Abbreviations

LASSO	Least Absolute Shrinkage and Selection Operator
SCAD	Smoothly Clipped Absolute Deviation
MCP	Minimax Concave Penalty
ISIS	Iterative Sure Independence Screening
CI	Confidence interval
CpG	Cytosine followed by guanine with a phosphate link
5-mC	5-methylcytosine
AHRR	Aryl Hydrocarbon Receptor Repressor gene
F2RL3	Coagulation factor II receptor-like 3 gene
CVD	Cardiovascular disease
CHD	Coronary Heart Disease
RNAseq	RNA sequencing
scRNAseq	Single cell RNA sequencing
FWER	Family-wise error rate
FDR	False Discovery Rate
EWAS	Epigenome-Wide Association Study
GWAS	Genome-Wide Association Study
SNPs	Single nucleotide polymorphisms
Aenet	Adaptive elastic-net
MSAenet	Multi-step adaptive elastic-net
MCMC	Markov Chain Montecarlo
SIS	Sure Independence Screening
MMLE	Maximum Marginal Likelihood Estimator
SHS	Strong Heart Study
BMI	Body Mass Index

NK	Natural killer cells
PCs	Principal Components
HbA1c	Glycosylated Hemoglobin
EPIC	Illumina Infinium MethylationEPIC Beadchip microarray
MSE	Mean Squared Error
C index	Concordance index
AUC	Area under the ROC curve
DMPs	Differentially Methylated Positions
ISIS-Aenet	Iterative Sure Independence Screening paired with adaptive elastic-net
ISIS-enet	Iterative Sure Independence Screening paired with elastic-net
450K	Illumina Infinium HumanMethylation450K Beadchip microarray
FHS	Framingham Heart Study
WHI	Women's Health Initiative
MESA	Multi-Ethnic Study of Atherosclerosis
DMRs	Differentially Methylated Regions
MMA	Monomethylarsonate
DMA	dimethylarsinate
LDL	Low-density lipoprotein
HDL	High-density lipoprotein
US	United States
EPA	Environmental Protection Agency
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
ApoE-/-	Apolipoprotein E knockout (apoE-/-)
IQR	Interquartile range
ncRNA	Non-coding RNA

SIMMA	Sequential Ignorability for Multiple Mediators Assumptions
SUTVA	Stable Unit Treatment Value Assumption
eQTM	Expression quantitative trait methylation
HR	Hazard Ratio
BCV	Biological coefficient of variation
TPR	True positive rate
kNN	k nearest neighbors
SAVER	Single-cell Analyses Via Expression Recovery
CDF	Cumulative distribution function

CHAPTER 1

Introduction

1.1 Motivation

The word *omics* comes from Greek and refers to the study of the whole or the totality of something. However, we use it to refer to the study of an organism in its different levels. Genomics, epigenomics, transcriptomics, proteomics and metabolomics complete the study of an organism from its genetic code to the metabolites it generates [2].

Epigenetic changes, or heritable phenotype changes that do not alter the DNA sequence, have shown to be highly influenced by environmental factors [3] and, in turn, have been proposed to influence chronic disease [4, 5]. The complexity of epigenomic data and the lack of appropriate statistical methods, though, have hindered the precise quantification of the association between epigenetic marks and chronic disease, including the potential intermediate role of epigenetic changes on the well-known association between environmental factors and chronic disease [6].

Several characteristics of omics data challenge the development of

appropriate statistical methods to analyze these data. First, the ultra-high dimensional nature of omics data requires effective dimensionality reduction techniques in order to select the features that are related to the outcome of interest, and focus subsequent extensive statistical analyses in those features. Shrinkage methods such as Least Absolute Shrinkage and Selection Operator (LASSO), which have been widely used for variable selection in high-dimensional settings [7], have shown to worsen their performance in ultra-high dimensional settings for both variable selection and effect estimation [8]. On the other hand, high correlations (both spatial and non-spatial) between features challenge the performance of traditional shrinkage methods. Thus, Fan and Lv proposed to use a variable selection technique fulfilling the sure screening property (i.e., high probability of selecting the optimal variable set) combined with shrinkage methods for variable selection in ultra-high dimensional settings [8]. This method, however, was paired with shrinkage methods (LASSO, Smoothly Clipped Absolute Deviation [SCAD] and Minimax Concave Penalty [MCP]) that present limitations, and had no way to quantify uncertainty. Statistical methods for variable selection in ultra-high dimensional settings that minimize the error in effect estimation and are able to quantify statistical uncertainty are needed.

Once we are able to select the optimal set of omics features associated with the outcome, a problem of interest in epidemiologic research is to quantify the amount of the effect of an exposure or treatment on an outcome that is mediated by changes in those omics features. Jerolon et al. proposed the multimediate algorithm [9], a quasi-bayesian algorithm that is able to conduct multiple mediation analysis in the context of correlated mediators. However, this algorithm was limited to continuous or dichotomous outcomes. Survival outcomes are widely used in epidemiologic research as they allow the incorporation of the time in which the event happened to the analysis of interest. To our knowledge, no statistical tools able to conduct multiple mediation analysis in the context of correlated mediators for survival analysis had been developed prior to this work.

Provided we find evidence of an intermediate role of DNA methy-

lation on the association between environmental factors and chronic disease, this would not be enough to establish a causal association. Whether DNA methylation dysregulations play a causal role in the development of chronic disease or are just biomarkers of other underlying disrupted biological processes would remain unclear. To study the biological processes that are influenced by DNA methylation and that could be important for disease development, the impact of DNA methylation in subsequent omics processes needs to be analyzed.

The aim of this dissertation was to develop a biostatistical toolkit to study the environmental epigenetics of chronic disease. We focused on high dimensional genome-wide DNA methylation, the most widely studied epigenetic mark, with special interest in variable selection and mediation analysis. In addition, given that DNA methylation directly influences gene expression, we present some preliminary work on this omics layer. In the following sections, we provide an introduction to DNA methylation and its association with environmental factors and chronic disease, as well as an introduction to transcriptomics. The main objectives of the thesis are presented in section 1.4.

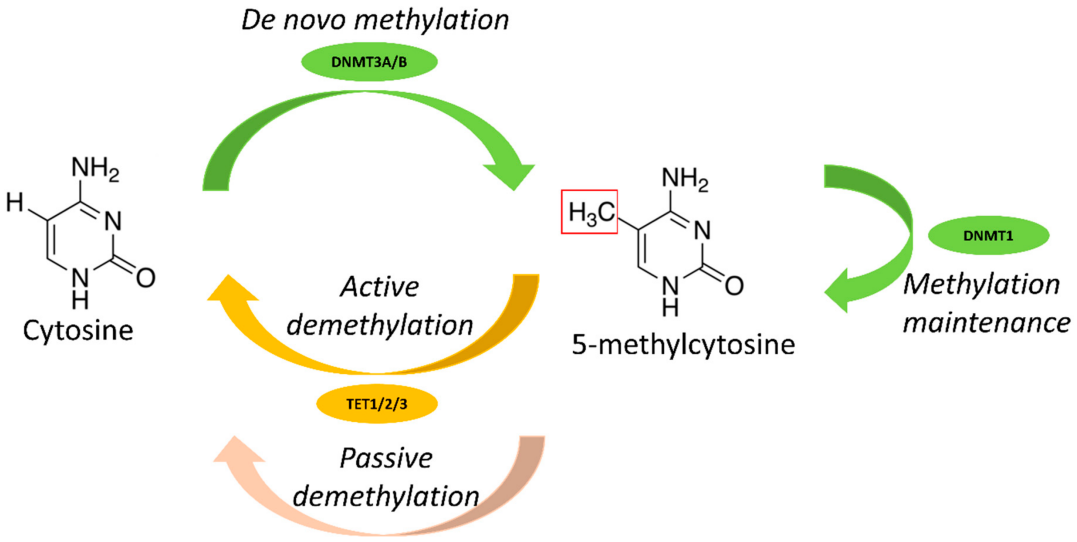
1.2 Environmental epigenetics and chronic disease

The term epigenomics refers to changes in gene regulation that do not affect the underlying DNA sequence. Although this term was not in the spotlight until the XXI-st century, it was first defined in 1942 as the concatenation of interactions between genotype and phenotype so that disturbances at early stages may cause far-reaching abnormalities in organs and tissues [10]. Nowadays, epigenetics is considered as the study of environmental and behavioral factors that alter gene expression in a heritable manner without changing the genomic sequence. Epigenetic changes have the ability to influence whether genes are more or less expressed. These modifications can last from a few minutes to a whole lifetime, therefore having direct impact in biological processes of human health. Thus, the use of epigenetics for both early detection of disease and disease treatment has been the focus of intense research in the area of biomedical sciences in the last years [11, 12].

DNA methylation is the most studied epigenetic mark [13]. Methylation is predominantly found at genomic sites presenting a cytosine nucleotide followed by a guanine, with a phosphate link (hereinafter referred to as CpG sites). It occurs through the attachment of a methyl group (CH_3) onto the C5 position of the cytosine, which leads to 5-methylcytosine (5-mC). The process is shown in Figure 1.1. Methylation is generally measured on a proportion scale (between 0 and 1), which represents the proportion of methylated cytosines for each genomic position.

DNA methylation is essential for normal cellular development and involved in several key biological processes. Other less studied, but relevant, epigenetic marks include DNA hydroxymethylation, histone modifications, RNA transcripts or microRNAs. Environmental factors such as exposure to chemicals, diet, physical exercise or stress are known regulators of epigenetic changes [14].

Figure 1.1: Summary of biological processes involved in DNA methylation.



Source: Lebecque et al. 2021 [15].

1.2.1 Environmental factors and DNA methylation

Environmental chemicals have shown to be major contributors to dysregulations of DNA methylation [16]. Among other explanations, they can influence one-carbon and citric acid metabolism pathways, leading to dysregulations of DNA methylation [17]. Smoking, a complex mixture of chemical compounds, has been robustly associated with DNA methylation in populations all over the world [18, 19, 20, 21]. On the other hand, metals such as arsenic are classified as group 1 carcinogens by the International Agency for Research on Cancer [22]. However, given that metals are poor mutagenics (with the exception of chromium), the biological processes that link metals exposure to disease are not well understood. DNA methylation has been proposed as a potential biological mechanism underlying the association between environmental exposures and chronic disease [6].

Cigarette smoke is the exposure that has reached the greatest consensus in terms of specific DNA methylation dysregulation patterns being commonly found and robust across populations, as stated in this meta-analysis [18]. For instance, DNA methylation dysregulations of

multiple CpGs annotated to the *AHRR* gene (Aryl Hydrocarbon Receptor Repressor, which mediates dioxin toxicity and is involved in cell growth and differentiation [23]) and the *F2RL3* gene (Coagulation factor II receptor-like 3, also known as *PAR-4*, which plays a role in blood coagulation, inflammation and response to pain [24]) have been associated with smoking in several studies [19, 20, 18, 21]. In addition to *AHRR* and *F2RL3*, other genes such as *PRSS23* and *GPR15* have also been consistently associated with smoking in methylome-wide epidemiologic studies [18]. Whether DNA methylation might be a causal mechanism through which smoking causes disease, however, remains unknown.

On the other hand, exposure to inorganic arsenic is a global health problem. Even at low exposure levels in water and food, arsenic has been related to multiple health outcomes including cardiovascular disease (CVD) [25, 26, 27]. CVD outcomes associated with arsenic in Bangladesh, Chile, Taiwan, Denmark, Spain and the United States include coronary heart disease (CHD) [28, 29, 26, 30, 31], stroke [26], peripheral arterial disease [32] and overall CVD mortality [26, 33, 34]. Arsenic has also been prospectively associated with changes in blood pressure levels [31, 35] and carotid atherosclerosis [31, 36, 37]. These epidemiological findings are consistent with data from animal models showing that arsenic can induce atherosclerosis at relatively low exposure levels [38, 39].

The recognition of arsenic as a CVD risk factor, however, remains hindered by limited understanding of the specific mechanisms involved. Growing evidence points to the importance of epigenetic dysregulation and its influence on gene transcription pathways as a potential mechanism for arsenic-related CVD. Indeed, arsenic has been associated with changes in DNA methylation in epigenome-wide association studies (EWAS) in human populations from Bangladesh [40, 41, 42, 43, 44], South America [45, 46], Taiwan [47], China [48], and the US [49, 50, 51]. Arsenic might influence DNA methylation through the inhibition of DNA methyltransferases by repressing expression of the DNA methyltransferase genes *DNMT1* and *DNMT3A* [52].

The well-documented influence of environmental factors on DNA methylation dysregulations might be a plausible explanation of part of the influence of environmental exposures in chronic disease. Indeed, DNA methylation dysregulations have been related to several chronic disease, with overwhelming evidence especially for different types of cancer [53, 54].

1.2.2 DNA methylation and chronic disease

DNA methylation dysregulations have shown to start several years before disease onset, which provides great opportunity for early detection of disease. In particular, extensive literature exists supporting the association of DNA methylation changes with several types of cancers including lung [55, 56, 57, 58, 59], colorectal [60, 61], liver [62, 63], kidney [64, 65], pancreatic [66, 67], esophagus and stomach [68, 69], and lymphatic-hematopoietic [70] cancers, among others. Aberrant DNA methylation occurs in early stages of tumorigenesis and has been associated with cancer-related biological processes including oxidative stress [71] and apoptosis [72]. Many types of human cancers show hypermethylation of regulatory regions of certain tumor-suppressor genes [73]. DNA methylation-based biomarkers have been a target for early detection of cancer [74, 75, 76] due to their early and frequent emergence in tumors, their high quality measurement by well-established methods, their stability over time, their presence in different body fluids, and their cell type specificity. In addition, DNA methylation has shown to be consistent across large genomic regions [77], thus enabling the use of multiple CpG sites for a more robust prediction. In fact, several diagnostic kits using DNA methylation-based epigenetic biomarkers for early detection are in clinical use nowadays for cervical, oral, colorectal, lung, breast, liver, ovarian and prostate cancers, among others [78].

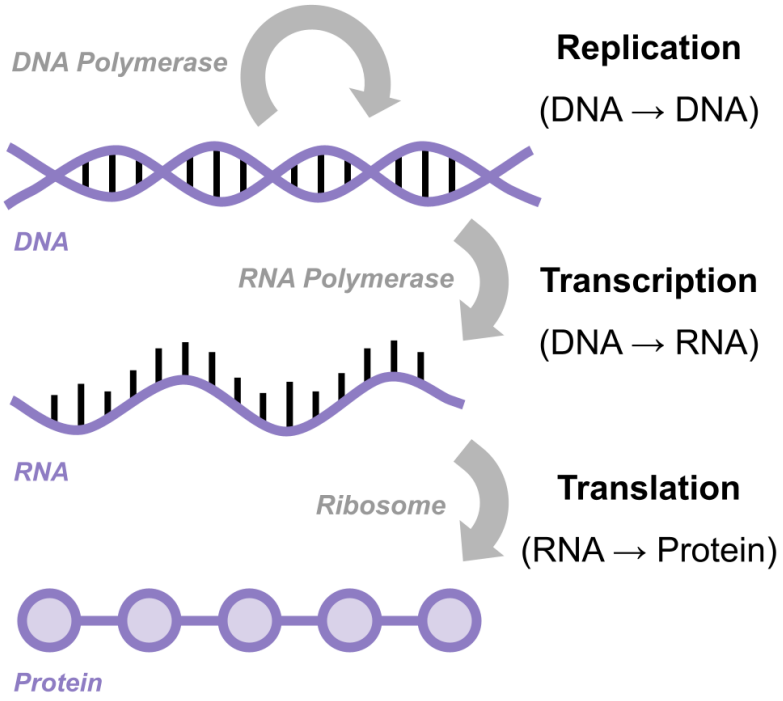
For other clinical traits, the evidence is less clear. For CVD, for instance, little consensus has been reached between studies for a common epigenomic signature. We recently conducted an EWAS of CHD including five cohorts. In this study, we found a complex and highly

population-specific epigenomic signature of CHD, with only few common differentially methylated positions (DMPs) across cohorts [79]. This might reflect that DNA methylation dysregulations associated with CVD are population-specific. More epidemiologic and experimental studies are needed to elucidate the potential role of DNA methylation on CVD.

1.3 Impact of DNA methylation in the genome structure: transcriptomics

Several biological processes might be involved on the association between DNA methylation and chronic disease, including the biological products affected by DNA methylation. Transcriptomics is the omics field that studies gene expression, the process by which information encoded in a gene is used to produce RNA, which will eventually lead to the synthesis of proteins or non-coding RNAs [80]. In transcription, DNA sequences are copied to RNA using an enzyme called RNA polymerase (Figure 1.2) [81].

Figure 1.2: From DNA sequence to translation into protein: the role of RNA transcription.



Source: National Center for Multiscale Modeling of Biological Systems (https://biologicalmodeling.org/motifs/transcription).

DNA methylation is known to influence gene expression [13]. Traditionally, DNA hypermethylation has been considered to repress transcription, especially when it happens in gene promoter regions, while

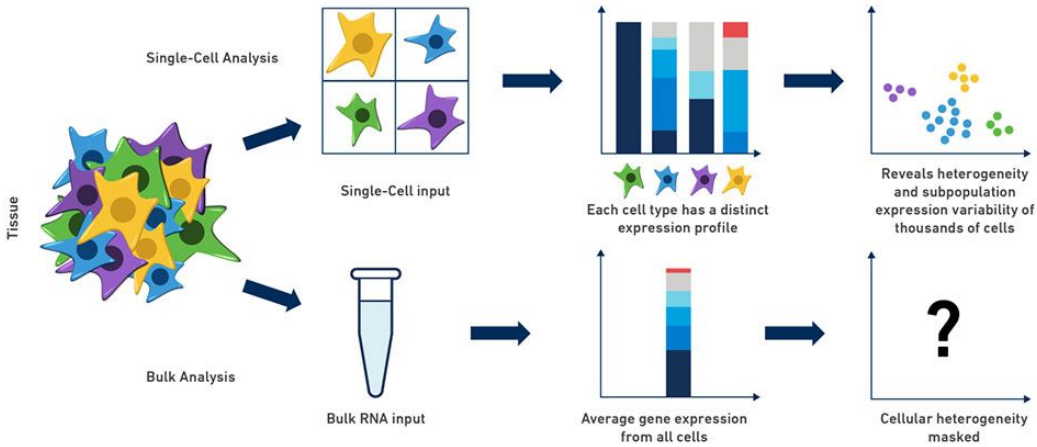
hypomethylation has been considered to increase gene expression [82]. Specifically in cancer, methylation in promoter regions of tumor suppressor genes has shown to lead to gene silencing [53]. However, more recent research has shown that the role of DNA methylation in transcription is more complex, and it differs through genomic positions [83, 84]. DNA methylation is thought to affect gene expression through remodeling of chromatine structure [85], however, establishing a direct correspondence between DNA methylation and gene expression is far from straightforward with the current understanding of the genome.

The most widely used method to measure gene expression is RNA sequencing (RNAseq) [86], which uses next-generation sequencing to measure the quantity of RNA in a biological sample. One read refers to a sequenced RNA fragment. The reads obtained from next generation sequencing are aligned to a reference genome, and the number of reads mapped to each gene are counted. The number of sequencing reads mapped to a given gene is an estimation of the expression level of that gene. This leads to a counts matrix, which is the matrix we statistically analyze after preprocessing.

RNAseq has led to extensive discovery and innovation in medicine over recent years. However, grouping large numbers of cells in biological samples results in loss of information and does not allow detailed assessment of the cells or the individual nuclei that package the genome. Current technologies allow to measure gene expression in single cells, which is useful to analyze cellular population heterogeneity, to identify cellular subtypes and to analyze the behavior of individual cells (Figure 1.3). This technology, known as single cell RNAseq (scRNAseq), allows the comparison of transcriptomes of individual cells, thus being useful to assess transcriptional differences and similarities within populations of cells [87]. This technique is able to reveal regulatory relationships between genes, and identify trajectories of different cell lineages in development, for example. In cancer, major heterogeneities between cells of the same tumor arise due to genetic and epigenetic factors, thus challenging treatment effectiveness. scRNAseq technologies would be able to characterize this heterogeneity, identify cell subtypes and measure mutation rates with the ultimate goal of

guiding diagnosis and treatment [88].

Figure 1.3: Comparison between single cell RNA-sequencing and bulk RNA-sequencing methods.



Source: 10x Genomics.

1.4 Objectives

The main objective of this thesis was to extend existing statistical methods to enable the evaluation of the role of DNA methylation in environment-related chronic disease. To do so, this thesis had two main objectives that focus on approaching different statistical challenges related to the analysis of differences in DNA methylation.

- **Objective 1: To pair the ISIS tool with Aenet, elastic-net and MSAenet to improve predictive accuracy and minimize the error in effect estimation, and to incorporate a bootstrap-based confidence interval approach to ISIS to quantify statistical uncertainty.**

  We used novel shrinkage methods such as elastic-net and Aenet to extend the existing Iterative Sure Independence Screening (ISIS) statistical tool [8] and incorporated statistical uncertainty by calculating bootstrap confidence intervals (CIs). This tool is able to conduct variable selection in ultra-high dimensional settings while dealing with multicollinearity.

  Two practical applications of the extension of this tool are included in this dissertation: "Comparison of regularization methods for the evaluation of blood DNA methylation as a marker of health endpoints", which is under journal review, and "Arsenic Exposure, Blood DNA Methylation and Cardiovascular Disease", which was published in the journal *Circulation Research* [89]. We also included the extension of this algorithm in the *SIS* R package, available in CRAN [90].

- **Objective 2: To develop an extension of the multimediate algorithm, which conducts mediation analysis for multiple correlated mediators, to time-to-event outcomes.**

  We extended the multimediate algorithm to time-to-event outcomes, and provided theoretical results as well as a simulation

study to show the improvements beyond single mediator models. The resultant paper is under journal review. We additionally adapted the algorithm to accommodate exposure-mediator interactions. A practical application of the extension of this tool to population-based data is included in this thesis: "Smoking, DNA methylation and smoking-related cancers", which is under journal review. We included the survival version of this algorithm in the *multimediate* R package, available in Github.

In addition to the two main objectives, we also conducted some subsequent work focused on gene expression, the process by which the information encoded in a gene is turned into a biological function. Gene expression is the direct biological consequence of DNA methylation. We particularly focused on single cell RNA sequencing (scRNAseq) gene expression, which is in the spotlight of the omics research community due to its potential to identify cell heterogeneity. In chapter 5, using simulated data, we analyzed whether existing methods to analyze transcriptional differences between groups for other omics data types are able to capture differences in transcriptional variability in scRNAseq data.

1.5 Study population

The population-based data applications included in this thesis use data from the Strong Heart Study (SHS), a prospective cohort study funded to investigate CVD and its risk factors in American Indian adults [91]. It is the largest and longest study of CVD in American Indian communities. In 1989–1991, a total of 4,549 men and women aged 45–75 years who were members of 13 tribes based in Arizona, Oklahoma, North Dakota, and South Dakota accepted invitations to participate. Participants without sufficient urine for metal analyses were excluded (N=576). Due to tribal request, samples from one of the tribes were not selected for DNA methylation analyses, leaving 3,515 participants. Among them, participants who were free of CVD and not missing urinary metals or other variables of interest at baseline (1989–1991) were eligible for blood DNA methylation analyses (N=3,105). Sufficient blood was available for DNA methylation analyses in 2,350 participants.

Trained and certified nurses and medical examiners collected information on sociodemographic factors (age, sex, study region, education level), medical history, smoking status (never, former, current), and cumulative smoking dose (cigarette pack-years) in a personal interview. Participants having smoked ≥ 100 cigarettes in their lifetime and smoking at the time of the interview were considered current smokers. Non current smokers who had smoked > 100 cigarettes in their lifetime were classified as former smokers. Cigarette pack-years were calculated as the number of 20-cigarette packs smoked per day times the number of years the person smoked, with zero assigned to never smokers. A physical exam was conducted, including anthropometric measures (height and weight to measure body mass index [BMI]), and collected fasting blood and spot urine samples. Height was measured standing in centimeters rounded to the nearest integer, and weight was measured in kilograms using a scale that was re-zeroed each day and calibrated against a known 22.68 kilograms weight every month.

Blood DNA methylation measurements and statistical preprocessing

DNA methylation was measured at the time of physical examination and interview for the assessment of baseline smoking status and sociodemographic variables. Buffy coats from fasting blood samples were collected in 1989–1991. Biological specimens were stored at 70° C. DNA from white blood cells was extracted and stored at the Penn Medical Laboratory, MedStar Health Research Institute under a strict quality-control system. In 2015, blood DNA was shipped to the analytical laboratory at the Texas Biomedical Research Institute for DNA methylation analysis. DNA was bisulfite-converted with the EZ DNA methylation kit (Zymo Research) according to the manufacturer's instructions. Bisulfite-converted DNA was measured using the Illumina MethylationEPIC BeadChip (referred to as 850K hereinafter), which provides a measure of DNA methylation at a single nucleotide resolution at $>$ 850,000 CpGs. Samples were randomized across and within plates to remove potential batch artifacts and confounding effects, and replicates and across-plate control samples were included on every plate.

All the preprocessing was conducted using R version 3.6.1. Data were read in six different batches (of $\sim$ 400 individuals each) and combined using the R package minfi [92]. Detection p-values were calculated using the *detectionP* function. This function calculates the total DNA signal (Methylated + Unmethylated) for each position as compared to the background signal level. The background is estimated using negative control positions, assuming a normal distribution. Then, a p-value of reliability of the DNA methylation signal is computed for each genomic position. Positions with high detection p-values should not be trusted. We removed CpGs with a p-detection value greater than 0.01 in more than 5 % of the individuals (6,159 CpGs).

Two different normalization procedures were applied. First, microarray data must be background corrected to remove the effects of non-specific binding or spatial heterogeneity across the array. Single sample noob normalization was conducted using the *preprocessNoob*

function in the R package *minfi* [93, 94], which includes a background correction with dye-bias normalization for Illumina Infinium methylation arrays. This method uses normal-exponential convolution for background correction. On the other hand, the EPIC microarray uses two types of probes—Infinium I (type I) and Infinium II (type II)—in order to increase genome coverage. However, differences in probe chemistries result in different type I and II distributions of methylation values, which might introduce bias in the downstream analyses. Thus, regression on correlated probes normalization was applied to correct for probe type bias using the R package *ENmix* [95]. This method uses the existing correlation between pairs of nearby type I and II probes to adjust the methylation proportions of all type II probes. As a result of these preprocessing preliminary analyses, we had data from 2,325 individuals and 860,079 CpGs.

Additionally, cross-hybridizing probes, sex chromosomes, and single nucleotide polymorphism (SNP) probes with minor allele frequency > 0.05 [96] were removed for analyses. The final number of CpGs for analyses was 790,026. Beta value calculation, which ranges from 0 to 1 and represents the proportion of unconverted cytosines in bisulfite-converted DNA at specific locations, was performed using the R package minfi [94]. M values, which refer to logit 2 transformed beta values and have better properties for statistical analyses (for example, they are more homoscedastic, and their range does not fall between 0 and 1), were also calculated.

Differences in cell type compositions can introduce bias in blood DNA methylation analyses. We estimated cell proportions (CD8 T cells, CD4 T cells, Natural Killer cells [NK], B cells, monocytes, and neutrophils) using the Houseman method [97], which uses regression calibration to estimate proportions of white blood cells from DNA methylation data. We used the R package *FlowSorted.Blood.EPIC* [98], which provides an adaptation of the Houseman method to the 850K microarray. We subsequently used those cell counts as adjustment variables in the regression models, leaving out one of the estimated cell type proportions, as they all add up to 1.

To account for population stratification, all models were addition-

ally adjusted for genetic principal components (PCs) [99]. Of 2,562 genotyped SHS participants as part of the CALiCo/PAGE Study, we identified 644 unrelated individuals (either founders of pedigrees or unrelated spouses of their descendants). Of 162,718 autosomal SNPs that passed quality control, we selected 15,158 based on the following criteria: minor allele frequency ≥ 0.05 (i.e., not rare variants), minimum physical separation of 1 kb and pairwise correlation of genotype scores ≤ 0.1 within a 100 kb sliding window. We performed PC analysis on the genotype scores within unrelated individuals. The first five PCs were kept for adjustment in the models as they explained most of the variance.

We detected and corrected for potential batch effects by sample plate, sample row, and DNA isolation time using the *combat* function (*sva* R package) [100]. This method uses an empirical Bayes framework described by [100] to correct for known batch effects.

We conducted annotation of CpGs to the nearest gene according to the Illumina Infinium MethylationEPIC Manifest File (version 1.0 B4). All this preprocessing resulted in data from 2,325 individuals and 788,368 CpG sites for our analyses.

In the following chapter, we present the state of the art on statistical methods for DNA methylation data analysis, and how our developed methods improve the current gold standards.

CHAPTER 2

Statistical methods for DNA methylation data analysis

Omics data, including DNA methylation, have the particular characteristic of being ultra high-dimensional. At least hundreds of thousands (and even tens of millions) of genomic positions need to be interrogated in order to have an adequate landscape of the whole epigenome. The analysis of ultra-high dimensional data, which further includes substantial between-feature correlations, poses a great computational and statistical challenge. Specifically, traditional data analysis methods in epidemiologic studies, such as linear regression for continuous endpoints or Cox proportional hazards regression for survival analysis, cannot accommodate large numbers of predictors at a time, especially in cases of multicollinearity when introducing highly correlated variables, which can lead to inflated standard errors, thus making the corresponding point estimates uncertain [101]. Therefore, the task of analyzing the data to look for patterns, associations and to potentially construct clinically useful scores and algorithms, implies the development of advanced statistical methods. Many efforts have been posed in the past few years to develop efficient statistical methods for the analysis of these data. In this section, we first describe the traditional approaches for epigenomic data analysis, which have effec-

tively set the bases for the development of more statistically efficient tools for these purposes.

2.1 One marker at a time approach

Traditionally, omics data analysis has been conducted evaluating genomic positions in separate regression models. The first and more obvious problem that arises from this approach is the multiple comparisons issue. The probability of identifying false positives increases with the number of statistical tests conducted, which, in the case of omics data, might be hundreds of thousands. Thus, p-values need to be corrected to make sure we are not identifying a high proportion of false positives among our statistically significant results. Bonferroni, family-wise error rate (FWER) and false discovery rate (FDR) are common methods to correct for multiple comparisons. The FDR refers to the rate that features called significant are truly null. For example, an FDR of 5 % (common threshold of significance) would mean that, among all features called significant, 5 % of these would be truly null. The FDR method has uniformly higher power, in terms of probability of rejecting the null hypothesis when the alternative is true, as compared to Bonferroni and FWER methods [102]. Thus, this method is lately preferred by researchers to account for multiple comparisons in EWAS.

The *limma* R package [103] has been considered as the standard for EWAS for several years. The limma *lmFit* function conducts linear regression for each CpG site individually, and then uses a quasi-Bayesian algorithm to shrink the standard errors towards a common value and gain robustness and stability of the test statistics. This tool fits the same statistical model to each available genomic position or gene, and ranks the features by evidence against the null hypothesis. This approach computes posterior values that shrink the observed variances towards the prior values that describe how the unknown coefficients and variances vary across features. Thus, moderated t-statistics that borrow information accross features are calculated, leading to more stable inference [104].

Another strength of the limma algorithm is that it is extremely fast in terms of computational efficiency. However, it has the huge limitation that the algorithm expects DNA methylation to be introduced as the outcome, as it conducts one regression per column of the multidimensional matrix that is introduced as the outcome. Thus, it is not appropriate for settings in which we want to evaluate the effect of DNA methylation on a clinical outcome (i.e. considering DNA methylation as the predictor). In addition, in omics data, and in particular, in epigenetics, it is well known that the potential effects of DNA methylation dysregulations on disease for each individual CpG are unlikely to be independent [105]. In fact, the identification of differentially methylated regions on the epigenome [77], the observed high correlations between CpGs [106] and the existence of complex regulatory networks in the genome [107], support that jointly studying all CpG sites (i.e., "multiple markers at a time") is a more informative approach. Even though efforts have been made in the limma algorithm to incoporate the common feature-wide structure of Genome-Wide Association Studies (GWAS) and EWAS, this approach still considers each feature in a separate regression model. To overcome this limitation, shrinkage methods have become a popular choice to approach the "all markers at a time" method in omics data analysis. We hereby describe the main frequentist and Bayesian shrinkage methods that have been developed in recent years.

2.2 Multiple markers at a time approach: frequentist shrinkage methods

The bias-variance trade-off [108] establishes that the variance of the parameters estimated accross samples can be reduced by increasing the bias in the estimated parameters. Shrinkage or regularization methods, such as LASSO or Ridge regression, decrease standard errors at the cost of introducing some bias in the simultaneously estimated effects. Thus, the first versions of these methods were considered very efficient for variable selection, while less efficient for effect estimation. Subsequent efforts have improved the effect estimation component of

shrinkage methods, as we will describe in this section. These tools have become popular approaches for variable selection in multi-dimensional DNA methylation data [109] and genome-wide SNPs analyses [110, 7]. We hereby describe the most widely used frequentist shrinkage methods.

2.2.1 Ridge Regression

Ridge regression, presented by Hoerl and Kennard in 1970 [111], was the first proposed shrinkage method. As all shrinkage methods, Ridge regression introduces bias with the aim of decreasing the mean squared errors (MSE-s). However, it does not conduct variable selection, i.e., it does not lead to a sparse solution. The Ridge estimator is obtained by solving the L_2 penalized least squares problem:

$$\widehat{\beta}_{Ridge} = \arg\min_{\beta}(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2)$$

where $\| \cdot \|_2$ is the L_2 norm, and $\lambda > 0$ is a tuning parameter that controls the amount of shrinkage. Ridge regression can be performed in R using the *glmnet* package [112].

2.2.2 Least Absolute Shrinkage and Selection Operator: LASSO

The LASSO estimator was the first popularized shrinkage method, and is even nowadays one of the most widely used ones. It was first developed in geophysics applications in 1986 by Fadil Santosa and William W. Symes [113], and was later rediscovered in 1996 by the statistician Robert Tibshirani [114]. LASSO was initially an extension of ordinary least squares, and is obtained by solving the L_1 penalized least squares problem:

$$\widehat{\beta}_{LASSO} = \arg\min_{\beta}(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1)$$

where $\| \cdot \|_1$ is the L_1 norm, and $\| \cdot \|_2$ is the L_2 norm. Because of the nature of the constraint, some of the coefficients will be set to exactly zero, which facilitates variable selection. Although widely used, the LASSO penalty has shown non-ignorable bias on effect estimation, which increases with the increase of the effect estimate in absolute value [115]. In addition, multicollinearity worsens the performance of the LASSO [116]. Moreover, it would not be the most suitable method for omics data given that it tends to select only one variable from a correlated set. In omics data settings, two highly correlated genes might have different biological functions. Thus, shrinkage methods for variable selection in omics data settings would ideally need to select more than one feature from a correlated set. LASSO can be implemented in R using the *glmnet* package [112].

2.2.3 Elastic-net

Elastic-net was proposed by Zou and Hastie [116] as an improvement of the LASSO for high-dimensional settings. It is a combination between Ridge and LASSO regressions, and enables selecting several variables from a correlated set, therefore improving prediction in highly correlated variables settings. However, the effect estimates are also subject to bias. The elastic-net estimator is defined as follows [117]:

$$\widehat{\beta}_{Enet} = \left(1 + \frac{\lambda_2}{n}\right) \left\{ \arg\min_{\beta} \left(\|Y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right) \right\},$$

being n the sample size, and λ_1 and λ_2 regularization parameters. When predictors are standardized to have mean 0 and standard deviation 1 (in most practical settings), $\left(1 + \frac{\lambda_2}{n}\right)$ can be replaced by $(1 + \lambda_2)$. The L_1 part of the elastic-net performs variable selection, whereas the L_2 part stabilizes the solution paths and thus improves predictive accuracy. The *glmnet* R package [112] uses an alternative formulation for the implementation of elastic-net:

$$\widehat{\beta}_{Enet} = \arg\min_{\beta} \left(\|Y - X\beta\|_2^2 + \lambda \left[\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right] \right).$$

According to this formulation, $\lambda > 0$ is a tuning parameter that controls the amount of shrinkage and can be selected via cross-validation. The $\alpha \in [0, 1]$ parameter controls the relative contribution of the L_1 and L_2 parts to the final solution. LASSO corresponds to $\alpha = 1$, whereas Ridge regression corresponds to $\alpha = 0$. Small α choices such as $\alpha = 0.05$, close to Ridge regression, are popular choices for the omics data setting and have shown to work well in terms of variable selection [109]. The reason is that, in the omics data setting, variable selection is generally conducted as a first screening step to reduce the dimensionality and subsequently do further evaluation of the selected markers (such as mediation analysis, for example). Thus, a more inclusive variable screening is generally preferable.

2.2.4 Smoothly Clipped Absolute Deviation (SCAD)

The SCAD [115] is a coupling of the concave convex procedure [118] and the LASSO. The SCAD penalty applies the same penalization rate (and bias) of the LASSO for small effect estimates, but continuously relaxes the rate of penalization as the absolute value of the effect estimate increases. Thus, it presents less bias in effect estimates that are high in absolute value as compared to the LASSO [119]. In addition, it has shown to be consistent in estimation and enjoys the oracle property [119]. This property states that, asymptotically, the model can perform as well in effect estimation as if the components of the true parameter that are restricted to zero were known in advance (see section 3.1.4). However, when the variables are strongly correlated, the performance of the SCAD is worsened. Similar to LASSO, it tends to select only one variable from a correlated set. The SCAD estimator is defined as follows [119]:

$$\widehat{\beta}_{SCAD} = \arg\min_{\beta} \left(\|Y - X\beta\|_2^2 + \sum_{j=1}^{p} p_{\lambda,\gamma}(\beta_j) \right)$$

where p is the dimension of β, and $p_{\lambda,\gamma}(\beta_j)$ is the SCAD penalty such that:

$$p_{\lambda,\gamma}(\beta_j) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda \\ \frac{2\gamma\lambda|\beta_j| - \beta_j^2 - \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < |\beta_j| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } |\beta_j| \geq \gamma\lambda \end{cases} \tag{2.1}$$

for $\lambda > 0$ and $\gamma > 2$. The tuning parameter γ controls the concavity of the penalty, which represents how rapidly the penalty tapers off. SCAD can be implemented in R using the *ncvreg* package [120].

2.2.5 Minimax Concave Penalty (MCP)

Similar to SCAD, MCP is another alternative to get less biased effect estimates for the non-zero coefficients in sparse models. This method also relaxes the penalization rate as the absolute value of the effect estimate increases, but MCP relaxes it immediately, while SCAD mantains the rate flat for a while before decreasing it [121]. MCP also enjoys the oracle property. Introduced by Cun-Hui Zhang [121], it is defined as follows:

$$\widehat{\beta}_{MCP} = \arg\min_{\beta} \left(\|Y - X\beta\|_2^2 + \sum_{j=1}^{p} q_{\lambda,\gamma}(\beta_j) \right)$$

where $q_{\lambda,\gamma}(\beta_j)$ is the MCP penalty such that:

$$q_{\lambda,\gamma}(\beta_j) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2\gamma}, & \text{if } |\beta_j| \leq \gamma\lambda \\ \frac{\gamma\lambda^2}{2}, & \text{if } |\beta_j| > \gamma\lambda \end{cases} \tag{2.2}$$

for $\lambda > 0$ and $\gamma > 1$. MCP can be implemented in R using the *ncvreg* package [120].

2.2.6 Adaptive elastic-net (Aenet)

Zou and Zhang [117] proposed an adaptive version of the elastic-net, in which adaptive weights are used in the L_1 penalty. The improvement with respect to LASSO and elastic-net is twofold. First, similar to SCAD and MCP, it enjoys the oracle property (see section 3.1.4). Second, it handles the multicollinearity issue and is able to select more than one predictor from a correlated set. Adaptive weights are constructed by fitting an elastic-net model to the data:

$$\begin{aligned} \widehat{\beta}_{Enet} &= (1+\tfrac{\lambda_2}{n}) \left\{ \arg\min_\beta \left(\|Y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right) \right\}, \\ \widehat{w_j} &= (|\widehat{\beta}_j(elastic\ net)|)^{-\gamma},\ j = 1, ..., p \end{aligned}$$

with γ being a positive constant (typically set to 1). Then, those weights are applied to the L_1 penalty of the elastic-net:

$$\widehat{\beta}_{Aenet} = (1+\tfrac{\lambda_2}{n}) \left\{ \arg\min_\beta \left(\|Y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \textstyle\sum_{j=1}^{p} \widehat{w}_j \|\beta_j\|_1 \right) \right\} \quad (2.3)$$

Aenet can be implemented in R using the *gcdnet* package [122] for continuous and binary outcomes, and the *Coxnet* [123] package for survival outcomes. To our knowledge, no implementations of Aenet for other outcome families such as Poisson have been conducted.

2.2.7 Multi-step adaptive elastic-net (MSAenet)

This method is an alternative approach presented by Xiao and Xu [124] and developed in the R package *msaenet* [125]. The formulation is similar to that of Aenet, however, instead of applying the adaptive

weights only to the L_1 penalty, it applies the weights to both the L_1 and L_2 penalties in an iterative way (see Xiao and Xu [124]). The oracle property has not been proven for this alternative version of Aenet. The estimation is as follows:

$$\widehat{\beta}_{Msaenet} = (1+\tfrac{\lambda_2}{n})\left\{\arg\min_\beta\left(\|Y-X\beta\|_2^2+\lambda_2\textstyle\sum_{j=1}^p \widehat{w}_j\|\beta_j\|_2^2+\lambda_1\sum_{j=1}^p \widehat{w}_j\|\beta_j\|_1\right)\right\}$$

2.3 Multiple markers at a time approach: Bayesian shrinkage methods

Bayesian shrinkage methods, like frequentist ones, assume sparsity, i.e., assume that only a relatively small number of predictors are related to the outcome. The difference between frequentist and Bayesian shrinkage methods is that Bayesian methods introduce a prior distribution to the regression parameters [126]. Bayesian versions of popular frequentist shrinkage approaches such as LASSO and elastic-net have been developed [127, 128, 129].

Unfortunately, Bayesian shrinkage methods are currently not feasible for direct application to omics data unless prior dimensionality reduction is conducted, due to its intensive computational cost. The currently implemented Bayesian shrinkage methods rely on Markov-chain Montecarlo (MCMC) methods. Markov chains are defined as stochastic processes for which the probability of an event depends only on the state attained in the previous event. Given that MCMC estimations depend on the estimations from the previous iteration, they cannot, in general, be parallelized. This makes these methods unfeasible when dealing with thousands, or even hundreds of thousands of variables, as in the case of omics data. Further research is needed to investigate how the computation of these methods could be sped-up.

Nevertheless, we included two Bayesian penalized methods in our

work as proof of concept and because it is still useful to compare the effect estimation obtained from other shrinkage methods, as well as from traditional methods such as linear regression and Cox proportional hazards models, to those obtained with Bayesian shrinkage methods. We hereby describe two Bayesian shrinkage methods implemented in R: the methodology implemented in the *bayesreg* package [130], which accomodates both continuous and binary outcomes, and the methodology implemented in the *psbcGroup* package [131], which accomodates survival outcomes.

2.3.1 Bayesian linear and logistic penalized models

The *bayesreg* R package [130] implements Bayesian regularized linear and logistic models based on sparsity inducing priors, which are implemented with exchangeable Gaussian variance mixture distributions. Let us consider an outcome y and a covariates matrix $X = (X_1, ..., X_p)$. Then, the Bayesian penalized regression model is set as follows, for the i-th individual:

$$\begin{gathered} z_i|X_i, \beta_0, \beta, \omega_i^2, \sigma^2 \sim \mathcal{N}_n(\beta_0 + X_i^T\beta, \sigma^2\omega_i^2), \\ \sigma^2 \sim \pi(\sigma^2)d\sigma^2, \\ \omega_i^2 \sim \pi(\omega_i^2)d\omega_i^2, \\ \beta_0 \sim d\beta_0, \\ \beta_j|\lambda_j^2, \tau^2, \sigma^2 \sim \mathcal{N}(0, \lambda_j^2\tau^2\sigma^2), \\ \lambda_j^2 \sim \pi(\lambda_j^2)d\lambda_j^2, \\ \tau^2 \sim \pi(\tau^2)d\tau^2, \end{gathered}$$

where $i = 1, ..., n$ corresponds to the individual, $j = 1, ..., p$ corresponds to the covariate, β_0 is the intercept and β are the regression coefficients. Statistical models for both the data and the prior distributions are constructed from exchangeable Gaussian variance mixture distributions [132]. The hyperparameter $\tau^2 > 0$ is the global variance parameter, which controls the amount of overall shrinkage

of the coefficients, and it is given the following prior distribution: $\tau \sim C^+(0,1)$, where C^+ is the half-Cauchy distribution. The hyper-parameters $(\lambda_1, ..., \lambda_n)$ correspond to the local variance components that determine the type of shrinkage penalty applied to the regression coefficients. Several prior distribution choices are available including LASSO, Ridge and Horseshoe, with details available in [130].

For linear regression, the Bayesian penalized regression model is adapted to a Bayesian linear regression model with Gaussian noise. The scale parameter $\sigma^2 > 0$ is given the scale invariant prior distribution $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$. The posterior distribution of σ^2 is given by the inverse Gamma distribution $IG(\alpha, \beta)$, with:

$$\alpha = \frac{n+p}{2},\ \beta = \frac{1}{2}\left(\sum_{i=1}^{n}\frac{e_i^2}{\omega_i^2} + \sum_{j=1}^{p}\frac{\beta_j^2}{\tau^2\lambda_j^2}\right);$$

being e_i the residuals of the linear model. One typical choice for the latent variables $(\omega_1^2, ..., \omega_n^2)$ is to set them to $\omega_i^2 = 1$ (other formulations for the latent variables are considered in [130]). The variables $z_1, ..., z_n$ can be set to $z_i = y_i$.

For binary outcomes (outcomes $y \in \{0, 1\}$), a logistic regression model is assumed:

$$p(y_i = 1 | x_i, \beta_0, \beta) = \frac{1}{1 + exp(-(\beta_0 + x_i^T\beta))},$$

and the variables $z_1, ..., z_n$ are set to $z_i = \omega_i^2\left(y_i - \frac{1}{2}\right)$.

The framework of the *bayesreg* R package models logistic regression using a Gaussian variance mixture distribution with a Pólya-gamma mixing density [133]. In this case, the scale parameter is fixed at $\sigma^2 = 1$. A Pólya-gamma prior distribution is considered for the latent variables $(\omega_1, ..., \omega_n)$:

$$\omega_i^2 \sim PG(0, 1).$$

The posterior distribution of the latent variables $\frac{1}{\omega^2}$ is the Pólya-gamma distribution $PG(1, \tilde{c}_i)$, where

$$\tilde{c}_i = \beta_0 + x_i^T \beta.$$

For both linear and logistic models, the point estimate is calculated as the median of the posterior distribution of β, whereas 95 % credibility intervals are calculated as the 2.5 and 97.5 percentiles of the posterior distribution.

2.3.2 Bayesian Cox penalized model

The *psbcGroup* R package [131] fits penalized semiparametric Bayesian Cox regression with an elastic-net prior. A Laplace prior is used for the regression coefficients as detailed in [131]:

$$\pi(\beta|\sigma^2) \propto exp\left(-\frac{\lambda_1}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda_2}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2\right)$$

To guarantee unimodality [131], a noninformative marginal prior $\pi(\sigma^2) = \frac{1}{\sigma^2}$ is assigned to σ^2. The conventional random walk Metropolis Hastings algorithm is used to update the parameters. The point estimates are calculated as the median of the posterior distribution of β, whereas 95 % credibility intervals are calculated as the 2.5 and 97.5 percentiles of the posterior distribution.

Different shrinkage methods might be adequate depending on whether the purpose is accurate effect estimation, high predictive ability, sparseness or computational efficiency. However, the performance of these tools might be especially worsened in ultra-high dimensional settings [8], meaning those in which, being n the sample size, the order of the number of predictors is $exp\{O(n^{\xi})\}$, for a given $\xi > 0$. Thus, Fan and Lv [8] proposed to apply an effective variable selection method in ultra-high dimensional settings, before applying shrinkage methods to the data: ISIS, which is described in section 3.

Once dimesionality has been reduced, the adequate set of features has been selected and associations with the outcome have been established, subsequent statistical analyses of interest regarding associations between features and the outcome might be conducted. One of them is mediation analysis, which aims to evaluate the potential intermediate role of a third variable on the association between two variables.

2.4 Evaluating the potential intermediate role of DNA methylation in environment-related disease: mediation analysis

The understanding of causal pathways underlying the association between an exposure or treatment and an outcome is a question of interest in epidemiologic research. Mediation analysis aims to quantify to which extent the relationship between two variables happens through a third variable called the mediator (indirect effect), and to which extent it happens through other not considered pathways (direct effect). Extensive literature, as well as many analytic tools, exist for the evaluation of simple mediation analysis [134, 135].

The counterfactual framework has been widely used in causal inference, including in mediation analysis [136]. Let us denote E as an exposure or treatment and Y as the outcome of interest. The counterfactual outcomes refer to the values Y would take under each of the potential values of E. Please note that some of those values of Y will be unobservable, which is the reason why they are called counterfactual (contrary-to-fact). For example, if the exposure E is dichotomous (exposed / unexposed), an individual will either be exposed or unexposed, thus, one of the counterfactual outcomes will not be observed.

For causal effects to be identified in observational studies, three conditions need to hold. The first condition is called consistency and ensures there are no multiple versions of the treatment or the exposure. The second condition is called exchangeability and refers to no unmeasured confounders. The last condition is called positivity, and refers to the probability of having exposed and unexposed (or treated and untreated) individuals in each strata of the covariates being greater than zero [136].

Focusing on causal mediation analysis, let us denote M as the mediator, which is dependent on the exposure E; X as a set of covariates and Y as the outcome of interest. Let us consider two different values of the exposure, e and e^*. Following the counterfactual framework [137], we consider $Y(e^*, M(e))$ as the counterfactual outcome, i.e., the

value the outcome would take had the exposure been set to e^* and the mediator been set to the value it would take when the exposure is set to e. For most of the purposes of this work, e will be a dichotomous variable (presence or absence of exposure), thus, $(e, e^*) \in \{0, 1\}^2$. We define the average indirect effect of changing the exposure from e^* to e when the covariates are set to $X = x$ as follows [138, 9, 139]:

$$\delta(e) = \mathbb{E}\left[Y(e^*, M(e))|X = x\right] - \mathbb{E}\left[Y(e^*, M(e^*))|X = x\right].$$

Similarly, the average direct effect, which refers to the effect of the exposure or treatment on the outcome which does not happen through the mediator, is quantified as:

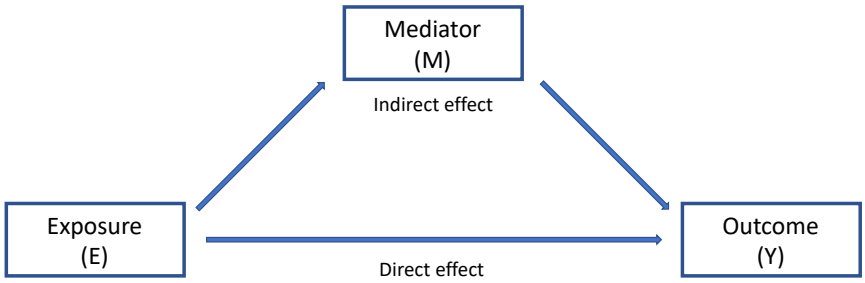
$$\zeta(e) = \mathbb{E}\left[Y(e, M(e))|X = x\right] - \mathbb{E}\left[Y(e^*, M(e))|X = x\right].$$

Please note that there is an indirect and a direct effect for each e. Last, the average total effect, which denotes the effect of the exposure or treatment on the outcome both through the mediator pathway and through other pathways, is quantified as:

$$\tau(e) = \mathbb{E}\left[Y(e, M(e))|X = x\right] - \mathbb{E}\left[Y(e^*, M(e^*))|X = x\right].$$

Please also note that, following these definitions, it holds that $\tau(e) = \zeta(e) + \delta(e)$, showing that the indirect and direct effects represent an exact decomposition of the total effect. Figure 2.1 depicts the framework of mediation analysis.

Figure 2.1: Overview of mediation analysis.



The total effect is the sum of the direct and the indirect effects.

The Sequential Ignorability Assumptions [140], related to no unmeasured confounding in the exposure-outcome, exposure-mediator and mediator-outcome relationship, are needed for these effects to be identified. In addition, the previously described causal inference assumptions of positivity and consistency [141] need to hold.

Given that survival outcomes are widely used in epidemiological research, Lange and Hansen [139] extended the general mediation framework to time-to-event outcomes, thus setting the bases for conducting simple mediation analysis in survival settings.

Simple mediation analysis in survival settings: additive risks model

Cox proportional hazards models are the most widely used in survival analysis. However, the coefficients of this model represent the log hazard-ratio, typically exponentiated to obtain the hazard ratio, which is a non-collapsible measure [142, 143, 144]. Non-collapsibility implies discrepancy between conditional and marginal effects even on the absence of confounding [142]. For non-collapsible measures, conditioning on a covariate that is related to the outcome would change the coefficient of the exposure, even if the covariate is unrelated to the

exposure. Conversely, measures from additive models are collapsible. These models quantify the effects on an additive scale (as rate differences), thus providing a more interpretable measure of impact that can be highly informative for public health [145]. For this reason, additive hazards models have been widely used in mediation analysis instead of Cox proportional hazards models.

The Aalen additive hazards model [146] assumes that the hazard function (or the rate) for the failure time t, dependent on an exposure E, a mediator M and a covariates matrix X, takes the form:

$$\gamma(t; E, M, X) = \lambda_0(t) + \lambda_1(t)E(t) + \lambda_2(t)^T X(t) + \lambda_3(t)M(t),$$

being λ_0 the baseline hazard. Lin and Ying [147] developed the semi-parametric additive risks model, in which the same form of the hazard function is assumed, but the covariates and coefficients can have either time-varying or constant effects. For this work, we will focus on time-invariant covariates and coefficients for simplicity, therefore using the Lin-Ying additive risks model. Only the baseline hazard λ_0 is dependent on time, and the hazard function would then be:

$$\gamma(t; E, M, X) = \lambda_0(t) + \lambda_1 E + \lambda_2^T X + \lambda_3 M$$

Effect definition

Lange and Hansen [139] proposed to define the direct, indirect and total effects in a survival context for a single mediator using an additive hazards model as the outcome model, in which the total effect of an exposure on an outcome is expressed as differences in rate. Let us assume that the mediator is continuous and use a linear model for the mediator model. Thus, being E the exposure, M the mediator and X a vector of p covariates, the mediator and outcome models are defined as follows:

$$\begin{cases} M(E,X) = \alpha_0 + \alpha_1 E + \alpha_2^T X + \epsilon \\ \gamma(t;E,X,M) = \lambda_0(t) + \lambda_1 E + \lambda_2^T X + \lambda_3 M \end{cases}$$

where $\alpha_0, \alpha_1, \lambda_1, \lambda_3 \in \mathbb{R}$; $\alpha_2, \lambda_2 \in \mathbb{R}^p$; $\lambda_0(t)$ is the time-varying baseline hazard and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the error in the linear model, with variance σ^2. The first equation is called the mediator model, whereas the second one is referred to as the outcome model.

In survival settings, the mediated effect, or indirect effect of changing the exposure from e^* to e, is quantified as:

$$\delta(e) = \gamma(t; e^*, M(e), X) - \gamma(t; e^*, M(e^*), X).$$

The direct effect is quantified as:

$$\zeta(e) = \gamma(t; e, M(e), X) - \gamma(t; e^*, M(e), X).$$

Last, the total effect is quantified as:

$$\tau(e) = \zeta(e) + \delta(e) = \gamma(t; e, M(e), X) - \gamma(t; e^*, M(e^*), X).$$

Please note that, here, the effects are defined as differences in hazard functions instead of differences of averages. In general, $\delta(e)$, $\zeta(e)$ and $\tau(e)$ are functions of t.

Sequential Ignorability Assumptions in survival analysis

We define $T(e, m)$ as the time to event when the exposure is set to e and the mediator is set to m. The following assumptions need to hold for the direct, indirect and total effects to be identifiable:

- H.1. First exchangeability assumption: No unmeasured confounding of the exposure-outcome relationship: $E \perp T(e, m) \mid X$.

- H.2. Second exchangeability assumption: No unmeasured confounding of the mediator-outcome relationship : $M \perp T(e,m) \mid X, E$.
- H.3. Third exchangeability assumption: No unmeasured confounding of the exposure-mediator relationship: $E \perp M(e) \mid X$.
- H.4. Consistency: $M(e) = M,\ T(e,m) = T$.
- H.5. Identifiability: $M(e^*) \perp T(e,m) \mid X$.

In *Theorem 1* of Lange and Hansen [139], it was proven that, under sequential ignorability assumptions, the total effect measured in the rate difference scale at time t is:

$$\tau(e) = \gamma(t; e, M(e), X) - \gamma(t; e^*, M(e^*), X) = \lambda_1(t)(e - e^*) + \lambda_3(t)\alpha_1(e - e^*),$$

where $\lambda_1(t)(e - e^*)$ is the direct effect of the exposure on the outcome, and $\lambda_3(t)\alpha_1(e - e^*)$ is the indirect efect through the mediator. The proof of this result can be found in the Appendix of Lange and Hansen [139]. Please note that, if λ_1 and λ_3 are time-independent, the three effects wil also be time-independent.

An application of simple mediation analysis to evaluate the potential intermediate role of DNA methylation on the association between arsenic and CVD is detailed in section 3.2.2. An extension to multiple mediation analysis for correlated mediators in survival settings is described in chapter 4.

CHAPTER 3

Variable selection in the omics data setting

Several challenges need to be faced when implementing variable selection methods in the omics data setting. In ultra-high dimensional settings, spurious correlations between some unimportant predictors (predictors that are not associated with the outcome) and the outcome can happen due to the fact that those unimportant predictors can be highly correlated with some predictors that are related to the outcome [8]. Other problems include the growing computational cost and the fact that the population covariance matrix can become ill conditioned (i.e. can have big changes when small changes happen among the predictor variables) when the sample size grows [8]. These drawbacks make it challenging to estimate the coefficients of the sparse parameter vector in ultra-high dimensional settings. To overcome these limitations, Fan and Lv proposed the Sure Independence Screening (SIS) tool and its variants, which are paired with a shrinkage method chosen by the user [8].

In this section, we describe this tool in detail, including its implementation in the *SIS* R package and our contributions to the extension of this package. The original SIS statistical framework has been published in [148], as well as its variants [149]. Our contributions include the incorporation of elastic-net, Aenet and MSAenet as

shrinkage methods to be paired with ISIS, as well as a bootstrap CI approach to quantify uncertainty of the estimated effects. We subsequently describe two applications of this tool to the omics data setting, particularly, using DNA methylation data. The first application is focused on the comparison of the performance of different regularization methods paired with ISIS. The second application uses the SIS tool to approach the problem of quantifying the association of arsenic exposure with DNA methylation and CVD.

3.1 Sure Independence Screening and its variants

The SIS method was proposed to reduce dimensionality from ultra-high to that below the sample size [8]. This method is based on component-wise regression (i.e., correlation learning), such that it ranks the importance of features according to their marginal correlation with the outcome and discards those variables that have weak marginal correlations with the outcome. This algorithm is implemented in the *SIS* R package for gaussian, dichotomous, time-to-event, Poisson and multinomial outcomes. Component-wise regression is conducted differently if the outcome follows a Gaussian, binomial, multinomial or Poisson distribution than if it is a time-to-event outcome. As detailed in [148], under a Gaussian outcome, we would consider the following model:

$$y = X\beta + \epsilon, \tag{3.1}$$

being $y = (Y_1, ..., Y_n)^T$ a vector of responses, $X = (x_1, ..., x_n)^T$ the $n \times p$ design matrix, $\beta = (\beta_0, ..., \beta_p)^T$ a vector of parameters and $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ a vector of independent identically distributed random errors. The maximum likelihood estimator is defined as:

$$\widehat{\beta}_{LS} = (X^T X)^{-1} X^T y.$$

For generalized linear models (including, in this case, binomial, multinomial and Poisson models), let us consider that we have observations $\{(x_i, y_i) : i = 1, ..., n\}$ from the population (x, y), where $x = (x_0, ..., x_p)^T$ is a $(p + 1)$-dimensional predictor vector, and y is the outcome. We assume that the distribution of y given x is from an exponential family taking the canonical form:

$$f(y, x, \beta) = exp\{yx^T\beta - b(x^T\beta) + c(y)\},$$

being $b(\cdot)$ and $c(\cdot)$ known functions, and β a $(p+1)$-dimensional regression coefficient vector. The maximum marginal likelihood estimator

(MMLE) $\widehat{\beta}$ is defined as:

$$\widehat{\beta}_{MMLE} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} \{y_i x_i^T \beta - b(x_i^T \beta)\}.$$

For time-to-event outcomes, we denote $\delta \in \{0, 1\}$ as the failure indicator, y as the time-to-event and x as the p-dimensional predictor vector such that $x = (x_1, ..., x_p)^T$. We additionally denote $R(t_j)$ as the risk set prior to time t_j: $R(t_j) = \{i : y_i \geq t_j\}$. Then, in order to obtain the MMLE for β, the Cox proportional hazards model uses the maximization of the partial log-likelihood:

$$\widehat{\beta}_{MMLE} = \underset{\beta}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \delta_i x_i^T \beta - \sum_{i=1}^{n} \delta_i \; log\Big\{ \sum_{k \in R(y_i)} exp(x_k^T \beta) \Big\} \right).$$

Let us denote M_* as the true model, and M_γ as the model selected in the variable selection process. The SIS method enjoys the sure screening property, which refers to all the important variables being selected with probability tending to 1 (i.e. $P(M_\gamma \subset M_*) \to 1$ as $n \to \infty$) under certain regularity conditions that we hereby describe.

3.1.1 Sure screening assumptions

We assume the linear model in (3.1). We define the standardized predictor vector $z = \Sigma^{-1/2} x$, being x a vector of p covariates and $\Sigma = cov(x)$. Note that z has covariance matrix I_p. Being X the design matrix, we also define the transformed design matrix $Z = X\Sigma^{-1/2}$. Please note that, provided the covariates are normally distributed, the n rows of Z are independent identically distributed copies of z, being n the number of observations.

Being λ_{max} and λ_{min} the largest and smallest eigenvalues of a matrix, respectively, a matrix is said to fulfill the concentration property

if there exist some $c, c_1 > 1$ and $C_1 > 0$ such that the deviation inequality:

$$P\{\lambda_{max}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) > c_1 \text{ or } \lambda_{min}(\tilde{p}^{-1}\tilde{Z}\tilde{Z}^T) < \tfrac{1}{c_1}\} \leq exp(-C_1 n)$$

holds for any $n \times \tilde{p}$ submatrix $\tilde{Z}$ of Z with $cn < \tilde{p} \leq p$. This property, which we call C property for simplicity hereinafter, intuitively means that, with high probability, the n non-zero singular values of the $n \times \tilde{p}$ matrix $\tilde{Z}$ are of the same order [8].

In order for the sure screening property to be fulfilled, these four conditions need to hold:

1. $p > n$ and $log(p) = O(n^{\xi})$, for some $\xi \in (0, 1 - 2\kappa)$, where κ is given by condition 3.
2. z has a spherically symmetric distribution, i.e., it is invariant with respect to rotations, and the Z matrix has property C. In addition, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.
3. $var(Y) = O(1)$, which implies that the variance does not grow with sample size, and, for some $\kappa \geq 0$ and $c_2, c_3 > 0$; $\min_{i \in \mathcal{M}^*} |\beta_i| \geq \frac{c_2}{n^{\kappa}}$ and $\min_{i \in \mathcal{M}^*} |cov(\beta_i^{-1}Y, X_i)| \geq c_3$. This condition would rule out the situation in which important variables are marginally uncorrelated, but jointly correlated with the outcome. The iterative variant of SIS, explained in Algorithm 1, aims to take this situation into account.
4. There exist some $\tau \geq 0$ and $c_4 > 0$ such that $\lambda_{max}(\Sigma) \leq c_4 n^{\tau}$. This condition would rule out the case of strong collinearity, which is, again, to be taken into account by the iterative variant of SIS.

Deeper explanations of conditions 1-4 have been published [8]. Under conditions 1-4, the SIS tool has been proven to fulfill the sure screening property. However, please note that these assumptions refer to the case in which the outcome is a linear model with Gaussian covariates, as if covariates do not follow a normal distribution, the

standardization does not guarantee independence between covariates, it would only guarantee they are uncorrelated. This would limit its use in practice to the cases in which there are no categorical variables in the design matrix, which is highly unlikely in most settings. Thus, Fan and Song [150] extended the SIS tool to a generalized linear model framework based on marginal maximum likelihood estimation. In addition, they extended the previously described sure independence conditions to the broader framework of generalized linear models and any kind of covariates.

On the other hand, in situations in which important variables are marginally uncorrelated, but jointly correlated with the outcome, or in situations of high correlations between predictors, these assumptions would not hold. In order to overcome this, an iterative variant was proposed by Fan and Lv as an extension of the SIS algorithm: ISIS [8]. Because this situation is quite common in the setting of omics data, this work is exclusively focused on ISIS, which is described in the following section, rather than on regular SIS.

3.1.2 Iterative Sure Independence Screening

ISIS goes one step further and evaluates the additional contribution of variables that have not initially been selected by the SIS algorithm. To do so, it conducts a multivariable regression considering all selected variables and each of the non-selected variables. The workflow of the ISIS algorithm is described in Algorithm 1.

Algorithm 1 Iterative Sure Independence Screening

Input: Response vector Y, design matrix X, pre-specified maximum number of variables to select a_1, maximum number of iterations k.

Output: Selected variables and coefficients.

1: Select the set $\widehat{M}_1$ using component-wise regression as the set of the a_1 largest marginal regression coefficients in absolute value: $\widehat{M}_1 = \{1 \leq i \leq p : abs(\widehat{\beta}_i)\ is\ among\ the\ a_1\ largest\}$

2: Obtain the subset $\widehat{S}_1$ by applying penalized regression to the set $\widehat{M}_1$.

3: **for** $i \leftarrow 2$ *to* k **do**

4: $\quad \widehat{M}_i = \{j \in \widehat{M}_{i-1}^c :\ abs(\widehat{\beta}_j)\ is\ among\ the\ a_1\ largest\}$

5: $\quad$ Obtain $\widehat{S}_i$ by applying penalized regression to the set $\widehat{S}_{i-1} \cup \widehat{M}_i$.

6: $\quad$ **if** $\left|\widehat{S}_i\right| = a_1$ **or** $\widehat{M}_i = \widehat{M}_s$ being $s < i$ **then**

7: $\quad\quad$ **break**

8: $\quad$ **end if**

9: **end for**

Step 4 in Algorithm 1 is conducted to cover the situation in which some variables might be marginally uncorrelated but jointly correlated with the outcome. To avoid leaving those variables outside the selected set, the additional contribution of the variables that have not been selected is evaluated.

The maximum number of variables to select, a_1, can be user-specified. The asymptotic theory [8] shows that, in the linear model, there exists some $\theta > 0$ with which the sure screening property should be obtained with $\left\lfloor n^{1-\theta} \right\rfloor < a_1 < n$. However, θ is unknown in practice. Thus, Fan, Samworth and Wu [149] recommended $a_1 = \left\lfloor \frac{n}{log(n)} \right\rfloor$, which led to good numerical results. Although choosing larger values of a_1 increases the probability of selecting all the important variables, including unimportant variables in the selected set tends to have a

detrimental effect on the performance of the effect estimation [149]. This detrimental effect is more evident in models with less informative responses than the real-valued Gaussian outcome. For this reason, the ISIS algorithm sets default a_1 values to $\left\lfloor \frac{n}{log(n)} \right\rfloor$ for Gaussian outcomes, to $\left\lfloor \frac{n}{2\ log(n)} \right\rfloor$ for Poisson outcomes and to $\left\lfloor \frac{n}{4\ log(n)} \right\rfloor$ for time-to-event and binary outcomes.

3.1.3 Variants of Iterative Sure Independence Screening

In order to reduce false selection rates of the ISIS algorithm, two variants of ISIS based on sample splitting have been proposed [149]. The first variant is called conservative variant. The reason of using sample splitting is that an unimportant predictor would need to be selected in both sets in order to be selected in the overall algorithm, which minimizes the probability of false selection. Algorithm 2 describes the workflow of the conservative ISIS variant.

Algorithm 2 Conservative Variant of Iterative Sure Independence Screening

Input: Response vector Y, design matrix X, pre-specified maximum number of variables to select a_1, maximum number of iterations k.

Output: Selected variables and coefficients.

1: Y and X are randomly split into two parts. Regular ISIS is applied to each of them. The number of features selected from $\widehat{M}_1^1$ and $\widehat{M}_1^2$ is the smallest number by which we can ensure that $\widehat{M}_1^1 \cap \widehat{M}_1^2$ has at least $\frac{2}{3}a_1$ features.
2: Obtain the subset $\widehat{S}_1$ by applying penalized regression to the set $\widehat{M}_1^1 \cap \widehat{M}_1^2$.
3: **for** $i \leftarrow 2$ *to* k **do**
4: $\widehat{M}_i^1 = \{j \in (\widehat{M}_{i-1}^1)^c : \ abs(\widehat{\beta}_j)$ *is among the* a_1 *largest*$\}$ and
5: $\widehat{M}_i^2 = \{j \in (\widehat{M}_{i-1}^2)^c : \ abs(\widehat{\beta}_j)$ *is among the* a_1 *largest*$\}$ ensuring that $\widehat{M}_i^1 \cap \widehat{M}_i^2$ has at least $a_1 - |\widehat{S}_{i-1}|$ features.
6: Obtain $\widehat{S}_i$ by applying penalized regression to the set $\widehat{S}_{i-1} \cup (\widehat{M}_i^1 \cap \widehat{M}_i^2)$.
7: **if** $\left|\widehat{S}_i\right| = a_1$ **or** $\widehat{M}_i^1 = \widehat{M}_s^1$ **or** $\widehat{M}_i^2 = \widehat{M}_s^2$ being $s < i$ **then**
8: **break**
9: **end if**
10: **end for**

As first estimations of the true sparse model, both sets $\widehat{M}_1^1$ and $\widehat{M}_1^2$ defined in Step 1 of Algorithm 2 may have large false discovery rates (FDR). However, under certain conditions [149], both sets should contain the most important features, therefore, the FDR is reduced as the most important features have been selected twice in different sets. In Step 1, the sets $\widehat{M}_1^1$ and $\widehat{M}_1^2$ are selected in a way that we ensure that the subset $\widehat{M}_1^1 \cap \widehat{M}_1^2$ has at least $\frac{2}{3}a_1$ predictors. Thus, this variant warrants that at least $\frac{2}{3}a_1$ predictors will be included in the penalized regression. This procedure is repeated in all iterations.

The second variant, called aggressive variant, is also based on sample splitting. However, it does not specify a minimum set size for $\widehat{M}_i^1$

and $\widehat{M}_i^2$. Algorithm 3 describes the workflow of the aggresive ISIS variant.

Algorithm 3 Aggresive Variant of Iterative Sure Independence Screening

Input: Response vector Y, design matrix X, pre-specified maximum number of variables to select a_1, maximum number of iterations k.

Output: Selected variables and coefficients.

1: Y and X are randomly split into two parts. Regular ISIS is applied to each of them, obtaining sets $\widehat{M}_1^1$ and $\widehat{M}_1^2$.
2: Obtain the subset $\widehat{S}_1$ by applying penalized regression to the set $\widehat{M}_1^1 \cap \widehat{M}_1^2$.
3: **for** $i \leftarrow 2$ *to* k **do**
4: $\quad \widehat{M}_i^1 = \{j \in {\widehat{M}_{i-1}^1}^c : \ abs(\widehat{\beta}_j)\ is\ among\ the\ a_1\ largest\}$
5: $\quad \widehat{M}_i^2 = \{j \in {\widehat{M}_{i-1}^2}^c : \ abs(\widehat{\beta}_j)\ is\ among\ the\ a_1\ largest\}$
6: $\quad$ Obtain $\widehat{S}_i$ by applying penalized regression to the set $\widehat{S}_{i-1} \cup (\widehat{M}_i^1 \cap \widehat{M}_i^2)$.
7: $\quad$ **if** $\left|\widehat{S}_i\right| = a_1$ **or** $\widehat{M}_i^1 = \widehat{M}_s^1$ **or** $\widehat{M}_i^2 = \widehat{M}_s^2$ being $s < i$ **then**
8: $\quad\quad$ **break**
9: $\quad$ **end if**
10: **end for**

The aggressive variant is more computationally efficient, but it might lead to undesirably small model sizes, which, as previously mentioned, are not desirable in an omics data setting. This variant does not guarantee that the intersection of the selected variables in each of the sets has a certain number of variables, unlike the conservative ISIS variant described in Algorithm 2. Therefore, this ISIS variant often leads to no variables being selected. For this reason and for simplicity, we will hereinafter focus on the conservative ISIS approach.

Once ISIS completes the screening process and the dimension of the variable vector is reduced, a shrinkage method is applied to the

reduced dataset. LASSO, SCAD and MCP (see section 2.2) were initially implemented in the *SIS* R package. In this PhD dissertation, we further implemented elastic-net and two versions of Aenet, which were not previously implemented in the package and, as described in detail in section 2.2, offer improved results in terms of variable selection and effect estimation.

The SIS algorithm performs effect estimation in addition to variable selection. Since we would like our algorithm to be consistent in both variable selection and parameter estimation, we hereby introduce the oracle property: a property that ensures this consistency.

3.1.4 The oracle property

An oracle-like estimator is a consistent estimator in both parameter estimation and variable selection. In 2009, Zou and Zhang [117] demonstrated that, under certain regularity conditions, Aenet fulfills the oracle property. Let Ω be the true sparse model, X_Ω the true design matrix and β^*_Ω its vector of coefficients, in which some components are exactly zero. Let us consider the Aenet estimator (see section 2.3). This estimator satisfies the oracle property, which ensures these two conditions:

- **Consistency in selection:** As $n \to \infty$, the selected variables are the ones included in the true sparse model, or $P(\{j : \widehat{\beta}_j \neq 0\} = \Omega) \to 1$
- **Asymptotic normality:**

$$\frac{\alpha^T}{1 + \frac{\lambda_2}{n}}(I + \lambda_2 \Sigma_\Omega^{-1})\Sigma_\Omega^{1/2}(\widehat{\beta}_{Aenet} - \beta^*_\Omega) \xrightarrow{d} N(0, \sigma^2)$$

where α is a vector of norm 1, $\xrightarrow{d}$ means convergency in distribution and $\Sigma_\Omega = X_\Omega^T X_\Omega$. From this formula, we can conclude that, asymptotically, the estimated coefficients are an unbiased estimation of the true coefficients.

The oracle property has also been proven for SCAD [119] and MCP [151]. However, neither LASSO nor elastic-net estimators fulfill the oracle property [152]. This property has not been proven to date for MSAenet either.

Once we have conducted effect estimation, we would also like to quantify the statistical uncertainty around those estimates. We extended the ISIS algorithm to incorporate a bootstrap-based confidence interval approach in order to quantify that uncertainty. Our approach is described in the following subsection.

3.1.5 Extension of the SIS R package: elastic-net, adaptive elastic-net and bootstrap confidence intervals

In this work, we extended the SIS R package to pair its algorithm with three previously described shrinkage methods that were not considered in the initial version of SIS: elastic-net, Aenet and MSAenet. Given that these methods have shown smaller errors and better predictive ability [117, 116], we hypothesized they would improve variable selection and effect estimation of the SIS tool. The inclusion of these new shrinkage methods is performed in the second step of SIS, when a shrinkage method is applied to the preselected features set by regular SIS or ISIS.

For elastic-net, we used the *cv.glmnet* function of the *glmnet* R package [112] to conduct cross-validation and select the optimal λ to obtain the most parsimonious model within 1 standard error from the value that minimizes the mean cross-validated error. We set the α parameter to 0.05 as SIS is already very restrictive on the number of variables selected, and we do not want to obtain undesirably minimal sizes of selected feature sets. The number of folds for cross-validation is automatically set to 10, but can be changed by the user.

For Aenet and MSAenet, we first fit an elastic-net model with the same specifications detailed on the above paragraph. We choose the appropriate λ and the set of coefficients obtained from that model.

For Aenet, two different R packages are internally called depending on whether the outcome is linear or binary, or time-to-event. If the outcome is linear or binary, the *gcdnet* R package [122] is used. If the outcome is time-to-event, the *Coxnet* package [123] is used. For MSAenet, the R package *msaenet* [125] is used for all outcomes. The type of penalty for the initial step of MSAenet is fixed to Ridge regression. For both Aenet and MSAenet, the coefficients previously calculated by elastic-net are considered as weights.

These three shrinkage methods were added to the *tune.fit* function of the *SIS* R package, which is available in the CRAN repository, as well as in the Github repository *https://github.com/yangfengstat/SIS/*.

In addition, given the need of quantification of uncertainty, we included a quantile bootstrap-based approach to calculate CIs of the obtained effect estimates. The non-parametric bootstrapping is a statistical approach that relies on resampling a dataset many times with replacement [153]. This tool has been previously used to calculate CIs for effect estimates obtained from penalized regression methods [154, 155]. In our method, bootstrapping is applied to the variable set selected by the ISIS method paired with penalized regression.

The quantile bootstrap approach uses the $\frac{\alpha}{2}$-th value of the bootstrap distribution of β_j as the lower bound CI of the j-th predictor, and the $\left(1-\frac{\alpha}{2}\right)$-th value as the upper bound CI. We set $\alpha = 0.05$ to obtain 95 % CIs. In order to ensure that the size of the bootstrap estimator is adequate, we first test 200 bootstrap samples, repeat the approach 10 times (leading to 2000 bootstrap estimates) and save the lower and upper CIs. We subsequently check if the standard error of the mean of those 10 estimations for both upper and lower CIs is lower than the 5 % of the mean length of the interval for each variable, i.e.:

$$\begin{cases} \frac{sd(LCI_i)}{\sqrt{10}} < 0.05 * mean(|LCI_i - UCI_i|), \ i = 1, ..., p \\ \frac{sd(UCI_i)}{\sqrt{10}} < 0.05 * mean(|LCI_i - UCI_i|), \ i = 1, ..., p \end{cases} \tag{3.2}$$

being LCI_i the 10 estimations of the lower CI for the i-th variable, and

UCI_i the 10 estimations of the upper CI for the i-th variable. If this condition is met for more than 95 % of the variables, the bootstrap sample is adequate and we use the constructed sample to calculate upper and lower CIs. If the condition is not met for more than 5 % of the variables, then the variability is too high and we need to increase the number of bootstrap repetitions. Thus, the process of testing 200 bootstrap samples 10 times is repeated and added to the previous bootstrap estimates, leading to 4000 bootstrap estimates. This process is repeated until the condition is met for more than 95 % of the variables. Bootstrap CIs were implemented in the *boot.sis* function of the *SIS* R package. Their calculation is only conducted if the value of the *boot_ci* parameter in the initial SIS call is set to *TRUE*.

In the following section, we illustrate two applications of our extension of SIS to population-based data from the SHS.

3.2 Data applications

In this section, we describe two applications of the ISIS algorithm to DNA methylation data. The first one is methodological and compares the performance of different shrinkage methods paired with ISIS, while also comparing them to traditional regression approaches and Bayesian shrinkage methods. The second one is epidemiology oriented, and applies the ISIS tool to a real problem of evaluating the effect of arsenic exposure on DNA methylation and CVD. Both applications use data from the SHS (see section 1.5).

3.2.1 Data Application 1: Comparison of regularization methods for the evaluation of blood DNA methylation as a marker of health endpoints

Given that different shrinkage methods might outperform others in computational efficiency, prediction or estimation, we empirically evaluated the performance -predictive accuracy, number of features selected and computational efficiency- of all penalties included in the *SIS* R package (LASSO, elastic-net, Aenet, MSAenet, SCAD and MCP) in combination with the ISIS tool using data from the SHS (see section 1.5). Our main outcome was continuous (BMI, measured in kg/m^2). However, we also report performance metrics for a survival outcome (lung cancer) and a binary outcome (diabetes incidence).

The reason why we consider survival and binary outcomes as secondary outcomes is that the real-valued outcome from a Gaussian model is more informative than the survival outcome from Cox (due to censoring) or the dichotomous outcome from binomial models (due to categorization) [150]. For this reason, as explained in section 3.1.2, being n the sample size, the maximum default number of variables selected by Cox and binomial ISIS models is smaller than that of the Gaussian outcomes. This implicitly means that we would need a bigger sample size for Cox and binomial models to obtain the same number of variables as for the Gaussian model. Poisson regression could constitute an alternative as an approximation to survival and bino-

mial models, as the maximum number of variables selected is set to $\left\lfloor \frac{n}{2\ log(n)} \right\rfloor$. However, to date, no R packages fitting Poisson regression for Aenet have been developed. Nevertheless, and given that time-to-event and binary outcomes are widely used outcomes in epidemiologic research, we present performance measures for those outcomes as well.

In addition, we compared the effect estimates and CIs obtained from ISIS for the three outcomes to those obtained from alternative shrinkage Bayesian methods and to traditional regression approaches (linear regression for continuous outcomes, Cox regression for survival outcomes and logistic regression for binary outcomes). We finally conducted bioinformatic pathway enrichment and network analyses to assess the extent of overlap and connection of biological pathways captured by the evaluated regularization methods.

Outcome Assessment

BMI was calculated as weight in kilograms divided by the square of height in meters. Lung cancer was defined as time to incident lung cancer after excluding those individuals that had prevalent lung cancer at the baseline visit, and it was assessed by interviews, death certificates and/or chart reviews which included pathology reports. We calculated the follow-up from the date of baseline examination to the date of cancer diagnosis or 31 December 2017, whichever occurred first. Diabetes was defined as a fasting glucose level of 126 mg/dL or higher, a 2-hour post-load plasma glucose level of 200 mg/dL or higher, a glycosylated hemoglobin (HbA1c) level of 6.5 % or higher, or the use of insulin or an oral hypoglycemic agent.

Statistical Analysis

To compare the different shrinkage methods combined with ISIS in the context of a continuous outcome using linear regression, we used BMI measured in kg/m^2. The dataset was randomly split into a training set (N=1676) and a test set (N=559). The Mean Squared Error (MSE),

defined as the average squared difference between observed and predicted values, was calculated for each method for both the training set and the test set. In addition to MSE-s, effect estimates, computational times and numbers of variables selected were reported for each shrinkage method. On the other hand, the ISIS algorithm internally relies on several random subset or observation selection processes such as cross-validation or bootstrap. In settings in which variables are highly correlated and have similar associations with the outcome, setting different seeds might lead to the selection of different sets of variables. In order to evaluate the impact of setting a different seed in the MSE-s of different shrinkage methods, we repeated the models changing the established seed.

For the time-to event and dichotomous outcomes, the dataset was randomly split into a training set (N=1677, 73 lung cancer cases, 694 diabetes cases) and a test set (N=558, 24 lung cancer cases, 232 diabetes cases), having about 3/4 of the cases in the training set and 1/4 of them in the test set. The concordance index (C index), which is defined as the proportion of concordant pairs between observed and predicted values (1 would be perfect prediction, whereas 0.5 would be random prediction), was used as the evaluation metric for the time-to-event outcome. The Area Under the ROC Curve (AUC), which represents the degree of separability of a binary classifier, was used as the evaluation metric for the dichotomous outcome.

For the three outcomes, the predictors were DNA methylation measurements at the 788,368 CpG sites, and models were adjusted for age, sex, study center (Arizona, Oklahoma, and North Dakota and South Dakota), smoking status (never, former, current), five genetic principal components and estimated cell counts (CD8T, CD4T, monocytes, B cells and NK cells). Lung cancer and diabetes models were further adjusted for BMI. The lung cancer model was additionally adjusted for cumulative smoking dose (cigarette pack-years). We refer to CpGs statistically significantly associated with the outcomes as Differentially Methylated Positions (DMPs).

Pathway enrichment analysis

The enrichment analysis aims to provide a global view of biological pathways associated to a list of genes, as it considers the accumulated biological knowledge of how the genes of interest work together, allowing the identification and quantification of over-represented genes in pre-specified pathways. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a public database for the systematic analysis of gene functions that links genomic information with high-level functions of the biological system [156]. To evaluate how the different ISIS-regularized methods compare in terms of capturing biological pathways, we conducted KEGG pathway enrichment analyses out of the list of the genes annotated to selected BMI-DMPs by each of the methods separately (method-specific networks), as well as for the union set of genes annotated to DMPs selected by the six methods together (overall network). We did not conduct enrichment analysis for lung cancer and diabetes endpoints as the number of selected variables was not sufficient. The significance threshold for KEGG pathway enrichment, based on a two-sided hypergeometric test, was set to a 0.05 In addition, the cut-off of the Kappa statistic, which is used to define KEGG terms interrelations (edges) and functional similarity groups based on shared genes between terms, was set to 0.6. For the overall network, to study the ontological contribution of each of the methods, we represented the nodes as slices according to the proportion of the genes from each method that contribute to the pathway. The pathway and network enrichment analyses were performed using Cytoscape v.3.8.2 [157] with ClueGO v.2.5.8 [158] and CluePedia v.1.5.8 [159] plugins.

Results

Predictive accuracy. Table 3.1 shows MSE-s (for the continuous outcome), C indexes (for the survival outcome) and AUC-s (for the dichotomous outcome) for each shrinkage method paired with ISIS. Aenet showed the smallest MSE in the test set for the continuous outcome and the highest AUC in the test set for the dichotomous outcome, which constitutes the best predictive accuracy. For the survival outcome, although elastic-net showed the best C index in the

test set, all methods performed similarly in terms of prediction except MSAenet, which had a worse C index in the test-set.

Table 3.1: Performance measures (predictive accuracy, number of variables selected and elapsed time) for each shrinkage method paired with Iterative Sure Independence Screening

Method	Aenet	LASSO	Elastic-net[a]	MSAenet	SCAD	MCP
Continuous outcome (body mass index)						
MSE train	21.3	14.7	17.2	22.5	12.6	11.8
MSE test	30.9	41.6	36.5	43.0	50.1	51.8
N variables selected	214	224	224	135	210	214
Elapsed time (hours)	146.4	124.8	108	81.6	67.2	69.6
Survival outcome (lung cancer)						
C index train	0.95	0.91	0.93	0.94	0.94	0.94
C index test	0.69	0.72	0.73	0.60	0.71	0.68
N variables selected	56	35	56	56	42	20
Elapsed time (hours)	40.5	77.4	17.5	39.6	66.2	47.2
Dichotomous outcome (diabetes)						
AUC train	0.83	0.86	0.83	0.83	0.86	0.88
AUC test	0.80	0.78	0.78	0.76	0.76	0.77
N variables selected	53	57	57	50	46	53
Elapsed time (hours)	32.9	35.8	29.3	81.4	38.6	62.6

Models adjusted for age (years), sex, study center (Arizona, Oklahoma or North Dakota / South Dakota), smoking status (never, former or current), five genetic PCs and estimated cell count proportions (CD8T, CD4T, monocytes, B cells and NK cells). Models for lung cancer and diabetes additionally adjusted for BMI. Model for lung cancer additionally adjusted for cigarette pack-years.

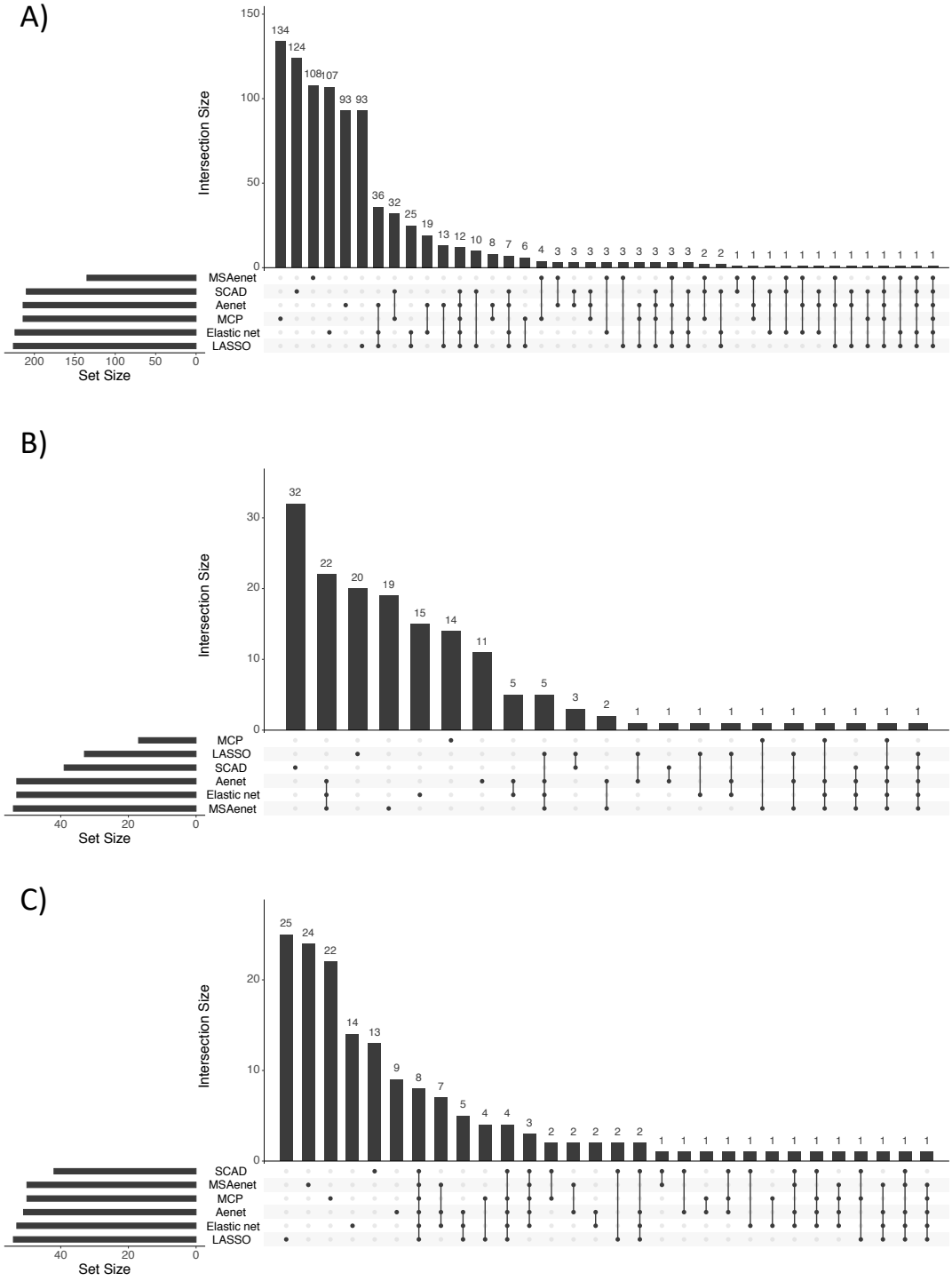
[a] $\alpha = 0.05$ ($\alpha = 0$ corresponds to Ridge regression, while $\alpha = 1$ corresponds to LASSO regression).

Lung cancer outcome: 73 cases in the training set, 24 cases in the test set.

Diabetes outcome: 694 cases in the training set, 232 cases in the test set.

Feature selection. For the continuous outcome, one feature was commonly selected for all methods (Figure 3.1A). There was an overlap of 12 DMPs for all methods except MSAenet, and additional overlap of 36 features for Aenet, elastic-net and LASSO. MSAenet was the method leading to the smallest set of features selected (135 variables). For the survival outcome, no features were commonly selected for all methods (Figure 3.1B). 22 features were selected in common for Aenet, MSAenet and elastic-net. MCP was the method leading to the smallest set of features selected (20 variables). For the dichotomous outcome, 8 features were commonly selected for all methods (Figure 3.1C). Nine more features were commonly selected for five methods.

Figure 3.1: Overlap of selected differentially methylated positions comparing different shrinkage methods for the A) Body mass index model (continuous outcome), B) Lung cancer model (survival outcome), C) Diabetes model (dichotomous outcome).



Bins in the upset plot are mutually exclusive. Thus, in order to obtain the intersection between two sets, the frequency of each of the bins in which those two sets are present need to be added.

Effect estimation. Effect estimates and 95 % CIs for the DMPs that were selected for each shrinkage method paired with ISIS, as well as effect estimates from the traditional models and the Bayesian model, are shown in Appendix A, in Tables A1 to A6 (for the continuous outcome), Tables A7 to A12 (for the survival outcome) and Tables A13 to A18 (for the dichotomous outcome). In general, the effect estimates from all methods went in the same direction and were quite consistent. The Aenet method showed attenuated coefficients as compared to other methods.

Computational efficiency. Computational times for each shrinkage method and each outcome are shown in Table 3.1. For the continuous outcome, the algorithm is much more computationally expensive because the default maximum number of selected variables is higher than for the survival outcome (see section 3.1.2). Major differences arise between shrinkage methods in terms of computational time for the continuous outcome. MCP and SCAD were the most computationally efficient methods (69.6 and 67.2 hours, respectively). For the survival and dichotomous outcomes, the most computationally efficient method was elastic-net (17.5 and 29.3 hours, respectively).

Performance measures for BMI with a different seed. The predictive accuracy in both the training set and the test set, as well as the numbers of variables selected, remained similar when changing the random seed in the ISIS algorithm, as shown in Table 3.2.

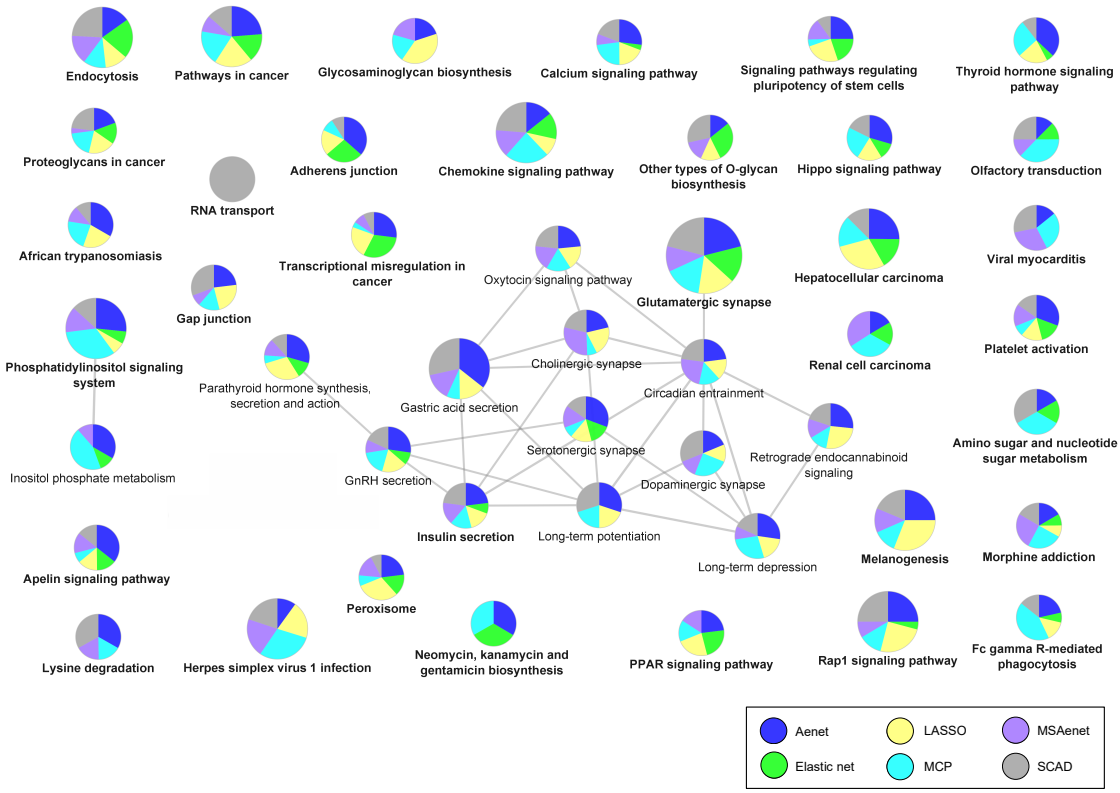
Table 3.2: Performance measures and number of variables selected for the continuous outcome using a different seed in ISIS.

Method	Aenet	LASSO	Elastic-net[a]	MSAenet	SCAD	MCP
MSE train	22.89	14.50	15.73	24.25	14.71	20.18
MSE test	31.78	38.29	35.38	42.0	41.65	33.77
N variables selected	222	221	222	113	93	42

Models adjusted for age (years), sex, study center (Arizona, Oklahoma or North Dakota / South Dakota), smoking status (never, former or current), five genetic PCs and estimated cell count proportions (CD8T, CD4T, monocytes, B cells and NK cells).
[a] $\alpha = 0.05$ ($\alpha = 0$ corresponds to Ridge regression, while $\alpha = 1$ corresponds to LASSO regression).

Pathway enrichment analysis. In the enrichment analysis of the union set of genes from method specific BMI-DMPs, Aenet was the method with the highest number of overlapping pathways (N=20) in common with at least one of the method-specific enrichment (Appendix A, Figure A1). Aenet was also the method that identified the highest number of enriched biological pathways (40 enriched pathways, as compared to 12 for elastic-net, 17 for LASSO, 14 for MCP, 12 for MSAenet and 14 for SCAD). In the overall network (Figure 3.2), Aenet was the method that contributed most genes to the enriched KEGG categories (i.e. genes annotated to BMI-DMPs from Aenet contributed to 45 out of the 46 biological pathways). All method-specific networks and enriched pathways for each method can be found in Appendix A (Figs. A2 to A7). While the pathways identified in the method-specific networks for LASSO, elastic-net, MSAenet, MCP and SCAD were not connected, Aenet showed many highly connected pathways (Appendix A, Figure A2).

Figure 3.2: Overall network of the significantly enriched pathways for the BMI outcome for the selected genes of the six methods.



KEGG pathways are represented as nodes and the node size represents the term enrichment significance. The size of the slices represents the proportion of the genes that contribute to the metabolic pathway for each method.

Discussion

We extended the SIS R package to pair the algorithm with Aenet, elastic-net and MSAenet, including the implementation of a quantile bootstrap-based approach to obtain CIs for the coefficients of selected features. In addition, we used DNA methylation microarray data from the SHS to compare the performance of all the available regularization methods combined with ISIS. We observed that, while there are specific ISIS-regularization method combinations that may be most suitable for prediction or estimation only, the ISIS-Aenet combination

can achieve the most-balanced compromise between both an optimal feature selection and estimation.

Regarding predictive ability, Aenet was the least likely method to overfit to the training data set for the continuous and binary outcomes (the MSE and AUC for the training set and the test set were similar). For the survival outcome, enet was the method with the highest C index, although all methods had similar C indexes except MSAenet, which had a much lower predictive ability for survival data. MSAenet did not provide as good results in terms of minimizing the error as did Aenet and elastic-net, and had, in general, less selected features in common with all the other methods. This suggests that applying weights to both L_1 and L_2 norms, as done in MSAenet, does not provide an improvement beyond applying them only to the L_1 norm part of the penalty, as done in Aenet.

In general, models that lead to sparser solutions might lead to selection of undesirably minimal predictor sets. In the omics data setting, variable selection is commonly used to obtain manageable variable sets in which associations with other traits will be tested. Thus, if the goal of the analysis is biological discovery and not prediction, the statistical analysis should be focused into not missing important features. In this sense, Aenet, elastic-net and LASSO lead to a higher number of selected variables. Given that LASSO is not able to select more than one variable from a correlated set, Aenet and elastic-net might be more prone to favor biological discovery.

The effect estimates were mostly consistent across different regularization methods and similar to those from traditional methods, suggesting that shrinkage methods, when combined with ISIS, are a reliable tool for effect estimation in addition to feature selection. However, the regression coefficients from Aenet were somewhat attenuated as compared to the regression coefficients from other methods. The adaptive weights used in Aenet, which provide a more precise effect estimate by incorporating prior information on the coefficients, might contribute to this attenuation. Consistently, Bayesian shrinkage methods, which also incorporate prior information, led to attenuated effect estimates, similar to Aenet. On the other hand, given that LASSO

and elastic-net estimators do not fulfill the oracle property, effect estimates from Aenet, SCAD or MPC are expected to be, theoretically, more reliable.

Interestingly, many DMPs that were selected by the shrinkage methods did not show statistically significant associations when using conventional methods (Appendix A, Tables A1 to A18). This might be related to multicollinearity after simultaneous introduction of the multi-markers in the model, which tends to inflate standard errors and undermine statistical significance. Linear regression provides unbiased effect estimates, but shrinkage methods are able to lower the variance of the estimators, thus generally reaching a better variance-bias trade-off [108] and reducing regression dilution bias, leading to smaller MSE-s than those of traditional methods.

Importantly, the pathway enrichment analysis of genes annotated to BMI-DMPs suggests that pairing ISIS with AEnet leads to the most robust selection of biologically relevant features as compared to other regularization methods. The overall network included several routes related to neurotransmitters and hormones release, which is consistent with the well-known role of high BMI on insulin resistance and diabetes [160]. Several other BMI-related pathways were related to cancer, which is consistent with the results of meta-analysis of 1000 observational studies supporting that excess body fat is associated with increased cancer risk [161]. Interestingly, in method-specific enrichment analysis, only Aenet (and to a lesser extent elastic-net) detected KEGG categories associated with CVD. Previous epidemiological studies consistently support significant associations between BMI and CVD [162, 163, 164].

Computational efficiency is a challenge when dealing with ultra-high dimensional data, as unmanageable computational times can be easily reached when using complex algorithms in datasets with thousands or millions of variables. Although the difference in computational efficiency between shrinkage methods was not large when the maximum number of variables selected was set to $\left\lfloor \frac{n}{4\ log(n)} \right\rfloor$ (for the time-to-event outcome), it became more evident when the maximum

number of variables was increased to $\left\lfloor \frac{n}{log(n)} \right\rfloor$ (for the continuous outcome). Future lines of research should evaluate performance of the ISIS Cox and logistic models with the implemented regularization methods in larger studies. On the other hand, we could not evaluate Bayesian shrinkage methods on the complete set of 788,368 CpG sites included in the microarray given the computational unfeasibility of MCMC-based approaches. The implementation of more computationally efficient Bayesian shrinkage methods is left for future research.

In summary, differences in feature selection across methods highlight the importance of selecting the most adequate shrinkage method depending on whether the objectives of the study are predictive accuracy, estimation, sparsity or computational efficiency. Our results suggest that pairing the ISIS tool with Aenet is a good compromise for feature selection, effect estimation and biological discovery, as compared to the regularization methods previously paired with ISIS.

3.2.2 Data Application 2: Arsenic Exposure, Blood DNA Methylation and Cardiovascular Disease

Epigenetic dysregulation has been proposed as a key mechanism for arsenic-related CVD. We hypothesized that epigenetics, measured based on DMPs in blood, can partially explain arsenic-related CVD. To test this hypothesis, we first used the ISIS-Aenet tool described in section 3.1.2 to select relevant DMPs for CVD. Subsequently, we conducted a simple mediation analysis in those selected DMPs.

Our main population was the SHS, described in section 1.5. Prior evidence in the SHS showed that baseline arsenic exposure from ground contamination, which was stable for decades, was associated with increased CVD risk [26] and with differentially methylated blood DNA in an EWAS [49]. In order to validate our findings, we also used data from three independent cohorts: the Framingham Heart Study (FHS), Women's Health Initiative (WHI) and Multi-Ethnic Study of Atherosclerosis (MESA) to assess if DMPs associated with arsenic-mediated CVD in the SHS were associated with incident CVD in those populations. Since MESA is, to our knowledge, the only other United States (US) cohort apart from the SHS that has data on arsenic, DNA methylation and CVD, we also used data from MESA to assess if the same DMPs were associated with arsenic exposure.

Our results were additionally validated in an animal model of apolipoprotein knockout mice [38], in which DNA methylation was measured in liver tissue and DMPs and differentially methylated regions (DMRs) were identified. Last, we conducted bioinformatic analyses to identify enriched biological pathways and assess the biological plausibility of our findings.

Arsenic measurements in the Strong Heart Study

Arsenic measurements in spot urine samples have been described in detail [165]. Briefly, arsenic species (inorganic arsenic, monomethylarsonate (MMA), dimethylarsinate (DMA), and arsenobetaine) were measured using high-performance liquid chromatography coupled to inductively coupled plasma mass spectrometry (Agilent 1100 HPLC

and Agilent 7700x ICP-MS; Agilent Technologies). Urinary creatinine was measured in the same urine sample used for arsenic measurement using an automated alkaline picrate methodology run on a rapid flow analyzer. As the biomarker of inorganic arsenic exposure (referred to as urinary arsenic in the manuscript for simplicity), we calculated the sum of inorganic and methylated arsenic species (MMA and DMA) concentrations (μg/L). This biomarker was divided by urinary creatinine (g/L) to account for urine dilution.

Outcome assessment in the Strong Heart Study

The endpoints were incident fatal and non-fatal CVD assessed during the follow-up by annual mortality and morbidity surveillance of medical records, which included evaluation of medical history and physical examinations, emergency room visits, medical consultations, electrocardiograms, laboratory assays, medical imaging, discharge summaries, operations, and other procedures from the Indian Health Service and other facilities. Mortality surveillance examined death certificates from state health departments, records from the Indian Health Service, autopsy and coroner's reports, and interviews with physicians or family members. Potential CVD-related deaths and events were reviewed by two independent physicians. In case of disagreement, they were adjudicated by a third independent physician. Incident CVD was defined as the first occurrence of fatal or non-fatal CHD, stroke or congestive heart failure, or other non-fatal CVD. CVD mortality was defined as any fatal CVD. Follow-up time was calculated as the time from blood drawn for DNA methylation measurements (1989-1991) to the time of CVD events (through 2009). For participants who did not develop CVD, follow-up was censored at the time of occurrence of non-CVD death, loss to follow-up, or December 31, 2009.

For this analyses, we restricted the follow-up through 2009 as water arsenic exposure, which was stable in the communities for decades [166], changed a few years after the enactment of the US Environmental Protection Agency (EPA) final arsenic rule in 2006 [167, 168].

Replication populations

We used data from the FHS, WHI, and MESA to replicate the DMPs associated with arsenic-mediated CVD in the SHS. All of them used follow-up procedures for CVD events and pre-processing of blood DNA methylation similar to those used by the SHS. Details on CVD outcome assessment and DNA methylation measurements for each cohort, and arsenic measurements for MESA, are provided in the Supplementary Material of Domingo-Relloso et al. [89].

Briefly, FHS recruited White adults of European descent from Framingham, Massachusetts starting in 1948 (original cohort). The children of the original cohort and their spouses were recruited into the Framingham Offspring study in 1971. The participants of exam 8 (2005-2008) of FHS offspring cohort were followed through 2014 (average follow-up of 7.7 years; range: 0.04 years – 9.8 years). DNA methylation was measured from whole blood samples using the Illumina Infinium HumanMethylation450K Beadchip array (referred to as 450K hereinafter). Among 2,631 FHS participants with blood DNA methylation data available in the FHS Offspring, we excluded those with prior CVD (N=316) and those missing information on CVD risk factors (N=325), leaving 1,990 participants with 408,254 CpG sites available. DNA methylation measurements in the FHS were conducted in two separate batches including 1879 and 111 participants, respectively. We conducted a sensitivity analysis excluding the 111 individuals in the second batch from the analysis.

WHI enrolled 161,808 women of diverse ethnicities (including White, African American, Native American, Hispanic, Asian and pacific Islanders) starting in 1993 as part of randomized control trials that were continued as a prospective cohort study. The participants of WHI were followed from baseline (1993-1998) to 2016 with an average follow-up time of 12.18 years (range: 0.003 – 21.3 years). DNA methylation was also measured in whole blood using the 450K array. Details regarding measurements, quality control and preprocessing have been published [169]. Among 2,096 WHI participants with blood DNA methylation for 434,113 CpG sites, we excluded those with missing information on traditional risk factors of CVD, leaving 1,487 participants.

MESA followed participants of diverse ethnicities (White, African-American, Hispanic and Asian) through 2017 with an average follow-up time of 15.56 years (range: 7.76 – 17.42 years). DNA methylation was measured in whole blood using the EPIC array. From 916 participants that had DNA methylation data and prospective CVD data, 20 were excluded due to missing covariates. The final sample size for DNA methylation and CVD analyses was 896. From 214 participants that had DNA methylation and urinary arsenic data, 8 were excluded due to missing covariates. The final sample size for DNA methylation and arsenic analyses was 206.

Statistical methods

DMPs associated with CVD. To identify DMPs associated with CVD incidence and mortality, we used Cox ISIS-Aenet. We entered all the 788,368 CpG sites simultaneously to select DMPs associated with CVD incidence and mortality (dependent variables, in separate models). CIs were calculated using the quantile bootstrap method as described in section 3.1.5. Models were adjusted for baseline covariates including age, sex, smoking status (never, former, current), BMI, low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, diabetes status (yes/no), hypertension medication (yes/no), systolic blood pressure (mm/Hg) and albuminuria (micro, macro, normal), which are established CVD risk factors in the SHS. Given the different characteristics of the three study centers (Arizona, Oklahoma, and North Dakota and South Dakota), models were also adjusted for study center. Models were also adjusted for estimated cell proportions (CD8T, CD4T, NK, B cells, and monocytes) and five genetic PCs (see section 1.5).

Mediation analysis. To identify DMPs that may explain arsenic-related CVD, we used additive hazards models for causal mediation analysis with survival outcomes as explained in section 2.4, similar to other mediation studies with time-to-event data [170, 171]. The DMPs tested as possible mediators included the DMPs identified as relevant for CVD by ISIS – Aenet, as well as 315 DMPs identified as associated

with arsenic exposure using an elastic-net model in the SHS in a previous study [49]. The additive hazards model included time to incident CVD (or CVD mortality, in a separate model) as the outcome, baseline urine arsenic (modeled as log2) as the exposure, and DNA methylation as the mediator (each DMP in a separate model). Our mediator model was a linear model with logit2-transformed methylation values (M values) as the outcome (each DMP in a separate model) and urine arsenic (modeled as log2) as the exposure. Both the outcome and mediator models included adjustment for the same covariates (age, sex, smoking status, BMI, LDL cholesterol, study center, cell counts and genetic PCs).

Mediated effects (natural indirect effects) were reported as the number of CVD cases per 100,000 person-years associated with a 2-fold increase in urinary arsenic that are attributable to DNA methylation changes in that CpG site. CIs were calculated using a resampling method that takes random values from multivariate normal distribution of the estimates, as described by Lange and Hansen [139]. Total effects, direct effects and indirect effects with CIs not including 0 were considered significant. To account for the withdrawal of one of the Tribal Nations, the primary mediation analyses used inverse probability weighting to reduce bias [172]. We weighted the participants remaining in the study with approximately 1/3 of weight for each center based on the baseline SHS cohort enrollment (33.0 % Arizona, 33.6 % Oklahoma, 33.4 % North Dakota / South Dakota). Unweighted models are presented in the Supplementary Material of Domingo-Relloso et al. [89].

Protein-protein interaction network to evaluate biological plausibility of identified DMPs. Arsenic-associated and CVD-associated DMPs were annotated to the nearest protein coding gene and included in a protein-protein interaction network. The interactions between nodes were obtained using the STRING database v11.0 [173], selecting all active interaction sources with a confidence score of 0.4. The confidence score (from 0 to 1) provided by the STRING database estimates the likelihood that an annotated interaction between a pair of proteins

is biologically meaningful, specific and reproducible. The network was analyzed and displayed using Cytoscape v3.8.2.61 [157].

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. We used the *missmethyl* R package [174] to conduct gene ontology enrichment and KEGG analyses. This package conducts enrichment analysis taking two sources of bias into account: the differing number of probes per gene, and CpGs that are annotated to multiple genes. We tested whether any GO terms or pathways were enriched for the set of DMPs that were significant in the mediation analysis for both CVD incidence and mortality, as compared to the total number of CpG sites that were tested in the mediation analysis (329 for CVD incidence and 338 for CVD mortality).

Cross-reference with the EWAS catalog to evaluate biological plausibility. For DMPs showing significant mediated effects for arsenic-related CVD incidence and/or mortality, we looked for previously known trait associations in the EWAS Catalog [1]. This catalog contains information on EWAS conducted across the literature and is regularly updated (we used the February 4, 2021 version). For DMPs with several traits in the EWAS catalog, either the most relevant trait or the study with the largest sample size were selected.

Sensitivity Analyses. Because diabetes and hypertension might be in the arsenic-CVD causal pathway, the main models were not adjusted for those variables. We repeated the mediation analyses for CVD incidence and CVD mortality adjusting for diabetes status and for hypertension treatment and systolic blood pressure.

Differentially Methylated Genomic Regions and Positions in Livers of Arsenic-Exposed Mice. Apolipoprotein E knockout (apoE-/-) mice are a well-established animal model of atherosclerosis, where genetic manipulation results in hyperlipidemia. Importantly, the model increases disease burden in response to dietary changes (i.e. high fat) [175] and environmental exposures (i.e. arsenic) [38]. This model is relevant for many human populations which diets are also lipid-rich, such as the

typical diet of many participants in the SHS [176, 177]. B6.129P2-ApoEtm1Unc/J (ApoE/) mice were obtained from the Jackson Laboratory (see Domingo-Relloso et al. 2022 [89] for further details). The male and female apoE-/- mice were assigned randomly into mating pairs prior to arsenic exposure. Arsenic exposure was then provided through drinking water or not to the female during the duration of pregnancy based on the random assignment of the mating pair. At endpoint, livers were harvested from the offspring (N=3 per sex, per treatment group). A total of 12 liver samples from randomly chosen offspring of each unique litter were sequenced. DNA was isolated from liver tissues, and bisulfite conversion and whole-genome bisulfite sequencing were performed.

The data were processed using the GemBS pipeline [178], using the MM9 mouse reference genome. A chromosome-wise matrix of methylation counts and read counts (after quality control filter) was created for all samples. The *BSmooth* function [179] from the *bsseq* bioconductor package was applied to smooth the data, and t-statistics were calculated. Finally, the *dmrfinder* function was used to identify genomic regions that were differentially methylated in the tissue samples from the offspring of exposed dams compared to the offspring of control dams. For the identification of differentially methylated CpG sites in the genes of interest, the R package *limma* [103] was used separately for male and female. The DMRs were annotated with the MM9 annotations using CHIPseeker [180] and Annotatr [181].

Results

A total of 847 participants developed incident CVD in the SHS (36.4 %), 208 in the FHS (10.4 %), 754 in the WHI (50.7 %) and 87 in MESA (9.7 %). In the SHS, individuals with incident CVD were older and more likely to have diabetes, higher LDL cholesterol, hypertension, higher systolic blood pressure and micro and macro albuminuria. Individuals who died of CVD had higher levels of urinary arsenic at baseline (Table 3.3). Participants' characteristics by CVD status for the replication cohorts are shown in Table 3.4.

Table 3.3: Baseline participant characteristics by cardiovascular disease incidence and mortality status in the Strong Heart Study.

	Non-incident CVD (N=1474)	Incident CVD (N=847)	CVD death (N=316)
Age (years), median (IQR)	53.1 (48.0, 60.0)	57.3 (51.0, 64.4)	58.4 (52.6, 66.2)
Sex, % Men	60.0	58.3	56.8
Smoking status, %			
Former	33.3	33.4	29.6
Current	32.3	36.4	34.3
BMI, median (IQR)	29.8 (26.3, 34.2)	30.4 (27.1, 34.5)	30.4 (27.1, 34.3)
LDL cholesterol (mg/dL), median (IQR)	114 (92, 135)	121 (99, 142)	121 (100, 144)
HDL cholesterol (mg/dL), median (IQR)	44 (38, 53)	42 (36, 50)	41 (36, 49)
Systolic blood pressure, median (IQR)	122 (111, 135)	129 (118, 141)	133 (120, 144)
Hypertension, %	15.3	30.1	34.5
Diabetes, %	40.3	61.9	69.2
Albuminuria, %			
Microalbuminuria	15.1	24.5	24.2
Macroalbuminuria	6.4	15.8	24.4
Urinary arsenic (μg/g creatinine)*	10.2 (5.9, 16.7)	10.3 (6.0, 17.3)	11.2 (6.6, 18.2)

CVD: Cardiovascular disease, IQR: interquartile range.

*Urinary arsenic corresponds to the sum of inorganic and methylated species (methylarsonic acid and dimethylarsinic acid) in the urine.

Table 3.4: Baseline participant characteristics by cardiovascular disease incidence status for the replication cohorts.

	Framingham Heart Study		Women's Health Initiative		Multi-Ethnic Study of Atherosclerosis	
	Non-incident CVD (N=1792)	Incident CVD (N=198)	Non-incident CVD (N=733)	Incident CVD (N=754)	Non-incident CVD (N=848)	Incident CVD (N=68)
Age (years), median (IQR)	64.0 (59.0, 70.0)	71.0 (64.0, 78.0)	64.0 (58.0, 69.0)	65.0 (60.0, 70.0)	68 (61, 77)	74.5 (65.0, 82.0)
Sex, % Men	41.6	52	-	-	46.7	57.4
Smoking status, %						
Former	-	-	6.68	11.27	49.9	53.7
Current	8.1	6.1	38.61	37.53	8.6	10.4
BMI, median (IQR)	27.3 (24.3, 30.7)	29.0 (25.7, 31.7)	28.6 (25.0, 32.6)	29.4 (25.8, 33.6)	28.3 (25.2, 32.2)	28.0 (24.2, 30.7)
LDL cholesterol (mg/dL), median (IQR)	190 (166, 214)	182 (161, 205)	139 (119, 162)	145 (121, 171.1)	106 (82, 127)	110 (80, 125)
HDL cholesterol (mg/dL), median (IQR)	57 (46, 70)	50 (40, 62)	54 (46, 64)	49 (43, 58)	52 (44, 63)	50 (41, 63)
Systolic blood pressure, median (IQR)	126 (115, 137)	133 (123, 144)	127 (115, 139)	133 (122, 146)	120 (110, 135)	129 (113, 144)
Hypertension, %	42.6	63.6	31.5	47.6	57.9	69.1
Diabetes, %	9.4	22.2	6.1	15.0	19.5	20.6
Albuminuria, %						
Microalbuminuria	0.28	2.0	-	-	11.1	21.2
Macroalbuminuria	5.6	14.7	-	-	2.4	7.6
Urinary arsenic (μg/g creatinine)*	-	-	-	-	2.9 (1.7, 4.9)	3.1 (1.8, 4.5)

CVD: Cardiovascular disease, IQR: interquartile range.
*Urinary arsenic corresponds to the sum of inorganic and methylated species (methylarsonic acid and dimethylarsinic acid) in the urine.

The Cox ISIS-Aenet model selected 70 and 72 DMPs as relevant for CVD incidence and mortality, respectively (Appendix B, Table B1 and Table B2).

In the mediation analysis for CVD incidence, which included the 70 DMPs selected by ISIS-Aenet and 315 DMPs associated with urinary arsenic in our previous study [49], we found statistically significant mediated effects for 21 DMPs (seven from ISIS–Aenet model, and 14 among those previously associated with arsenic) (Table 3.5). For CVD mortality, which included 72 DMPs selected by ISIS-Aenet and 315 DMPs associated with urinary arsenic in our previous study, we found statistically significant mediated effects for 15 CpG sites (five from the ISIS–Aenet model and 10 previously associated with arsenic) (Appendix B, Table B3). The DMPs cg05779585 (*LOC286083*), cg19693031 (*TXNIP*), cg06716655 (*ADAR*), cg17608381 (*HLA-A*),

cg22294740 (*LINGO3*), cg11946459 (*HLA-A*), cg03362418 (*TYMP*) and cg06970472 (*APBB2*) were common significant mediators for arsenic-related CVD incidence and mortality (two from the ISIS–Aenet model and four from those previously associated with arsenic).

Table 3.5: Incident CVD cases per 100,000 person-years for the doubling of urinary arsenic levels not attributable (direct effect) and attributable (indirect effect) to changes in DNA methylation for each CpG (one marker at a time approach).

CpG	Gene	Function	Cases attributable to a doubling of urinary As (95% CI) (direct effect)	Cases attributable to a doubling of urinary As through DNAm (95% CI) (indirect effect)	% cases attributable to a doubling of urinary As explained by DNAm (95 % CI)
cg19693031	*TXNIP*	Binding partner for redox signaling protein thioredoxin	137.6 (-61.2, 335.9)	95.7 (43.8, 158.8)	41.0 (14.5, 183.0)
cg05779585	*LOC286083*	Unknown function	200.2 (5.8, 394.2)	69.2 (5.8, 161.2)	25.7 (1.8, 83.6)
cg03497652	*ANKS3*	Vasopressin signaling in the kidney	181.7 (-14.4, 377.5)	46.1 (12.9, 86.5)	20.2 (3.8, 97.4)
cg01270753	*TGFBR1**	Aortic disease and altered cardiovascular development	200.3 (8.7, 391.4)	43.9 (13.6, 82.9)	18.0 (4.7, 70.6)
cg22294740	*LINGO3*	Unknown function	185.3 (-11.5, 381.9)	43.3 (7.0, 8.4)	18.9 (1.3, 92.4)
cg03362418	*TYMP**	Angiogenesis in vivo. Possible therapeutic target for CVD	190.3 (-3.8, 383.8)	40.1 (9.1, 78.6)	17.4 (2.8, 78.0)
cg23027596	*UBAC1**	Glucose-induced insulin synthesis and secretion	186.3 (-6.0, 378.1)	39.9 (11.1, 74.6)	17.6 (3.5, 80.4)
cg17608381	*HLA-A*	Central role in the immune system	196.3 (-0.4, 392.4)	35.9 (5.5, 72.9)	15.5 (1.1, 74.9)
cg09956442	*ARRDC2*	Unknown function	195.2 (1.6, 388.4)	35.3 (10.3, 67.9)	15.3 (3.4, 68.2)
cg06668829	*EPPK1**	Cytoskeletal linker protein involved in response to stress	203.4 (10.9, 395.5)	33.2 (10.1, 63.8)	14.0 (3.4, 60.5)
cg14827056	*EIF2C2*	RNA-mediated gene silencing	193.8 (-0.3, 387.5)	31.0 (5.5, 63.8)	13.8 (1.2, 67)
cg18032342	*NISCH*	Cell growth and death in cardiac tissue	197.2 (3.3, 390.8)	30.1 (2.2, 63.9)	13.2 (-0.4, 61.5)
cg13092901	*TYMP**	Angiogenesis in vivo. Possible therapeutic target for CVD	200.1 (6.4, 393.3)	30.3 (3.2, 62.7)	13.1 (0.2, 59.4)
cg11946459	*HLA-A*	Central role in the immune system	206.4 (11.6, 400.7)	27.2 (1.9, 58.8)	11.7 (-0.1, 55.5)
cg06970472	*APBB2**	Beta cell function, insulin secretion	205.7 (13.7, 397.3)	27.8 (7.7, 54.8)	11.9 (2.6, 52.3)
cg06716655	*ADAR2*	RNA editing enzyme involved in innate immunity	203.3 (7.0, 399.2)	25.7 (3.9, 56.5)	11.2 (0.9, 55.7)
cg18618815	*COL1A1**	Extracellular matrix. As-induced remodeling mice model	198.5 (3.1, 393.4)	23.7 (4.8, 49.8)	10.7 (1.2, 54.9)
cg01178924	*LMO7*	Development of muscle and heart tissues. Pancreatic cancer.	208.7 (13.6, 403.4)	23.7 (0.4, 54.7)	10.2 (-0.8, 48.8)
cg01542019	*TECR*	Sphingolipid synthesis and oxidoreductase activity	202.1 (7.7, 396.1)	21.4 (2.3, 48.4)	9.6 (0.2, 48.8)
cg02047803	*RELL2*	Apoptosis	206.3 (13.3, 398.8)	18.7 (0.7, 45.6)	8.3 (-0.3, 43.5)
cg16335098	*SMOC2*	Angiogenesis in tumor growth and myocardial ischemia	219.2 (25.7, 412.2)	13.1 (2.7, 26.9)	5.7 (0.8, 25.4)

Abbreviations: As, arsenic; DNAm, DNA methylation; CI, confidence interval.

The sum of the direct and indirect effect represents the total effect for a doubling of urinary arsenic in CVD incidence.

Models adjusted for age, sex, smoking status, BMI, LDL cholesterol, study center (Arizona, Oklahoma or North and South Dakota), cell counts (CD8T, CD4T, NK, B cells and monocytes) and genetic PCs.

*CpG sites selected by ISIS – Aenet as predictive of CVD incidence. Other CpG sites were originally identified as associated with arsenic exposure in previous research [49].

To account for the withdrawal of one of the Tribal Nations, models were weighted with approximately 1/3 of weight for each center (33.0 % Arizona, 33.6 % Oklahoma, 33.4 % North Dakota / South Dakota) using inverse probability weighting.

The adjustment for diabetes in the mediation models attenuated the indirect effects for arsenic-related CVD incidence and mortality for all DMPs, although most of them remained statistically significant for both CVD incidence and mortality. Two CpG sites that were not significant in non-diabetes-adjusted models had significant indirect effects when adjusting for diabetes; cg25371036 (annotated to *AMOTL1*) showed an indirect effect of 13.5 (0.1, 31.4) CVD incidence cases per 100,000 person-years (i.e., of 71 CVD cases per 100,000 person-years associated with a doubling of arsenic exposure, 13 cases were attributed to DNA methylation). In addition, cg22130008 (annotated to *FGG*), showed an indirect effect of 18.8 (0.53, 46.35) for CVD incidence. The adjustment for hypertension and systolic blood pressure in the mediation models lead to similar results as the primary analysis.

Among the 21 DMPs associated with arsenic-mediated incident CVD in the SHS, all of the CpG sites were available in MESA (which also used the EPIC microarray for DNA methylation measurements) and 14 were available in FHS and WHI. Among the 14 common CpG sites, six had hazard ratios in the same direction for the four populations (annotated to *LINGO3*, *TXNIP*, *HLA-A*, *EIF2C2*, *ANKS3* and *TECR*), and five more had hazard ratios in the same direction for all populations except one (Table 3.6). Results for FHS were similar

when excluding the 111 individuals from the second DNA methylation batch.

Table 3.6: Replication: hazard ratios (95 % CI) of the differentially methylated positions identified in the mediation analysis in the Strong Heart Study in three diverse US populations (Framingham Heart Study, Women's Health Initiative, and Multi-Ethnic Study of Atherosclerosis).

CpG	Gene	Strong Heart Study	Framingham Heart Study	Women's Health Initiative	Multi-Ethnic Study of Atherosclerosis
cg01178924	*LMO7*	0.86 (0.73, 1.02)	0.83 (0.60 ,1.14)	1.15 (0.94, 1.40)	1.03 (0.57, 1.85)
cg01270753	*TGFBR1*	0.60 (0.50, 0.73)	-	-	1.03 (0.52, 2.03)
cg01542019	*TECR*	1.14 (0.96, 1.36)	1.06 (0.74, 1.52)	1.26 (1.03, 1.54)	1.59 (0.78, 3.25)
cg02047803	*RELL2*	0.77 (0.65, 0.92)	0.76 (0.54, 1.07)	1.02 (0.82, 1.26)	1.77 (0.91, 3.42)
cg03362418	*TYMP*	0.60 (0.48, 0.74)	-	-	3.36 (1.44, 7.83)
cg03497652	*ANKS3*	1.50 (1.24, 1.82)	2.32 (1.58, 3.40)	1.15 (0.91, 1.44)	2.36 (1.10, 5.06)
cg05779585	*LOC286083*	0.89 (0.84, 0.95)	0.87 (0.69, 1.09)	1.18 (0.99, 1.40)	4.02 (1.89, 8.57)
cg06668829	*EPPK1*	1.44 (1.21, 1.72)	0.77 (0.53, 1.11)	1.15 (0.92, 1.44)	1.96 (0.92, 4.20)
cg06716655	*ADAR2*	0.76 (0.64, 0.9)	-	-	0.57 (0.27, 1.17)
cg06970472	*APBB2*	0.72 (0.59, 0.88)	0.64 (0.41, 0.99)	0.93 (0.73, 1.18)	3.97 (1.93, 8.19)
cg09956442	*ARRDC2*	0.71 (0.59, 0.85)	-	-	0.89 (0.45, 1.76)
cg11946459	*HLA-A*	0.76 (0.63, 0.92)	0.65 (0.46, 0.92)	0.86 (0.70, 1.06)	1.41 (0.71, 2.83)
cg13092901	*TYMP*	0.59 (0.48, 0.72)	0.54 (0.34, 0.87)	0.80 (0.63, 1.00)	1.19 (0.53, 2.67)
cg14827056	*EIF2C2*	1.41 (1.17, 1.69)	1.47 (1.01, 2.13)	1.21 (0.95, 1.54)	1.41 (0.68, 2.89)
cg16335098	*SMOC2*	0.89 (0.80, 0.99)	-	1.08 (0.94, 1.25)	0.89 (0.62, 1.28)
cg17608381	*HLA-A*	0.77 (0.64, 0.92)	0.62 (0.45, 0.87)	0.88 (0.72, 1.07)	0.93 (0.50, 1.73)
cg18032342	*NISCH*	1.27 (1.07, 1.50)	-	-	1.99 (1.06, 3.75)
cg18618815	*COL1A1*	0.63 (0.52, 0.76)	0.52 (0.35, 0.78)	1.05 (0.85, 1.30)	0.85 (0.41, 1.79)
cg19693031	*TXNIP*	0.51 (0.43, 0.59)	0.72 (0.50, 1.02)	0.76 (0.62, 0.92)	0.93 (0.50, 1.70)
cg22294740	*LINGO3*	1.42 (1.19, 1.69)	1.84 (1.31, 2.59)	1.21 (0.97, 1.50)	3.87 (2.03, 7.38)
cg23027596	*UBAC1*	0.65 (0.54, 0.79)	-	-	0.90 (0.42, 1.95)

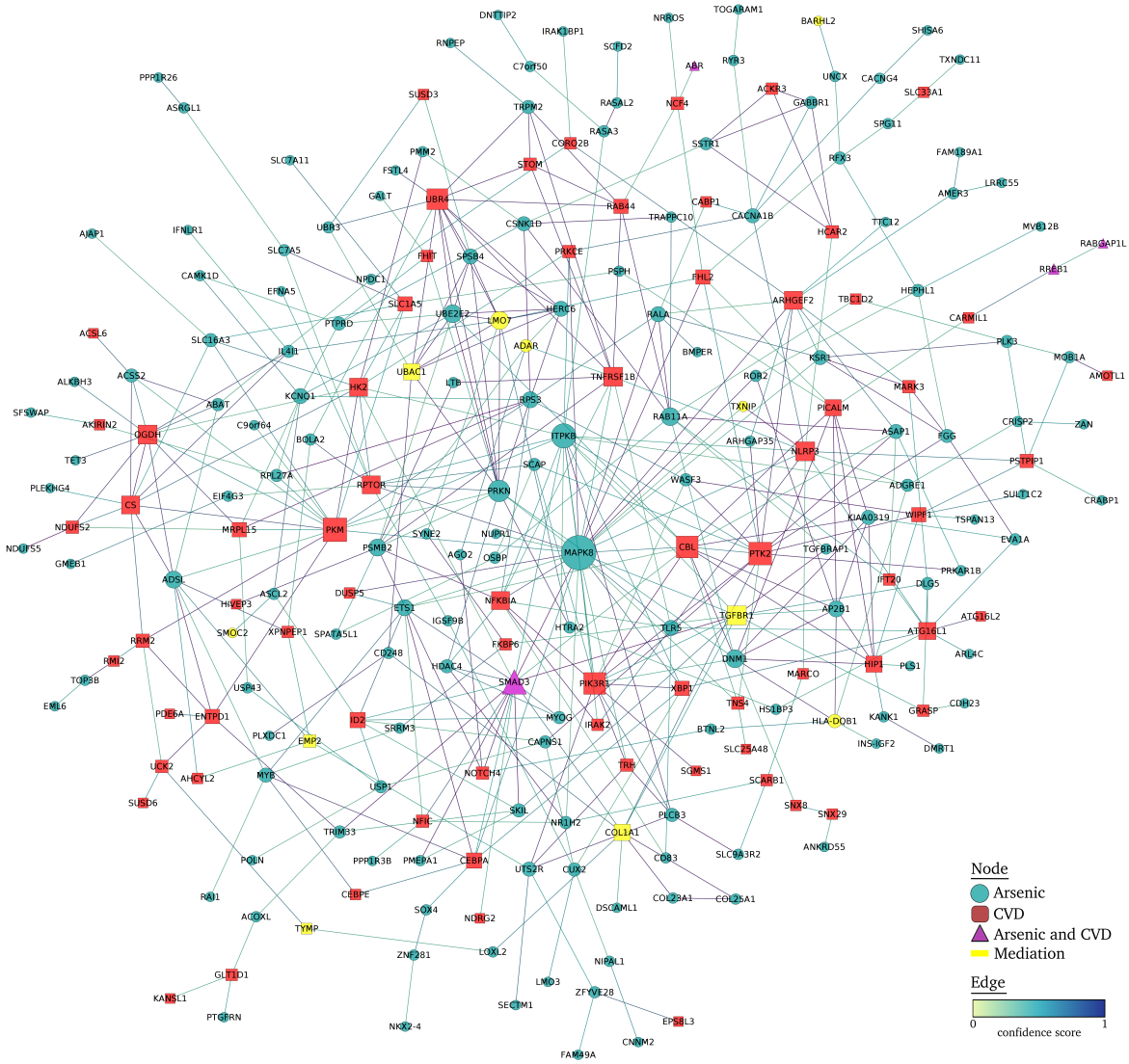
Models adjusted for age, sex, smoking status, BMI and cell counts (CD8T, CD4T, NK, B cells [eosinophils for MESA] and monocytes) for all populations. Additionally adjusted for total cholesterol in the FHS, for LDL cholesterol, study center (Arizona, Oklahoma or North and South Dakota) and genetic PCs in the SHS, for LDL cholesterol, technical covariates (plate number and pull ID) and race in the WHI, and for race, site and LDL cholesterol in MESA.

In the SHS and MESA, DNA methylation was measured using the EPIC array. In FHS and WHI, the 450K array was used.

In MESA, the only cohort with urine arsenic data available (N=206), one DMP was associated with arsenic at 0.05 p-value cut-off, and two more were associated with arsenic at 0.1 p-value cut-off. These DMPs were annotated to *EPPK1* (mean difference [SE] in methylation M values -0.018 [0.008] for one log-unit change in arsenic), *ANKS3* (mean difference [SE]: -0.018 [0.01]) and *ARRDC2* (mean difference [SE]: 0.013 [0.007]). A DMP annotated to *TXNIP* associated with arsenic before adjustment for cell counts (mean difference [SE]: 0.027 [0.008]), was no longer significantly associated after adjustment for cell counts (mean difference [SE]: -0.014 [0.02]).

In the protein-protein interaction network, we analyzed a list of 405 unique genes (from 315 genes tagged to DMPs associated with arsenic and 70 and 72 genes tagged to DMPs associated respectively with CVD incidence and mortality). Of these, 168 non-coding RNA (ncRNA) genes or unconnected nodes were discarded, obtaining a network with 237 nodes and 460 interactions (Figure 3.3). MAPK8, ITPKB and SMAD3 were the most connected nodes in the network with 28, 17 and 17 interactions, respectively, and all nodes associated with arsenic and SMAD3 were also associated with CVD. Other highly connected nodes associated with CVD were TGFBR1 or PKM, with more than 10 interactions. TGFBR1, LMO7, UBAC1 and COL1A1, with 11, 10, 8 and 8 interactions respectively, were significant in the mediation analysis.

Figure 3.3: Protein-protein interaction network of differentially methylated positions associated with CVD and with arsenic in the Strong Heart Study.



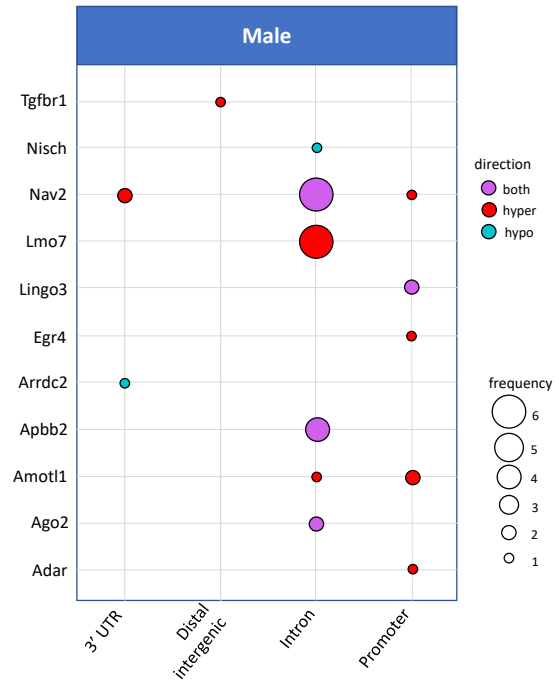
Arsenic-associated and CVD-associated differentially methylated positions were annotated to the nearest protein coding gene and included in a protein-protein interaction network. The interactions between nodes were obtained using the STRING database v11 [173] selecting all active interaction sources with a confidence score of 0.4. The network was analyzed and displayed using edge weighted spring embedded layout with Cytoscape v3.8.2 [157].

In the GO enrichment analysis, we found 110 enriched terms for CVD incidence, and 86 enriched terms for CVD mortality, at a cut-off of nominal p-value 0.05, none of them significant when adjusting for multiple comparisons using the FDR approach. The GO terms with nominal p-value < 0.1 are showed in Appendix B, Tables B4 and B5. Most of the top GO terms were related to immune function for CVD mortality and to gene silencing for CVD incidence. In the KEGG analysis, no pathways were enriched for CVD incidence, while 12 pathways were enriched for CVD mortality at a 0.05 nominal p-value significance threshold, including a diabetes mellitus pathway (Appendix B, Table B6).

Cross referencing with the EWAS Catalog, 17 of the 29 DMPs that were significant in the mediation analysis for either CVD incidence or mortality showed previous associations with other traits (Appendix B, Table B7). The most frequently found traits were type II diabetes, smoking, and alcohol consumption.

We next investigated whether DNA methylation marks were conserved in a mouse model of early-life arsenic exposure. We first interrogated DMRs within the 29 genes that showed significant indirect effects in the mediation analysis and were present in the animal model. We observed most (20 out of 29 DMRs) were related to arsenic-induced atherosclerosis in the animal model (Appendix B, Table B8). Further, we assessed whether individual DMPs within the 29 genes were significantly different between controls and arsenic-exposed mice. In this more stringent analysis, 43 (42 in males and one in females) DMPs mapped to 10 of 26 genes (Figure 3.4 and Appendix B, Table B8). Of note, six DMPs were annotated to *Lmo7* in males, but not females, correlating with more profound arsenic-induced changes in atherosclerotic plaques found in males. The gene *Nav2*, significant in the mediation analysis for CVD mortality, had eight and one differentially methylated positions for male and female, respectively.

Figure 3.4: Summary of significant differentially methylated positions in a mouse model of in utero arsenic exposure by gene element and the direction of differential methylation.



Discussion

In this population-based study of American Indian adults chronically exposed to arsenic in drinking water, our extended SIS R package enabled us to identify differentially methylated CpG sites associated with CVD incidence and mortality. Furthermore, among 70 and 72 DMPs associated with CVD incidence and mortality, respectively, and 315 previously associated with arsenic in the SHS [49], we found significant mediated effects for 21 and 15 DMPs for CVD incidence and mortality, with up to 41 % of individual mediated effects. Among the 21 DMPs associated with arsenic-mediated incident CVD, six of them were associated with incident CVD in the same direction in three independent cohorts. In MESA, the only cohort with arsenic measured in a subset, despite the small sample size, the direction of association between arsenic and CVD was replicated in 13 of the 21 DMPs

(N=896), and three DMPs were associated with urinary arsenic levels (N=206).

The biological functions of genes annotated to the significant DMPs in the mediation analysis are relevant for CVD development and provide additional supportive evidence on the potential role of inorganic arsenic exposure on CVD through DNA methylation. However, our mediation analysis was conducted considering CpG sites one by one rather than all together. The fact that the sum of each individual mediated percentage (% cases or deaths attributable to a doubling of urinary arsenic explained by DNA methylation, in Tables 3.5 and Appendix, Table B3) goes beyond 100 % reflects that several pathways might be intertwined and individual mediated effects might be inflated for some CpG sites. To overcome this limitation, in section 4, we present a multimediator algorithm that conducts mediation analysis in presence of correlated mediators and is able to identify both the individual and the joint mediated effect taking correlations between mediators into account.

In addition to diabetes, the EWAS catalog linked some DMPs with smoking and alcohol intake. Smoking is a known source of arsenic [182], although it is generally not the main source. Some alcoholic beverages are known to contain arsenic, however, the estimated amount of arsenic exposure via those beverages is low [183]. The EWAS catalog did not identify DMPs associated to other traits. However, this catalog is not balanced as no blood DNA methylation EWAS have been conducted for variables that might be important for arsenic-induced CVD, such as hypertension. Hypertension is one of the most important risk factors for CVD, and it has been associated with arsenic [35]. In our mediation analysis, the results did not change when adjusting for hypertension treatment and systolic blood pressure. Other EWAS are needed to evaluate the potential role of hypertension in arsenic-induced CVD.

Some of the genes in our mediation analysis have been evaluated as therapeutic targets for CVD. Mutations in the gene *TGFBR1* have been associated with aortic diseases [184, 185] and perturbations in cardiovascular development [186]. This gene has also been proposed

as a prognostic biomarker after myocardial infarction [187]. The DMP annotated to *TYMP* was consistently inversely associated with CVD in the four populations. *TYMP* encodes an angiogenic factor which promotes angiogenesis in vivo and contributes to endothelial cells growth in vitro. Platelets are a major source of TYMP and platelet-mediated clot formation is a key process for several types of CVD [188]. The *ADAR2* gene, from the ADAR gene family, has been suggested to play a vital role in preventing cardiovascular defects [189].

A recent study conducted in the same mouse model used for replication in this work showed that an in utero and early-life arsenic exposure can enhance atherosclerosis later in life in apoE-/- mice [190]. Comparing the DNA methylation data from the livers harvested in that study to the top hits from our population-based study, we observed differential DNA methylation in the genes of interest. The fact that these DMPs and DMRs are validated in a different tissue (blood vs. liver) that is equally important to CVD, in particular in the context of cardiometabolic disease, provides supporting evidence of a potential causal relationship between arsenic-induced DNA methylation changes and atherosclerosis.

One of the methodological strengths of this work is the implementation of ISIS–Aenet to evaluate the association of DNA methylation with CVD. ISIS has proven to be very efficient for variable selection, reducing the FDR. It has been used in other studies paired with other shrinkage methods such as LASSO or elastic-net, however, to our knowledge, this is the first study that has incorporated Aenet to the ISIS algorithm for a survival problem. Of note, in Tables B1 and B2 from Appendix B, which show the DMPs selected by ISIS for CVD incidence and CVD mortality, respectively, several bootstrap CIs include the null value of 1. This means that, when repeatedly applying Aenet to the feature set previously selected by ISIS-Aenet, the coefficients went in opposite directions, which results in less evidence that that feature is truly associated with the outcome. We decided to keep those features as selected given that the variable selection process of ISIS-Aenet already lowers the dimensionality substantially.

Other strengths of this study include replication in three indepen-

dent cohorts and in an animal model, having methylation data in one of the largest microarrays available with nowadays technology (850K), the prospective study design, and the high quality of the study protocol and CVD ascertainment, as well as urinary arsenic measurements.

This work has some limitations. First, water arsenic levels changed a few years after the implementation of the US EPA final arsenic rule in 2006 [167]. However, the SHS does not have updated information on urinary arsenic levels in recent years, and data from Chile support that CVD incidence changes a few years after exposure changes [191]. Longitudinal studies with repeated measurements of arsenic and DNA methylation are needed to assess the reduction of CVD risk after arsenic exposure decreases. Second, DNA methylation is highly cell-type specific and results from blood cells might not be comparable to DNA methylation in other tissues. Blood DNA methylation, however, is emerging as a relevant tissue for CVD, probably because many of the immune cells in blood are involved in CVD pathogenesis. Also, it is unknown if CpG sites in human blood are comparable to mouse liver cells; indeed, there is limited homology between human and murine CpG sites. A genetically-modified mouse that induces hyperlipidemia had to be used, as wild-type mice do not develop atherosclerosis, even on a high-fat diet. Thus, arsenic exposure cannot be studied in the absence of hyperlipidemia. This model might be well suited for the populations we studied such as SHS and MESA, but may not be representative for populations exposed to arsenic in Bangladesh and other parts of the world where high-fat diets are less common.

In conclusion, differential methylation of CpG sites annotated to genes relevant for arsenic-related health effects might be part of the biological link between inorganic arsenic exposure and CVD. Diabetes might be a relevant mechanism for arsenic-induced cardiovascular risk in populations with a high diabetes burden, or alternatively arsenic and diabetes might share common pathways for CVD. Replication was observed for several DMPs across diverse US populations. The interspecies comparison supports that arsenic exposure modifies methylation of the same genes in the liver of an animal model of atherosclerosis compared to unexposed animals. Additional experimental studies are

needed to assess whether changes in these epigenetic signatures depending on arsenic exposure influence CVD development.

CHAPTER 4

Mediation analysis for uncausally correlated mediators in the context of survival analysis

In section 2.4, we provide an introduction to simple mediation analysis, in which only one mediator is present. Nevertheless, the fact that the effect of an exposure on an outcome will happen through only one mediating feature is unlikely in practice. The setting in which several mediators, or underlying biological pathways, exist from one variable to another is more plausible. Some work has been conducted for settings in which multiple mediators exist [192, 193, 194, 195].

The identification of the joint indirect effect for all mediators is straightforward. However, individual indirect effects cannot be identified using traditional methods in presence of correlated mediators. Jerolon et al. [9] recently developed a quasi-bayesian algorithm to conduct multiple mediation analysis in the setting of uncausally correlated mediators. They implemented this algorithm in the R package *multimediate* for continuous and binary outcomes.

On the other hand, one of the main advantages of the counterfactual mediation framework [196] as compared to traditional mediation

methods such as the difference of coefficients and the product of coefficients methods [197] is that it provides valid estimates even in presence of interactions between the exposure and the mediator [198].

The study of the effect of exposure or treatment variables on time-to-event outcomes (i.e. survival outcomes) is a common research question in epidemiology. Cox proportional hazards models are widely used in epidemiologic research. However, due to the lack of collapsibility of the hazard ratio [143], these models are not, in general conditions, the most suitable for mediation analysis. Conversely, most of the literature of mediation analysis in survival settings relies on additive hazards models [199].

In this work, we extend the *multimediate* R package to the survival setting, and provide the generalization to survival analysis of the theoretical results proved by Jerolon et al. [9] for continuous and binary outcomes. We additionally adapted the multimediate algorithm to accommodate exposure-mediator interactions. The code for the extension to survival outcomes is available in the Github repository *https://github.com/AllanJe/multimediate*. The code enabling exposure-mediator interactions will soon be available in the same repository.

We also present two data applications of the multimediate algorithm. In the first application (section 4.4.1), we applied the algorithm to simulated data in order to compare the results obtained using the multimediate algorithm to those obtained using simple mediation. The second application (section 4.4.2) is an epidemiologic study in which we aimed to study the potential mediating role of several DMPs on the association between smoking and cancer.

4.1 Multiple mediation analysis

Imai and Yamamoto [194] extended the effect definition for simple mediation analysis to the multiple mediators setting. Let us assume that $Z = (M_1, ..., M_K)^T$ is the vector of all mediators, with $K \geq 2$. Considering M_k as the mediator of interest, $k = 1, ..., K$, let us define W_k as the vector of all mediators except M_k. We also consider $Y(e^*, M_k(e), W_k(e^*))$ as the counterfactual outcome, i.e., the value the outcome would take had the exposure been set to e^*, the mediator of interest been set to the value it would take when the exposure is set to e and the other mediators been set to the value they would take when the exposure is set to e^*. In the multiple mediator setting, with $K \geq 2$ mediators, the average mediated effect of the k-th mediator is given by:

$$\delta_k(e) = \mathbb{E}\left[Y(e^*, M_k(e), W_k(e^*))|X = x\right] - \mathbb{E}\left[Y(e^*, M_k(e^*), W_k(e^*))|X = x\right],$$

being X the covariate vector. The joint indirect effect of all mediators is defined as:

$$\delta_Z(e) = \mathbb{E}\left[Y(e^*, Z(e))|X = x\right] - \mathbb{E}\left[Y(e^*, Z(e^*))|X = x\right].$$

The direct effect is defined as:

$$\zeta(e) = \mathbb{E}\left[Y(e, Z(e))|X = x\right] - \mathbb{E}\left[Y(e^*, Z(e))|X = x\right].$$

Last, the total effect is defined as:

$$\tau(e) = \zeta(e) + \delta_Z(e) = \mathbb{E}\left[Y(e, Z(e))|X = x\right] - \mathbb{E}\left[Y(e^*, Z(e^*))|X = x\right].$$

Jerolon et al. [9] defined the direct and indirect effects for con-

tinuous and binary outcomes in multiple mediation settings with uncausally correlated mediators. As in simple mediation analysis, in order for the direct, indirect and total effects to be identifiable in multiple mediators settings, several assumptions need to hold. The authors rely on the following hypothesis.

4.1.1 Sequential Ignorability for Multiple Mediators Assumptions (SIMMA)

We define $Y(e, m, w)$ as the value the outcome would take when the exposure is set to e and the mediator is set to m. The SIMMA hypothesis are the following:

1. $\{Y(e, m, w), M(e^*), W(e^{**})\} \perp E|X = x$.
2. $Y(e^*, m, w) \perp (M(e), W(e))|E = e, X = x$
3. $Y(e, m, w) \perp (M(e^*), W(e))|E = e, X = x$

In addition, the authors assume both the positivity assumption: $P(E = e|X = x) > 0$ and $P(M = m, W = w|E = e, X = x) > 0$ $\forall x, e, e^*, m, w$; and the Stable Unit Treatment Value Assumption (SUTVA), or no-interference assumption, which implies that:

1. Potential mediator and outcome values of individual i are not dependent on exposures of other individuals, i.e.: $M_{ik}(E) = M_{ik}(E_i)$ and $Y_i(E, M_k, W_k) = Y_i(E_i, M_{ik}, W_{ik})$.
2. There are no multiple versions of exposures, i.e. $E_i = E_i^*$ implies $M_{ik}(E_i) = M_{ik}(E_i^*)$ and
   $Y_i(E_i, M_{ik}(E_i), W_{ik}(E_i)) = Y_i(E_i^*, M_{ik}(E_i^*), W_{ik}(E_i^*))$.
3. There are no multiple versions of mediators, i.e. if $M_{ik} = M_{ik}^*$, then $Y_i(E_i, M_{ik}, W_{ik}) = Y_i(E_i, M_{ik}^*, W_{ik})$.

4.1.2 Multiple mediation analysis for continuous outcomes

In the case of continuous outcomes and K independent or uncausally correlated mediators, Jerolon et al. [9] assume the following linear models for both the mediators and the outcome:

$$\begin{cases} Z(E,X) = \alpha_0 + \alpha_1 E + \alpha_2 X + \epsilon_1 \\ Y(E,X,Z) = \lambda_0 + \lambda_1 E + \lambda_2^T X + \lambda_3^T Z + \epsilon_2 \end{cases}$$

where $\alpha_0, \alpha_1, \lambda_3 \in \mathbb{R}^K$, $\alpha_2 \in \mathbb{R}^K \times \mathbb{R}^p$, $\lambda_2 \in \mathbb{R}^p$, $\lambda_0, \lambda_1 \in \mathbb{R}, \epsilon_1 \sim \mathcal{N}_K(0,\Sigma)$ is the vector of residuals with covariance matrix $\Sigma \in \mathbb{R}^K \times \mathbb{R}^K$, and $\epsilon_2 \sim \mathcal{N}_K(0,\sigma^2)$, with $\sigma^2 \in \mathbb{R}$.

Under SIMMA, *Corolary 3.2* in Jerolon et al. [9] shows that the indirect effect of the k-th mediator is given by:

$$\delta_k(e) = \lambda_{3k}\alpha_{1k}(e - e^*).$$

In addition, the joint indirect effect of all mediators is given by:

$$\delta_Z(e) = \sum_{k=1}^{K} \delta_k(e).$$

Last, the direct effect is given by:

$$\zeta(e) = \lambda_1(e - e^*).$$

4.1.3 Multiple mediation analysis for binary outcomes

In the case of binary outcomes and K independent or uncausally correlated mediators, Jerolon et al. [9] assume linear models for the

mediators and a logistic or probit model for the outcome. Assuming a logistic regression model for the outcome:

$$\begin{cases} Z(E,X) = \alpha_0 + \alpha_1 E + \alpha_2 X + \epsilon_1 \\ Y^*(E,X,Z) = \lambda_0 + \lambda_1 E + \lambda_2^T X + \lambda_3^T Z + \epsilon_2 \end{cases}$$

where $Y = 1_{\{Y^*>0\}},\ \alpha_0, \alpha_1, \lambda_3 \in \mathbb{R}^K,\ \alpha_2 \in \mathbb{R}^K \times \mathbb{R}^p,\ \lambda_2 \in \mathbb{R}^p,\ \lambda_0, \lambda_1 \in \mathbb{R}$, and $\epsilon_1 \sim \mathcal{N}_K(0, \Sigma)$ is the vector of residuals with covariance matrix $\Sigma \in \mathbb{R}^K \times \mathbb{R}^K$. For logistic regression, $Y^* = logit(Pr(Y = 1|E, X, Z))$ and $\epsilon_2 \sim logit(0, 1)$.

Under SIMMA, *Corolary 3.3* in Jerolon et al. [9] shows that, in the case of logistic regression, the indirect effect of the k-th mediator is given by:

$$\begin{aligned} \delta_k(e) \;=\; & \int_{\mathbb{R}^p} F_U\Bigg(\Bigg(\lambda_0 + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{0j}\Bigg) + \Bigg(\lambda_1 + \textstyle\sum_{j=1,j\neq k}^k \lambda_{3j}\alpha_{1j}\Bigg)e^* + \lambda_{3k}\alpha_{1k}e + \\ & \Bigg(\lambda_2 + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{2j}\Bigg)x\Bigg) - F_U\Bigg(\Bigg(\lambda_0 + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{0j}\Bigg) + \Bigg(\lambda_1 + \textstyle\sum_{j=1,j\neq k}^K \lambda_{3j}\alpha_{1j}\Bigg)e^* + \\ & \lambda_{3k}\alpha_{1k}e^* + \Bigg(\lambda_2 + \sum_{j=1}^K \lambda_{3j}\alpha_{2j}\Bigg)x\Bigg)dF_X(x). \end{aligned}$$

In addition, the joint indirect effect of all mediators is given by:

$$\begin{aligned} \delta_Z(e) \;=\; & \int_{\mathbb{R}^p} F_U\Bigg(\Bigg(\lambda_0 + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{0j}\Bigg) + \lambda_1 e^* + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{1j}e + \Bigg(\lambda_2 + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{2j}\Bigg)x\Bigg) - \\ & F_U\Bigg(\Bigg(\lambda_0 + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{0j}\Bigg) + \lambda_1 e^* + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{1j}e^* + \Bigg(\lambda_2 + \textstyle\sum_{j=1}^K \lambda_{3j}\alpha_{2j}\Bigg)x\Bigg)dF_X(x). \end{aligned}$$

Last, the direct effect is given by:

$$\zeta(e) = \int_{\mathbb{R}^1} F_U\left(\left(\lambda_0 + \sum_{j=1}^K \lambda_{3j}\alpha_{0j}\right) + \lambda_1 e + \sum_{j=1}^K \lambda_{3j}\alpha_{1j} e + \left(\lambda_2 + \sum_{j=1}^K \lambda_{3j}\alpha_{2j}\right)x\right) -$$

$$F_U\left(\left(\lambda_0 + \sum_{j=1}^K \lambda_{3j}\alpha_{0j}\right) + \lambda_1 e^* + \sum_{j=1}^K \lambda_{3j}\alpha_{1j} e + \left(\lambda_2 + \sum_{j=1}^K \lambda_{3j}\alpha_{2j}\right)x\right) dF_X(x),$$

where

$$F_U(z) = \int_{\mathbb{R}} \Phi\left(\frac{z-\epsilon_2}{\sqrt{\sum_{k=1}^K \sum_{j=1}^K \lambda_{3k}\lambda_{3j}\ cov(\epsilon_{1k},\epsilon_{1j})}}\right) \frac{e^{\epsilon_2}}{(1+e^{\epsilon_2})^2}\ d\epsilon_2.$$

The proof of these expressions, as well as the equivalent expressions for probit regression, can be found in Jerolon et al. [9]. In the following section, we extend these results to the case of survival outcomes.

4.2 Multiple mediation in survival analysis

4.2.1 Effect definition

Following Lange and Hansen [139], we define the indirect effect of the mediator $M_k, k = 1, ..., K$ changing the exposure from e^* to e as:

$$\delta_k(e) = \gamma(t; e^*, M_k(e), W_k(e^*), X) - \gamma(t; e^*, M_k(e^*), W_k(e^*), X),$$

being γ the hazard, or rate, function, which is given, for each (e, e^*, e^{**}), by:

$$\gamma(t; e, M_k(e^*), W_k(e^{**})) = \lim_{dt \to 0} \frac{1}{dt} P(T(e, M_k(e^*), W_k(e^{**})) \in [t, t + dt] \mid T(e, M_k(e^*), W_k(e^{**})) \geq t).$$

We define the joint indirect effect of all mediators as:

$$\delta_Z(e) = \gamma(t; e^*, Z(e), X) - \gamma(t; e^*, Z(e^*), X).$$

The direct effect is defined as:

$$\zeta(e) = \gamma(t; e, Z(e), X) - \gamma(t; e^*, Z(e), X).$$

Last, the total effect is defined as:

$$\tau(e) = \gamma(t; e, Z(e), X) - \gamma(t; e^*, Z(e^*), X).$$

By the above definitions, $\tau(e) = \delta_Z(e) + \zeta(e)$.

4.2.2 Hypothesis

Adapting Lange and Hansen's hypothesis [139] to Jerolon et al.'s notation [9], the following set of SIMMA assumptions is obtained. Let us consider $T(e, m, w)$ as the time to event when the exposure is set to e, the mediator of interest is set to m and the other mediators are set to w.

1. $E \perp (T(e, m, w),\ M_k(e^*), W_k(e^{**})) \mid X, \quad \forall k = 1, ..., K.$
2. $T(e^*, m, w) \perp Z(e) \mid X, E.$
3. $T(e, m, w) \perp (M_k(e^*), W_k(e^{**})) \mid X, E, \quad \forall k = 1, ..., K.$
4. $M_k(E) = M_k,\ W_k(E) = W_k,\ T(E, Z) = T.$

We also assume that $P(E = e|X = x) > 0$ and $P(M = m, W = w|E = e, X = x) > 0\ \forall\ e, e^*, x, m, w$; and that SUTVA holds.

In addition to SIMMA and SUTVA, we assume that the mediators follow a multivariate multiple linear normal homoscedastic model, and that hazard functions follow the additive risk model, with time independent coefficients. Therefore, the outcome and mediator models in survival settings with multiple mediators are defined as follows:

$$\begin{cases} Z(E, X) = \alpha_0 + \alpha_1 E + \alpha_2 X + \epsilon \\ \gamma(t; E, X, Z) = \lambda_0(t) + \lambda_1 E + \lambda_2^T X + \lambda_3^T Z \end{cases} \tag{4.1}$$

where $\alpha_0, \alpha_1, \lambda_3 \in \mathbb{R}^K$, $\alpha_2 \in \mathbb{R}^K \times \mathbb{R}^p$, $\lambda_2 \in \mathbb{R}^p$, $\lambda_1 \in \mathbb{R}$, $\lambda_0(t)$ is the time-varying baseline hazard and $\epsilon \sim \mathcal{N}_K(0, \Sigma)$ is the error vector of the multivariate linear regression, with covariance matrix $\Sigma \in \mathbb{R}^K \times \mathbb{R}^K$.

We also assume, following Jerolon et al. [9], that, either the mediators are independent, or the correlations between the k mediators are

not causal, i.e., that the dependence between them does not have a causal order. In this latter case, we assume that pairwise correlations between mediators are independent of the exposure:

$$cor(M_i(e), M_j(e^*)|E, X) = \rho_{ij},\ \forall e, e^* \in \{0,1\},\ \forall i, j \in 1, ..., k.$$

4.2.3 Main theoretical results

Proposition 1 *Under the previous conditions, it holds that the hazard function takes the following value:*

$$\gamma(t; e, M_k(e^*), W_k(e^{**})) = C(t) + \lambda_1 e + \lambda_{3k}\alpha_{1k}e^* + \sum_{j\neq k}^{K} \lambda_{3j}\alpha_{1j}e^{**},$$

$\forall (e, e^*, e^{**}) \in \{0,1\}^3$, *being* $C(t)$ *a function that does not depend on the exposure values* e, e^* *or* e^{**}.

Proof 1 (Proof of Proposition 1) *Without loss of generality, we consider* M_1 *as the mediator of interest. Let us call* T^* *the random variable* $T(e, M_1(e^*), W_1(e^{**}))$, *being* $(e, e^*, e^{**}) \in \{0,1\}^3$. *Then, the rate can be expressed as:*

$$\gamma(t; e, M_1(e^*), W_1(e^{**})) = \lim_{dt\to 0} \frac{1}{dt}\ P(T^* \in [t, t+dt] \mid T^* \geq t).$$

It holds that:

$$P(T^* \in [t, t+dt] \mid T^* \geq t) = \mathbb{E}_{\{X|T^*\geq t\}}[P(T^* \in [t, t+dt] \mid X = x, T^* \geq t)]$$

and, similarly, being $F(m_1, w_1)$ *the distribution function of* $M_1(e^*), W_1(e^{**})$, *given that* $X = x$ *and* $T^* \geq t$:

$$\begin{aligned} P(T^* \in [t, t+dt] | X = x, T^* \geq t) &= \int_{\mathbb{R}^K} P(T^* \in [t, t+dt] | X = x, M_1 = m_1, W_1 = w_1, T^* \geq t) dF(m_1, w_1) \\ &= \int_{\mathbb{R}^K} P(T(e, m_1, w_1) \in [t, t+dt] | X = x, M_1 = m_1, W_1 = w_1, T(e, m_1, w_1) \geq t) dF(m_1, w_1) \\ &= \int_{\mathbb{R}^K} P(T(e, m_1, w_1) \in [t, t+dt] | X = x, E = e, M_1 = m_1, W_1 = w_1, T(e, m_1, w_1) \geq t) dF(m_1, w_1). \end{aligned}$$

Using the bounded convergency theorem and the additive risk hypothesis:

$$\begin{aligned} \gamma(t; e, M_1(e^*), W_1(e^{**})) &= \mathbb{E}_{\{X|T^* \geq t\}} \Big[\int_{\mathbb{R}^k} (\lambda_0(t) + \lambda_1 e + \lambda_2^T x + \lambda_3^T (m_1, w_1^T)^T) \; dF(m_1, w_1) \Big] \\ &= \lambda_0(t) + \lambda_1 e + \lambda_2^T \; \mathbb{E}(X | T^* \geq t) + \mathbb{E}_{\{X|T^* \geq t\}} \Big[\int_{\mathbb{R}^K} \lambda_3^T (m_1, w_1^T)^T \; dF(m_1, w_1) \Big]. \end{aligned}$$

In addition,

$$\begin{aligned} \int_{\mathbb{R}^K} \lambda_3^T (m_1, w_1^T)^T dF(m_1, w_1) &= \int_{\mathbb{R}^K} \lambda_3^T (m_1, w_1^T)^T \; f(M_1(e^*) = m_1, W_1(e^{**}) = w_1 \mid X = x, T^* \geq t) \; dm_1 dw_1 \\ &= \int_{\mathbb{R}^K} \lambda_3^T (m_1, w_1^T)^T \; \frac{P(T^* \geq t \mid M_1 = m_1, W_1 = w_1, X = x) \; f(M_1(e^*) = m_1, W_1(e^{**}) = w_1 \mid X = x)}{P(T^* \geq t | X = x)} \; dm_1 dw_1. \end{aligned}$$

Following the same arguments and taking into account that additive hazards models have been used for a time-to-event setting:

$$\begin{aligned} P(T^* \geq t \mid M_1 = m_1, W_1 = w_1, X = x) &= P(T(e, m_1, w_1) \geq t \mid M_1 = m_1, W_1 = w_1, X = x, E = e) \\ &= exp\{ - \textstyle\int_0^t \lambda_0(u) du - \lambda_1 et - \lambda_2^T xt - \lambda_3^T (m_1, w_1^T)^T t \}. \end{aligned}$$

Also,

$$\begin{aligned} P(T^* \geq t | X = x) &= \int_{\mathbb{R}^K} P(T^* \geq t \mid M_1 = m_1, W_1 = w_1, X = x) \; dF(m_1, w_1) \\ &= exp\Big\{ - \textstyle\int_0^t \lambda_0(u) du - \lambda_1 et - \lambda_2^T xt \Big\} \; \mathbb{E} \; (exp\{ - \lambda_3^T (m_1, w_1^T)^T t \}) \end{aligned}$$

Hence, defining $V = \lambda_3^T (m_1, w_1^T)^T$ *and putting together the above results, it holds that:*

$\int_{\mathbb{R}^K} \lambda_3^T (m_1, w_1^T)^T \; dF(m_1, w_1) = \int_{\mathbb{R}^K} V \; dF(m_1, w_1) = \frac{\mathbb{E}(V exp\{-tV\} \mid X=x)}{\mathbb{E}(exp\{-tV\} \mid X=x)}$.

The distribution of $(M_1(e^*), W_1(e^{**}) \mid X = x)$ *is multivariate normal [9] with covariance matrix* Σ*, which does not depend on the exposures, and with expected values:*

$$\mathbb{E}(M_1(e^*) \mid X = x) = \alpha_{01} + \alpha_{11} e^* + \alpha_{21} x$$

$$\mathbb{E}(M_j(e^{**}) \mid X = x) = \alpha_{0j} + \alpha_{1j} e^{**} + \alpha_{2j} x, \;\; j = 2, ..., K.$$

Thus, the distribution of V *is normal with:*

$$\mathbb{E}(V \mid X = x) = \lambda_3^T \alpha_0 + \lambda_{31} \alpha_{11} e^* + \sum_{j=2}^{K} \lambda_{3j} \alpha_{1j} e^{**} + \lambda_3^T \alpha_2 x$$

$$Var(V \mid X = x) = \lambda_3^T \Sigma \lambda_3.$$

Following Lange and Hansen [139],

$$\begin{aligned} \frac{\mathbb{E}(V exp\{-tV\}|X=x)}{\mathbb{E}(exp\{-tV\}|X=x)} &= \mathbb{E}(V|X=x) - tVar(V|X=x) \\ &= \lambda_3^T \alpha_0 + \lambda_{31} \alpha_{11} e^* + \sum_{j=2}^{K} \lambda_{3j} \alpha_{1j} e^{**} + \lambda_3^T \alpha_2 x - t \lambda_3^T \Sigma \lambda_3, \end{aligned}$$

and it holds that the counterfactual rate can be expressed as:

$$\gamma(t; e, M_1(e^*), W_1(e^{**})) = C(t) + \lambda_1 e + \lambda_{31} \alpha_{11} e^* + \sum_{j=2}^{K} \lambda_{3j} \alpha_{1j} e^{**},$$

being

$$C(t) = \lambda_0(t) + \lambda_2^T \, \mathbb{E}(X \mid T^* \geq t) + \lambda_3^T \alpha_0 + \lambda_3^T \alpha_2 \, \mathbb{E}(X \mid T^* \geq t) - t\lambda_3^T \Sigma \lambda_3$$

a function of t that does not depend on the exposures.

The proof would be equivalent for any of the K mediators.

Once the hazard function is obtained, the following theorem shows how to obtain the different effects.

Theorem 1 *Under the conditions described in proposition 1, it holds that the indirect effect of the mediator $M_k, k = 1, ..., K$ changing the exposure from e^* to e is:*

$$\delta_k(e) = \gamma(t; e^*, M_k(e), W_k(e^*), X) - \gamma(t; e^*, M_k(e^*), W_k(e^*), X) = \lambda_{3k}\alpha_{1k}(e - e^*)$$

Moreover, the joint indirect effect of all mediators Z is the sum of individual mediated effects:

$$\delta_Z(e) = \gamma(t; e^*, Z(e), X) - \gamma(t; e^*, Z(e^*), X) = \sum_{j=1}^{K} \lambda_{3j}\alpha_{1j}(e - e^*)$$

The direct effect is:

$$\zeta(e) = \gamma(t; e, Z(e), X) - \gamma(t; e^*, Z(e), X) = \lambda_1(e - e^*),$$

and the total effect equals the sum of the joint indirect effect and the direct effect:

$$\tau(e) = \gamma(t; e, Z(e), X) - \gamma(t; e^*, Z(e^*), X) = (\sum_{j=1}^{k} \lambda_{3j}\alpha_{1j} + \lambda_1)(e - e^*)$$

Please note that, if we consider $e^* = 0$ *and* $e = 1$*, the factor* $(e - e^*)$ *can be removed in all formulas. In addition, please note that* $\delta_k(1) = -\delta_k(0)$*,* $\delta_Z(1) = -\delta_Z(0)$*,* $\zeta(1) = -\zeta(0)$ *and* $\tau(1) = -\tau(0)$*.*

Proof 2 (Proof of Theorem 1) *Without loss of generality, we consider* M_1 *as the mediator of interest. The effect mediated by the* M_1 *mediator would be given by:*

$$\begin{aligned}
\delta_1(e) &= \gamma(t; e^*, M_1(e), W_1(e^*), X) - \gamma(t; e^*, M_1(e^*), W_1(e^*), X) \\
&= C(t) + \lambda_1 e^* + \lambda_{31}\alpha_{11} e + \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e^* - C(t) - \lambda_1 e^* - \lambda_{31}\alpha_{11} e^* - \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e^* \\
&= \lambda_{31}\alpha_{11}(e - e^*).
\end{aligned}$$

The effect mediated by all mediators $M_1, ..., M_K$ *would be:*

$$\begin{aligned}
\delta_Z(e) &= \gamma(t; e^*, Z(e), X) - \gamma(t; e^*, Z(e^*), X) \\
&= C(t) + \lambda_1 e^* + \lambda_{31}\alpha_{11} e + \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e - C(t) - \lambda_1 e^* - \lambda_{31}\alpha_{11} e^* - \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e^* \\
&= \sum_{j=1}^{K} \lambda_{3j}\alpha_{1j}(e - e^*),
\end{aligned}$$

which equals the sum of the mediated effects for each of the mediators.

The direct effect would be:

$$\begin{aligned}\zeta(e) &= \gamma(t; e, Z(e), X) - \gamma(t; e^*, Z(e), X) \\ &= C(t) + \lambda_1 e + \lambda_{31}\alpha_{11} e + \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e - C(t) - \lambda_1 e^* - \lambda_{31}\alpha_{11} e - \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e \\ &= \lambda_1(e - e^*)\end{aligned}$$

Last, the total effect would be:

$$\begin{aligned}\tau(e) &= \gamma(t; e, Z(e), X) - \gamma(t; e^*, Z(e^*), X) \\ &= C(t) + \lambda_1 e + \lambda_{31}\alpha_{11} e + \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e - C(t) - \lambda_1 e^* - \lambda_{31}\alpha_{11} e^* - \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} e^* \\ &= \Big(\lambda_1 + \sum_{j=1}^{K} \lambda_{3j}\alpha_{1j}\Big)(e - e^*)\end{aligned}$$

which equals the sum of the direct and the indirect effects. The proof would be equivalent for any of the K mediators.

In addition to extending the multimediate algorithm to survival settings, we adapted this algorithm to accommodate exposure-mediator interactions. Let us consider the following mediator and outcome models, which are the ones used in (4.1), but introducing an interaction term between the exposure and the k-th mediator.

$$\begin{cases} Z(E, X) = \alpha_0 + \alpha_1 E + \alpha_2 X + \epsilon \\ \gamma(t; E, X, Z) = \lambda_0(t) + \lambda_1 E + \lambda_2^T X + \lambda_3^T Z + \lambda_4 E M_k \end{cases} \tag{4.2}$$

Please note that, in presence of an interaction between the exposure and one of the mediators, the effects would be different for different strata of the exposure or treatment. We hereby provide an extension of theorem 1 in presence of exposure-mediator interactions.

The risk function in (4.2) is identical to that in (4.1) when $E = 0$. When $E = 1$, the coefficient λ_{3k} becomes $\lambda_{3k} + \lambda_4$. Thus, following proposition 1, provided there are interactions between the exposure and the k-th mediator, it holds that:

$$\gamma(t; 1, M_k(e^*), W_k(1), X) = C(t) + \lambda_1 + \alpha_{1k}(\lambda_{3k} + \lambda_4)e^* + \sum_{j \neq k}^{K} \lambda_{3j}\alpha_{1j},$$

$$\gamma(t; 0, M_k(e^*), W_k(0), X) = C(t) + \alpha_{1k}\lambda_{3k}e^*,$$

$$\gamma(t; 1, Z(e^*), X) = C(t) + \lambda_1 + \alpha_{1k}(\lambda_{3k} + \lambda_4)e^* + \sum_{j \neq k}^{K} \lambda_{3j}\alpha_{1j}e^*,$$

and

$$\gamma(t; 0, Z(e^*), X) = C(t) + \sum_{j=1}^{K} \lambda_{3j}\alpha_{1j}e^*,$$

which leads to the following corollary:

Corollary 1 *Under the hypothesis of theorem 1, in presence of an interaction between the exposure and the k-th mediator, it holds that the indirect effect is:*

$$\delta_k(1) = \lambda_{3k}\alpha_{1k}$$

$$\delta_1(0) = -(\lambda_{3k} + \lambda_4)\alpha_{1k}$$

Likewise, the joint indirect effect of all mediators is:

$$\delta_Z(1) = \sum_{j=1}^{K} \lambda_{3j}\alpha_{1j}$$

$$\delta_Z(0) = -(\lambda_{3k} + \lambda_4)\alpha_{1k} - \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j}$$

The direct effect is:

$$\zeta(1) = \lambda_1 + \lambda_4\alpha_{1k}$$

$$\zeta(0) = -\lambda_1$$

and the total effect is:

$$\tau(1) = \lambda_1 + (\lambda_{3k} + \lambda_4)\alpha_{1k} + \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j}$$

$$\tau(0) = -\lambda_1 - (\lambda_{3k} + \lambda_4)\alpha_{1k} - \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j}$$

Proof 3 (Proof of Corollary 1) *Without loss of generality, we consider M_1 as the mediator of interest. The effect mediated by the M_1 mediator would be given by:*

$$\begin{aligned} \delta_1(1) &= \gamma(t; 0, M_1(1), W_1(0), X) - \gamma(t; 0, M_1(0), W_1(0), X) \\ &= C(t) + \lambda_{31}\alpha_{11} - C(t) = \lambda_{31}\alpha_{11} \end{aligned}$$

$$\begin{aligned} \delta_1(0) &= \gamma(t; 1, M_1(0), W_1(1), X) - \gamma(t; 1, M_1(1), W_1(1), X) \\ &= C(t) + \lambda_1 + \textstyle\sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} - C(t) - \lambda_1 - (\lambda_{31} + \lambda_4)\alpha_{11} - \sum_{j=2}^{K} \lambda_{3j}\alpha_{1j} \\ &= -(\lambda_{31} + \lambda_4)\alpha_{11} \end{aligned}$$

The effect mediated by all mediators $M_1, ..., M_K$ would be:

$$\begin{aligned}\delta_Z(1) &= \gamma(t;0,Z(1),X) - \gamma(t;0,Z(0),X) = C(t) + \lambda_{31}\alpha_{11} + \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j} - C(t) \\ &= \sum_{j=1}^{K}\lambda_{3j}\alpha_{1j}\end{aligned}$$

$$\begin{aligned}\delta_Z(0) &= \gamma(t;1,Z(0),X) - \gamma(t;1,Z(1),X) \\ &= C(t) + \lambda_1 - C(t) - \lambda_1 - (\lambda_{31} + \lambda_4)\alpha_{11} - \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j} \\ &= -(\lambda_{31} + \lambda_4)\alpha_{11} - \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j}\end{aligned}$$

The direct effect would be:

$$\begin{aligned}\zeta(1) &= \gamma(t;1,Z(1),X) - \gamma(t;0,Z(1),X) \\ &= C(t) + \lambda_1 + (\lambda_{31} + \lambda_4)\alpha_{11} + \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j} - C(t) - \lambda_{31}\alpha_{11} - \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j} \\ &= \lambda_1 + \lambda_4\alpha_{11}\end{aligned}$$

$$\zeta(0) = \gamma(t;0,Z(0),X) - \gamma(t;1,Z(0),X) = C(t) - C(t) - \lambda_1 = -\lambda_1$$

Last, the total effect would be:

$$\begin{aligned}\tau(1) &= \gamma(t;1,Z(1),X) - \gamma(t;0,Z(0),X) = C(t) + \lambda_1 + (\lambda_{31}+\lambda_4)\alpha_{11} + \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j} - C(t) \\ &= \lambda_1 + (\lambda_{31}+\lambda_4)\alpha_{11} + \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j}\end{aligned}$$

$$\begin{aligned}\tau(0) &= \gamma(t;0,Z(0),X) - \gamma(t;1,Z(1),X) = C(t) - C(t) - \lambda_1 - (\lambda_{31}+\lambda_4)\alpha_{11} - \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j} \\ &= -\lambda_1 - (\lambda_{31}+\lambda_4)\alpha_{11} - \sum_{j=2}^{K}\lambda_{3j}\alpha_{1j}\end{aligned}$$

Please note that, in this context, $\delta_k(1) \neq -\delta_k(0)$, $\delta_Z(1) \neq -\delta_Z(0)$ and $\zeta(1) \neq -\zeta(0)$, however, $\tau(1) = -\tau(0)$. Thus, the estimators of the effects should be considered separated by strata of the exposure instead of calculating average estimator effects.

4.3 Extension of the multimediate algorithm to a survival setting

We used an adapted version of the quasi-bayesian algorithm developed by Jerolon et al. [9] to obtain point estimates of the effects of interest, as well as confidence intervals and p-values. Let us consider the scenario of K mediators and n observations.

1. We fit the observed mediator model using linear regression, and the observed outcome model using the Lin-Ying model [147] fitted with the *aalen* function from the R package *timereg*, which allows to specify that all coefficients are time-invariant except the baseline hazard.

2. We estimate the covariance matrix Σ of the errors of the mediator models by extracting the residuals $\epsilon_1^k, \ldots, \epsilon_n^k$ for each of the K mediator models and computing pairwise correlations between $\epsilon_1^i, \ldots, \epsilon_n^i$ and $\epsilon_1^j, \ldots, \epsilon_n^j$ for each $i \neq j$, obtaining the matrix $\hat{\Sigma}$. This matrix will be used later to incorporate the correlations between mediators to the simulation algorithm.

3. For each parameter of each of the models, we sample J values from the multivariate sampling distribution of their maximum likelihood estimators: $\widehat{\Theta}_j^Z = (\widehat{\Theta}_j^1, ..., \widehat{\Theta}_j^K)$ for the mediator models and $\widehat{\Theta}_j^Y$ for the outcome model. For the mediator models, we use the multivariate normal distribution. For the additive model, the baseline hazard is not taken into account as all effect estimations imply a substraction in which the baseline hazard is cancelled (see section 4.2). According to Lin and Ying [147], all coefficients of the additive hazards model are also asymptotically normal. Thus, we also sample from the multivariate normal distribution for the outcome model. We use the estimates of the parameters as the mean, and the asymptotic covariance matrix between the estimators as the covariance.

4. In order to take into account the correlations between mediators, we jointly simulate the residuals of all the mediator models using

a multivariate normal distribution with mean zero and covariance matrix $\hat{\Sigma}$.

5. For each simulation $j = 1, \ldots, J$:
   (a) We calculate the counterfactual values of each mediator under each exposure or treatment. For each of the K mediators, each pair of exposures $(e, e^*) \in \{0,1\}^2$ and each individual $i = 1, \ldots n$; $Z_{ij}(e, e^*) = (M_{ik}(e), W_{ik}(e^*))$.
   (b) Given the simulated values of the counterfactual mediators, we calculate the counterfactual outcomes, i.e., for each individual $i = 1, \ldots n$ and $(e, e^*, e^{**}) \in \{0,1\}^3$, we calculate $Y_{ij}(e, Z_{ij}(e^*, e^{**})) = \gamma_{ij}(e, Z_{ij}(e^*, e^{**}))$.
   (c) We estimate the causal mediation effects. Our proposed estimators are the sample mean of the effects obtained in the previous simulation process:

   - Indirect effect for each mediator:
     $\widehat{\delta}_j^k(e) = \left(\frac{1}{n}\sum_{i=1}^n \gamma_{ij}(e^*, Z_{ij}(e, e^*)) - \gamma_{ij}(e^*, Z_{ij}(e^*, e^*))\right) * 100000$

   - Joint indirect effect:
     $\widehat{\delta}_j^Z(e) = \left(\frac{1}{n}\sum_{i=1}^n \gamma_{ij}(e^*, Z_{ij}(e)) - \gamma_{ij}(e^*, Z_{ij}(e^*))\right) * 100000$

   - Direct effect:
     $\widehat{\zeta}_j(e) = \left(\frac{1}{n}\sum_{i=1}^n \gamma_{ij}(e, Z_{ij}(e)) - \gamma_{ij}(e^*, Z_{ij}(e))\right) * 100000$

   - Total effect:
     $\widehat{\tau}_j(e) = \left(\frac{1}{n}\sum_{i=1}^n \gamma_{ij}(e, Z_{ij}(e)) - \gamma_{ij}(e^*, Z_{ij}(e^*))\right) * 100000$

   Each effect is calculated for both $e = 0$ and $e = 1$. In the previous section, we proved that, in absence of interactions, $\delta_k(1) = -\delta_k(0)$ but, in general, $\hat{\delta}_k(1) \neq -\hat{\delta}_k(0)$, as they represent two different estimators of the same parameter. In absence of interactions, we propose to use $\frac{\hat{\delta}_k(e) - \hat{\delta}_k(1-e)}{2}$ as the estimator of $\hat{\delta}_k(e)$. Similarly for direct and total effects. Please

note that we multiply each estimator by $100,000$ in order to get an estimation of the number of cases attributable to the exposure through the mediator per $100,000$ person-years (this number could be changed according to the users preferences). Also note that, for time-invariant covariates, the effects do not depend on the time t.

6. From the empirical distribution of each effect above, we calculate the estimator of the effect as well as confidence intervals. The 50-th percentile is taken as the average effect of interest, and the 2.5-th and 97.5-th percentiles of the sample distribution of each estimator are taken as the 95 % confidence intervals' lower and upper bounds, respectively.

4.4 Data applications

4.4.1 Data application 1: a simulation study

We conducted a simulation study in order to assess the performance of the multimediate algorithm in survival settings, and compare it to simple mediation analysis. For the purposes of this simulation study, we assume the setting of three mediators ($K = 3$). Following Jerolon et al.'s simulation framework [9], we first simulate a database of 10^6 observations for exposure $e \in \{0, 1\}$, for the counterfactual mediators M_1, M_2 and M_3 and the counterfactual value of the linear predictor $\Psi(E, X, Z) = \lambda_1 E + \lambda_2^T X + \lambda_3^T Z$ (hereinafter referred to as Ψ for simplicity), which equals the definition of the rate γ in additive models except for the baseline hazard, which is removed as all effect calculations require substractions and the baseline hazard is cancelled. We will subsequently use this linear predictor to calculate survival times for each individual. We then calculate the direct, indirect and total effects as described in section 4.3, substracting means of the counterfactual values of the linear predictor in different scenarios. The large size of the database guarantees that those estimates are sufficiently close to the true values of the effects. We fixed the number of simulations to 600. In each simulation, a random sample of 2000 observations of the full database is taken, and the effects of interest are calculated in that subsample.

The MSE, the bias, the variance and the % coverage of the 95 % CIs are calculated comparing the true effects (calculated in the full simulated database) to those estimated by simple mediation analysis and by the multimediate algorithm.

In order to simulate survival times, we use the inverse transformation method. Please note that the survival distribution function is $S(t) = exp(-\Lambda(t))$, being $\Lambda(t)$ the cumulative hazard function, in our case $\Lambda(t) = \int_0^t [\lambda_0(s) + \Phi(E, X, Z)]\ ds = \Lambda_0(t) + t\Phi(E, X, Z)$, where $\Lambda_0(t)$ is the cumulative baseline hazard function. Hence, a simulated time t is obtained as the solution of the equation $u = exp(-\Lambda_0(t) - t\Phi(E, X, Z))$, being u a number randomly generated from a $U(0, 1)$ distribution [200].

We consider three different scenarios for the baseline hazard: constant baseline hazard, monotonic baseline hazard dependent on time, and non-monotonic baseline hazard. In addition, we consider three different correlation scenarios for the mediators: negative correlation ($\rho = -0.4$), no correlation ($\rho = 0$) and positive correlation ($\rho = 0.4$). In the next sections, we present the results of the metrics (MSE, bias, variance and CI coverage) of the simulations in each of the scenarios.

Constant baseline hazard

We assume that the baseline hazard takes the constant value $\lambda_0 = 0.1$. Given that in this case $\Lambda_0(t) = 0.1t$, the survival time for a given individual would be simulated as:

$$t = \frac{-log(u)}{0.1 + \Psi}$$

being $u \sim U(0, 1)$. Tables 4.1, 4.2 and 4.3 show the MSE, variance and bias for the total, direct and indirect effects comparing simple mediation to the multimediate algorithm. While both frameworks present similar results for the total effect, the multimediate algorithm presents, in general, smaller MSEs for the direct and indirect effects. For the direct effect, the MSE is smaller even for the setting of no correlations. The reduction in bias of the multimediate algorithm drives the reduction in MSE.

Table 4.1: Simulation results for the total effect in a constant baseline risk scenario

	Correlation = -0.4				Correlation = 0				Correlation = 0.4			
	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim
MSE	0.23	0.23	0.23	0.23	0.26	0.26	0.26	0.27	0.25	0.25	0.25	0.25
Var	0.23	0.23	0.23	0.23	0.26	0.26	0.26	0.27	0.25	0.25	0.25	0.25
Bias	-0.033	-0.034	-0.035	-0.033	-0.029	-0.029	-0.027	-0.029	-0.012	-0.011	-0.013	-0.014

Table 4.2: Simulation results for the direct effect in a constant baseline risk scenario

	Correlation = -0.4				Correlation = 0				Correlation = 0.4			
	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim
MSE	0.68	1.79	0.72	0.71	0.51	0.95	0.50	0.36	0.39	0.47	0.35	0.31
Var	0.26	0.30	0.25	0.71	0.29	0.29	0.29	0.36	0.28	0.29	0.28	0.31
Bias	0.65	1.22	0.68	-0.014	0.47	0.81	0.46	-0.048	0.33	0.42	0.26	-0.005

Table 4.3: Simulation results for the indirect effects (simple mediation / multimediate) in a constant baseline risk scenario

	Correlation = -0.4		
	Med 1	Med 2	Med 3
MSE	0.056 / 0.049	0.21 / 0.074	0.078 / 0.063
Var	0.022 / 0.049	0.039 / 0.075	0.031 / 0.063
Bias	-0.18 / -0.019	-0.42 / -0.0027	-0.22 / 0.002
	Correlation = 0		
	Med 1	Med 2	Med 3
MSE	0.026 / 0.026	0.038 / 0.037	0.029 / 0.030
Var	0.026 / 0.026	0.038 / 0.038	0.029 / 0.030
Bias	-0.0007 / -0.003	-0.0015 / 0.0031	0.016 / 0.019
	Correlation = 0.4		
	Med 1	Med 2	Med 3
MSE	0.051 / 0.031	0.21 / 0.051	0.084 / 0.039
Var	0.025 / 0.031	0.042 / 0.051	0.034 / 0.039
Bias	0.16 / -0.011	0.40 / -0.0068	0.22 / 0.0094

Tables 4.4 and 4.5 show the empirical coverage of 95 % CIs in terms of proportions of simulations that contain the real value of the different effects (calculated in the full database of 1,000,000 observations). While the total effect has great empirical coverage for both simple mediation and the multimediate algorithm, direct and indirect effects clearly worsen their empirical coverage in simple mediation models in

settings of correlated mediators. Conversely, the multimediate algorithm remains with good and similar coverage in both correlated and uncorrelated settings.

Table 4.4: Empirical coverage of the confidence interval with theoretical coverage of 95 % (proportion of simulations including the true value) of simple mediation models in a constant baseline risk scenario

	Mediator 1		
	Indirect	Direct	Total
Correlation=-0.4	0.76	0.75	0.96
Correlation=0	0.95	0.84	0.95
Correlation=0.4	0.82	0.90	0.95
	Mediator 2		
	Indirect	Direct	Total
Correlation=-0.4	0.44	0.37	0.96
Correlation=0	0.97	0.67	0.95
Correlation=0.4	0.46	0.87	0.95
	Mediator 3		
	Indirect	Direct	Total
Correlation=-0.4	0.76	0.75	0.96
Correlation=0	0.95	0.85	0.95
Correlation=0.4	0.75	0.92	0.95

Table 4.5: Empirical coverage of the confidence interval with theoretical coverage of 95 % (in proportions of simulations) of the multimediate algorithm in a constant baseline risk scenario

	Indirect M1	Indirect M2	Indirect M3	Direct	Total
Correlation=-0.4	0.94	0.96	0.94	0.95	0.94
Correlation=0	0.94	0.95	0.95	0.93	0.93
Correlation=0.4	0.93	0.92	0.95	0.94	0.95

Monotonic baseline hazard dependent on time

We now assume that the baseline hazard takes the value $\lambda_0 = t$. Thus, the cumulative hazard function would be defined as:

$$\Lambda(t) = \int_0^t (u + \Psi)\ du,$$

and the survival function would be defined as:

$$S(t) = exp\{-(\int_0^t (u + \Psi)du)\} = exp\{-\frac{t^2}{2} - \Psi t\}$$

$$\frac{t^2}{2} + \Psi t + logU = 0 \implies t = -\Psi + \sqrt{\Psi^2 - 2logU}.$$

Please note that, given that $0 < U < 1$, it always holds that $\sqrt{\Psi^2 - 2logU} > |\Psi|$. Therefore, $-\Psi - \sqrt{\Psi^2 - 2logU}$ is not considered as a possible solution as survival times are always positive.

Tables 4.6, 4.7 and 4.7 show the MSE, variance and bias for the total, direct and indirect effects comparing simple mediation to the multimediate algorithm. A similar tendency to that of the constant baseline hazard case can be observed. Again, both frameworks present similar results for the total effect and the multimediate algorithm presents, in general, a smaller MSE for the direct effect, even in the context of no correlation between mediators. For the indirect effect, the error is again similar in the context of no correlation between mediators, and smaller for the multimediate algorithm in contexts of correlated mediators.

Table 4.6: Simulation results for the total effect in a monotonic time-dependent baseline risk scenario

	Correlation = -0.4				Correlation = 0				Correlation = 0.4			
	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim
MSE	64.8	64.9	64.5	64.1	62.1	62.5	62.5	61.4	66.3	66.5	66.3	65.4
Var	50.8	50.9	50.6	50.9	48.1	48.1	48.1	48.1	50.1	50.4	50.4	49.9
Bias	-3.7	-3.7	-3.7	-3.6	-3.7	-3.8	-3.8	-3.7	-4.0	-4.0	-4.0	-3.9

Table 4.7: Simulation results for the direct effect in a monotonic time-dependent baseline risk scenario

	Correlation = -0.4				Correlation = 0				Correlation = 0.4			
	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim
MSE	4551.9	15205.0	5191.4	5882.7	2587.5	6769.4	2610.5	1456.6	1455.2	2349.3	1168.5	879.3
Var	360.0	671.3	499.3	5887.6	364.9	626.1	537.7	1425.0	375.9	661.3	515.3	880.7
Bias	64.7	120.6	68.5	-2.2	47.1	78.4	45.5	-5.8	32.9	41.1	25.6	0.30

Table 4.8: Simulation results for the indirect effects (simple mediation / multimediate) in a monotonic time-dependent baseline risk scenario

	Correlation = -0.4		
	Med 1	Med 2	Med 3
MSE	650.4 / 641.3	2236.9 / 1109.2	949.6 / 872.9
Var	308.8 / 639.0	613.1 / 1110.8	455.5 / 874.3
Bias	-18.5 / -1.8	-40.3 / 0.56	-22.2 / -0.16
	Correlation = 0		
	Med 1	Med 2	Med 3
MSE	309.2 / 310.3	598.4 / 597.8	491.0 / 492.6
Var	308.9 / 309.7	596.1 / 594.3	491.4 / 492.3
Bias	-0.91 / -1.1	1.8 / 2.1	0.67 / 1.09
	Correlation = 0.4		
	Med 1	Med 2	Med 3
MSE	508.3 / 458.6	2114.0 / 728.9	874.0 / 575.9
Var	337.1 / 445.5	603.5 / 729.5	457.7 / 576.9
Bias	13.1 / -3.7	38.9 / -0.76	20.4 / 0.26

Tables 4.9 and 4.10 show the empirical coverage of 95 % CIs. As for the constant baseline risk scenario, total effects have similar empirical coverage for both simple mediation and the multimediate algorithm. However, the empirical coverage is much better for the multimediate algorithm for both direct and indirect effects. Direct effects have sometimes null empirical coverage in the simple mediation models, and the empirical coverage is also clearly worse in contexts of correlated settings. The multimediate model maintains good and similar empirical coverage for all effects.

Table 4.9: Empirical coverage of the confidence interval with theoretical coverage of 95 % (proportion of simulations including the true value) of simple mediation models in a monotonic time-dependent baseline risk scenario

	Mediator 1		
	Indirect	Direct	Total
Correlation=-0.4	0.81	0.07	0.91
Correlation=0	0.95	0.32	0.92
Correlation=0.4	0.88	0.59	0.89
	Mediator 2		
	Indirect	Direct	Total
Correlation=-0.4	0.60	0.005	0.91
Correlation=0	0.94	0.12	0.91
Correlation=0.4	0.60	0.60	0.90
	Mediator 3		
	Indirect	Direct	Total
Correlation=-0.4	0.81	0.11	0.91
Correlation=0	0.94	0.46	0.92
Correlation=0.4	0.81	0.76	0.91

Table 4.10: Empirical coverage of the confidence interval with theoretical coverage of 95 % (in proportions of simulations) of the multimediate algorithm in a monotonic time-dependent baseline risk scenario

	Indirect M1	Indirect M2	Indirect M3	Direct	Total
Correlation=-0.4	0.96	0.94	0.94	0.95	0.90
Correlation=0	0.95	0.92	0.93	0.94	0.90
Correlation=0.4	0.92	0.94	0.95	0.93	0.89

Non-monotonic baseline hazard

Let us now define the baseline hazard as the following piecewise function:

$$\lambda_0(t) = \begin{cases} 1, & t < 1 \\ 2, & 1 \leq t < 2 \\ 1, & t \geq 2 \end{cases}$$

Then, the cumulative risk would be defined as:

$$\Lambda(t) = \begin{cases} \int_0^t (1+\Psi)du = t + \Psi t, & t < 1 \\ 1 + \Psi + \int_1^t (2+\Psi)du = 2t + \Psi t - 1, & 1 \leq t < 2 \\ 2\Psi + 3 + \int_2^t (1+\Psi)du = t + \Psi t + 1, & t \geq 2 \end{cases}$$

Thus, the survival function would be defined as:

$$S(t) = \begin{cases} exp\{-(t + \Psi t)\}, & t < 1 \\ exp\{-(2t + \Psi t - 1)\}, & 1 \leq t < 2 \\ exp\{-(t + \Psi t + 1)\}, & t \geq 2 \end{cases}$$

and, following simple inequalities calculations, the survival time t would be simulated as:

$$t = \begin{cases} \dfrac{-logU}{1+\Psi}, & U > exp(-1-\Psi) \\ \dfrac{-logU+1}{2+\Psi}, & exp(-1-\Psi) \geq U > exp(-3-2\Psi) \\ \dfrac{-logU-1}{1+\Psi}, & U \leq exp(-3-2\Psi) \end{cases}$$

Tables 4.11, 4.12 and 4.13 show the MSE, variance and bias for the total, direct and indirect effects comparing simple mediation to the multimediate algorithm. Tables 4.14 and 4.15 show the empirical coverage of CIs. The patterns are essentially similar to those observed in the previous two baseline hazard scenarios.

Table 4.11: Simulation results for the total effect in a non-monotonic baseline risk scenario

	Correlation = -0.4				Correlation = 0				Correlation = 0.4			
	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim
MSE	44.8	45.1	44.9	45.1	43.5	43.5	43.4	43.6	40.9	40.7	40.7	40.9
Var	44.9	45.1	44.9	45.0	43.6	43.5	43.4	43.6	40.9	40.8	40.8	40.9
Bias	0.27	0.28	0.28	0.34	0.11	0.11	0.10	0.18	0.06	0.11	0.07	0.17

Table 4.12: Simulation results for the direct effect in a non-monotonic baseline risk scenario

	Correlation = -0.4				Correlation = 0				Correlation = 0.4			
	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim	Med 1	Med 2	Med 3	Multim
MSE	3620.3	1603.6	3066.6	2135.9	2264.4	669.3	2639.5	605.5	1298.5	584.7	2329.7	586.6
Var	122.3	505.2	55.3	2098.3	124.9	498.2	50.1	606.3	126.4	539.0	52.5	586.7
Bias	59.1	33.2	54.9	6.4	46.3	13.1	50.9	-0.43	34.2	-6.8	47.7	-0.94

Table 4.13: Simulation results for the indirect effects (simple mediation / multimediate) in a non-monotonic baseline risk scenario

	Correlation = -0.4		
	Med 1	Med 2	Med 3
MSE	253.8 / 201.6	846.5 / 988.8	21.8 / 20.9
Var	88.3 / 196.9	452.2 / 979.1	8.8 / 20.7
Bias	-12.9 / -2.3	-19.9 / -3.4	-3.6 / -0.46
	Correlation = 0		
	Med 1	Med 2	Med 3
MSE	88.5 / 89.1	445.0 / 450.4	8.5 / 8.5
Var	88.6 / 89.2	445.7 / 450.9	8.4 / 8.4
Bias	-0.15 / -0.23	0.001 / 0.57	0.21 / 0.27
	Correlation = 0.4		
	Med 1	Med 2	Med 3
MSE	224.1 / 105.2	874.6 / 617.9	20.9 / 12.5
Var	84.5 / 105.4	478.0 / 616.6	9.7 / 12.4
Bias	11.8 / -0.11	19.9 / 1.5	3.3 / -0.3

Table 4.14: Empirical coverage of the confidence interval with theoretical coverage of 95 % (proportion of simulations including the true value) of simple mediation models in a non-monotonic baseline risk scenario

	Mediator 1		
	Indirect	Direct	Total
Correlation=-0.4	0.71	0	0.96
Correlation=0	0.94	0.01	0.95
Correlation=0.4	0.73	0.15	0.96
	Mediator 2		
	Indirect	Direct	Total
Correlation=-0.4	0.83	0.66	0.96
Correlation=0	0.96	0.91	0.95
Correlation=0.4	0.83	0.92	0.96
	Mediator 3		
	Indirect	Direct	Total
Correlation=-0.4	0.78	0	0.96
Correlation=0	0.97	0	0.95
Correlation=0.4	0.81	0	0.96

Table 4.15: Empirical coverage of the confidence interval with theoretical coverage of 95 % (in proportions of simulations) of the multimediate algorithm in a non-monotonic baseline risk scenario

	Indirect M1	Indirect M2	Indirect M3	Direct	Total
Correlation=-0.4	0.93	0.94	0.94	0.94	0.95
Correlation=0	0.94	0.93	0.97	0.94	0.94
Correlation=0.4	0.95	0.93	0.94	0.93	0.95

Discussion

In this work, we extended the quasi-bayesian multimediate algorithm to a time-to-event setting using the semiparametric additive hazards model. We theoretically demonstrated that, under certain assumptions, indirect, direct and total effects can be calculated using the counterfactual framework in survival settings. We additionally conducted a simulation study under different baseline risk scenarios and different levels of correlations between mediators to show that the multimediate algorithm has a better performance, in terms of MSEs and CI coverage, than simple mediation analysis, especially in the setting in which mediators are correlated. This work has been added to Github as part of an extension of the original R package *multimediate* developed by Jerolon et al. [9].

Our simulation study shows that, in general, and regardless of the baseline risk definition, the MSEs are smaller for both direct and indirect effects for the multimediate algorithm as compared to those of the simple mediation framework, especially in settings of correlated mediators. Of note, the empirical coverage of the CIs in the multimediate algorithm is far better than that of simple mediation analysis, in which the empirical coverage is worsened for both direct and indirect effects in the context of correlated mediators.

Survival analysis is widely used in mediation analysis applied to medical settings, in which one might be interested in evaluating the potential mediating effect of a biological process on the association between an exposure or treatment and a health outcome. Traditionally, mediation analysis has been conducted using additive hazards models [199], however, to our knowledge, no multi-mediator algorithms for correlated mediators with survival endpoints have been developed to date. Additive hazards models have several advantages as compared to Cox proportional hazards models. Rate differences provide a more straightforward interpretation in attributable cases per person-years and, unlike hazard ratios, are collapsible [142], meaning that the magnitude of the coefficient of the exposure would not change when adjusting the model for a variable that is unrelated to the exposure. In addition, in settings in which the proportional hazards assumption

is not fulfilled [201], the additive models are more appropriate.

However, this model is not without complications. Convergency issues might arise with this survival version of the multimediate algorithm in settings of small sample sizes or very high inverse correlations between mediators, as the Lin-Ying model might present more convergency issues than the Cox model. In our setting, inverse correlations between mediators lower than -0.4 presented convergency issues even for sample sizes greater than $10,000$. On the other hand, given that survival models in general are less informative than linear models due to censoring, larger sample sizes are needed for a survival model than for a linear model to obtain similar results in terms of robustness. This is the reason why we chose larger sample sizes for the simulation study as compared to the simulation study conducted in Jerolon et al. for continuous outcomes [9].

Furthermore, the context of this work requires two important assumptions. First, as stated in Jerolon et al. [9], this work is restricted to the setting in which the correlation between counterfactual mediators is independent of the exposure or treatment. Relevant future work should include the development of methods for addressing the situation in which the correlation between mediators is dependent on the exposure. Second, we assume that the joint distribution of the mediators is a multivariate normal. This is not necessarily true in settings in which mediators are not independent. However, this is not feasible to prove in practice as all linear combinations of the mediators should follow a normal distribution in order to conclude that the joint distribution of the mediators is a multivariate normal. Deviations from multivariate normality should be studied in future work.

Of note, the multimediate algorithm uses the counterfactual framework to identify direct, indirect and total effects. Traditional mediation approaches such as the product of coefficients and the difference of coefficients [197] approaches can lead to biased effect estimates in presence of exposure-mediator interactions. As stated by Richiardi et al. [196], the natural direct effects and natural indirect effects as defined by the counterfactual framework can provide valid estimates even in the case of exposure-mediator interactions. Our extension of the

multimediate algorithm provides direct, indirect and total effect estimates in all strata of the exposure, thus, potential exposure-mediator interactions can be identified.

In conclusion, the multimediate algorithm is able to conduct multiple mediation analysis in presence of correlations between mediators. Unlike multiplicative models, the semiparametric additive risks model provides the effect in a rate difference scale, which is a more interpretable measure in a survival setting and can be highly informative for public health.

4.4.2 Data application 2: contribution of blood DNA methylation to explain the association between smoking and smoking-related cancer

Differential patterns in blood DNA methylation are associated with lung cancer, the main cause of cancer death worldwide [55, 57, 56, 58, 59], suggesting that DNA methylation changes may play a key role in tumorigenesis [202]. However, studies investigating the role of DNA methylation in smoking-related lung cancer are unclear [203]. Hypomethylation of CpGs annotated to smoking-related genes including *AHRR* and *F2RL3* has been associated with lung cancer [204]. An in-vitro study showed that smoking-induced epigenetic changes in the *KRAS* oncogene might lead to sensitization of bronchial epithelial cells for malignant transformation [205]. However, two Mendelian randomization studies have provided little evidence in favor of a causal role of DNA methylation in lung cancer [206, 207]. In most studies of smoking, DNA methylation and cancer are limited by the lack of time to incident (i.e. newly diagnosed) cancer or the lack of formal mediation analysis.

On the other hand, smoking is associated with at least 11 types of cancer beyond lung cancer [208]. Although the evidence is weaker compared to lung cancer, differences in DNA methylation have also been related to other smoking-related cancers such as liver [63, 62] esophageal [209], stomach [68, 69], colorectal [61, 60] pancreatic [67, 66] bladder [210, 211] prostate [212, 213] and kidney [65, 64] cancer. Large prospective studies are needed to evaluate whether effects of smoking in smoking-related cancers beyond lung cancer could be partially mediated by differential DNA methylation.

In this study, we investigated whether the association of current and cumulative smoking with lung cancer and smoking-related cancer risk might be explained by differences in human blood DNA methylation. We used data from the SHS as described in section 1.5 (discovery population), and the FHS (replication population). we used the extended multimediate algorithm described in section 4.3 to jointly assess mediated effects in a way that can account for correlations across

DNA methylation sites, which enabled the evaluation of the most impactful DMPs potentially driving smoking-related cancer risk.

In addition, we explored potential functional implications of the DMPs identified in our study using whole blood gene expression in a subset of FHS participants. Bioinformatic pathway enrichment analysis enabled the exploration of potential biological pathways that might be involved on the association between smoking and cancer through DNA methylation differences.

Outcome assessment in the Strong Heart Study

Cancer incidence was assessed by self-report during interviews, death certificates and/or chart reviews and pathology reports if available. Smoking-related cancers included lung cancer, esophageal-stomach cancer, colorectal cancer, liver cancer, pancreatic cancer, and kidney cancer. We calculated follow-up from the date of baseline examination to the date of cancer diagnosis or 31 December 2017, whichever occurred first.

Replication study population: The Framingham Heart Study

Cancer incidence was assessed by interviews, death certificates, and/or chart reviews that included pathology reports, and crosschecked with official medical records whenever possible. We included lung, bladder and prostate cancers. We calculated the follow-up from the date of baseline examination to the date of cancer diagnosis or December 31, 2016, whichever occurred first.

DNA methylation measurements in the Framingham Heart Study

DNA methylation was measured in 2,648 participants who participated in the 8th visit (2005-2008) and 1,522 Generation III participants who participated in the second visit (2006-2009). Details of microarray DNA methylation measurements have been published [18].

Briefly, DNA methylation was assayed from whole blood using the 450K array, which contains 485,512 CpG sites. There were 4,170 samples passing quality control, 2,648 belonging to the Offspring cohort and 1,522 to the third Generation cohort. Please note that the FHS data used in this project is different from that used in section 3.2.2, as only data from the Offspring cohort was used in the project described in section 3.2.2. Raw methylated and total probe intensities were extracted using the Illumina Genome Studio methylation module. Preprocessing of the methylated signal and unmethylated signal was conducted using the *DASEN* function of the R package *wateRmelon2* [214]. Further details regarding DNA methylation data preprocessing can be found in [18]. The final sample size was N=4170.

Gene expression measurements in the Framingham Heart Study

Gene expression from paired whole blood RNA was sequenced at $> \times$ 30 depth of coverage using RNA-SeQC v1.1.9. according to TOPMed RNA-Seq pipeline v242 [215]. Expression quantitative trait methylation (eQTM) refers to CpGs associated with gene expression levels of some transcript. To explore whether DNA methylation changes in significant CpGs in the mediation analysis influence gene expression, we conducted an eQTM analysis in the FHS. This analysis was not conducted in the SHS due to lack of gene expression data. The RNA for the gene expression came from whole blood. Further details regarding gene expression assessment are presented in [216]. Gene expression data was normalized using the *edgeR* package [217] and log-2 transformed. PC regression [218] was performed to identify technical covariates (i.e. batch effects) in the RNASeq dataset, which included RNA integrity number, batch, RNA concentration, and shipping boxes. Identified potential batch effects were subsequently corrected as needed by obtaining residualized expression from batch-adjusted regression.

Cis-eQTMs were defined as eQTMs in which the CpG site falls within a 1 Mb distance from the gene transcription start site. Trans-

eQTMs were defined as those with target genes on other chromosomes or genes outside the contiguous cis-blocks.

Statistical Methods

Association of smoking with smoking-related cancers. We used Cox proportional hazards models and additive hazard models [147] to estimate relative hazards and hazard differences for cancer in the SHS. Models accounted for potential confounding due to age, sex, BMI and study center (Arizona, Oklahoma, or North Dakota and South Dakota). Former smoking has been associated with cancer mortality in the SHS [219]. Consequently, we kept the regression coefficient for former smoking status in the models (i.e. two indicator variables were simultaneously introduced in the regression models, for mutually exclusive former and current smoking status categories, with the never smoking category being the reference). Former smoking indicator was thus considered an adjustment variable. Cumulative smoking models were additionally adjusted for current smoking status using an indicator variable.

Differential Methylation Analysis by ISIS-enet. We first conducted a screening among the CpG sites that were associated with smoking in previous work in the SHS (303 CpGs in total) [220], by using a Cox ISIS coupled with elastic-net (ISIS-enet, as conducted by the extended SIS R package described in section 3.1.5) to select CpG sites associated with time to lung cancer and time to smoking-related cancers. In differential methylation analysis, we used the same adjustment models as in the association analysis of smoking with smoking-related cancer, but additionally including DNA methylation-related variables such as cell counts (CD8T, CD4T, NK, B cells and monocytes) and five genetic PCs as described in section 1.5.

Mediation analysis based on additive hazards models. We calculated natural direct, indirect and total effects based on the product of coefficients method for survival mediation analysis using additive hazards models as described in section 2.4. Our outcome model was an

additive hazards model with time to incident cancer as outcome, current smoking and cumulative smoking as exposures, and logit-2 transformed DNA methylation proportions (M values) as mediators. Given that DNA methylation changes are reversible upon smoking cessation [206], and that cancer risk decreases over time in former smokers [207], we did not consider former smoking as an exposure of interest in our mediation analysis. Our mediator model was a linear model with the same logit-2 transformed DNA methylation proportions as outcomes, and smoking-related variables as the exposure. Models were adjusted for the same variables used in differential methylation analysis.

First, separate mediation models were run for each of the DMPs selected by the ISIS-enet model for each of the two endpoints and two exposures (current smoking status or cumulative smoking) in the SHS. For statistically significant CpGs identified in the SHS that were included in the 450K array, we subsequently reproduced single mediator models using FHS data.

Mediated effects were reported as differences in cancer cases comparing current to never smokers, or differences in cancer cases per a 10 cigarette pack-years increase, attributable to smoking-related blood DNA methylation differences per 100,000 person-years. The corresponding 95 % CIs were calculated using resampling from the multivariate normal distribution as described in Lange and Hansen [139].

Expression quantitative trait methylation (eQTM) analysis. To quantify the association between DNA methylation and gene expression, we conducted an eQTM analysis. We fitted a linear model for DMPs that were significant in the single-DMP mediation analysis both in the SHS and the FHS. The final regression model included batch effect-corrected expression as the dependent variable, batch effect-corrected DNA methylation as an independent variable, and adjustment for sex, age, predicted blood cell fraction to account for signal heterogeneity from multiple sample types [221], five expression PCs and 10 DNA methylation PCs.

Multimediator model. In presence of correlated mediators, traditional mediation analysis methods might lead to individual relative mediated effects that add up to more than 100 %, which suggests that some pathways are overlapping and the joint and individual effects remain unidentifiable. To address this limitation, we extended the multimediate algorithm to the survival data setting using additive hazards models as described in section 4.3. Our novel multimediate algorithm is able to identify individual mediated effects of several mediators simultaneously while taking into account correlated mediators. In this setting, relative mediated effects could never add up to more than 100 %. The multimediator model was only evaluated for current versus never smoking, as it has not yet been extended to continuous exposures or treatments. Mediated effects with p-values lower than 0.05 were considered statistically significant.

Individual indirect effects cannot be correctly identified in presence of correlated mediators using the traditional "difference of coefficients" method [197]. However, the joint mediated effect for a given set of correlated mediators as calculated by the "difference of coefficients" method and the joint indirect effect as calculated by the multimediate algorithm should yield similar results. We thus ran post-hoc sensitivity analyses using the traditional "difference of coefficients method" to provide additional support to our newly developed multi-mediator model.

Enrichment analysis. We conducted a KEGG enrichment analysis out of the genes annotated to cis- and trans- eQTMs to explore possible biological implications of our findings. We considered a given KEGG pathway as significantly enriched if the enrichment p-value was ≤ 0.01 based on a two-sided hypergeometric test and at least 10 eQTM-related genes were contributing to that pathway. The Kappa statistic, which is used to define KEGG terms interrelations (edges) and functional groups based on shared genes between terms, was set to 0.4. The enrichment analysis was performed using Cytoscape (version.3.8.2) [158] with the ClueGO (version 2.5.8) and CluePedia (version 1.5.8) plugins [159].

Sensitivity analysis. Oncogenic transformations can happen several years before cancer diagnosis. Thus, as an attempt to discard cases where DNA methylation may have been measured after oncogenic transformations started, we repeated the mediation analysis excluding individuals with cancer that was diagnosed in the first 5 follow-up years (10 lung cancer and 27 smoking-related cancer cases excluded). Given the non-statistically significant inverse association between smoking and liver cancer in the SHS, we conducted an additional sensitivity analysis excluding the liver cancer cases from the smoking-related cancer endpoint in the mediation analysis.

Results

Association of smoking and smoking-related cancers. Participants with lung cancer and smoking-related cancers were older, had higher cumulative smoking and were mostly current smokers, especially for lung cancer (Table 4.16). Adjusted hazard ratios (HR) (95 % CIs) in the SHS for current versus never smoking and cumulative smoking for different cancers can be found in Table 4.17.

Table 4.16: Participant characteristics for the Strong Heart Study and the Framingham Heart Study by cancer status.

	Strong Heart Study			Framingham Heart Study		
	Smoking-related cancer (N=222)	Lung cancer (N=97)	Non-cases (N=2013)	Smoking-related cancer (N=251)	Lung cancer (N=56)	Non-cases (N=3919)
Age (years), median (IQR)	57 (51.2, 64.6)	57.6 (52.8, 64.7)	54.7 (49.0, 61.6)	69 (62, 75)	68 (61, 74.3)	59 (48, 68)
Sex, % Male	47.3	53.6	40.5	80.5	44.6	44.2
Smoking status						
Former, %	26.6	15.5	30.2	67.3	71.4	45.2
Current, %	54.5	75.3	37.7	11.2	23.2	10.5
Pack-years, median (IQR)	12.5 (1, 34)	26 (9, 44)	3 (0, 17)	0.63 (0, 16.6)	18.1 (1.8, 37.5)	0.25 (0, 10.5)
BMI, median (IQR)	28.7 (25.3, 32.7)	27.5 (24.5, 30.8)	29.7 (26.3, 33.7)	27.8 (25.6, 30.5)	27.4 (23.4, 30.6)	27.3 (24.2, 31.0)

Table 4.17: Hazard ratios and rate differences (cases/100,000 person-years) (95 % CI) of smoking-related cancer by current and cumulative smoking in the Strong Heart Study (N=2235).

		Smoking status (current versus never)		Cumulative smoking[a]	
	N cases/non-cases	HR (95 % CI)	RD (95 %CI), cases/ 100000 person-year	HR (95 % CI)	RD (95 %CI), cases/ 100000 person-year
Smoking-related cancers	222 / 2013	2.5 (1.7, 3.6)	440.2 (280.1, 600.3)	1.2 (1.1, 1.3)	152.0 (76.1, 228.0)
Non-lung smoking-related cancers	125 / 2107	1.4 (0.9, 2.2)	101.3 (-18, 220.6)	1.1 (1.0, 1.2)	17.9 (-25.4, 61.2)
Lung	97 / 2138	5.7 (2.8, 11.7)	334.4 (227.5, 441.3)	1.3 (1.2, 1.4)	132.4 (68.1, 196.8)
Colorectal	46 / 2189	1.7 (0.8, 3.6)	61.4 (-13.8, 136.5)	1.1 (1, 1.3)	6.5 (-17.2, 30.2)
Kidney	24 / 2211	1.4 (0.4, 4.3)	17.2 (-30.3, 64.7)	1.1 (0.9, 1.3)	4.6 (-15.9, 25.1)
Pancreatic	23 / 2212	1.7 (0.5, 5.5)	15.5 (-24.6, 55.6)	1.1 (0.9, 1.4)	4.0 (-16.2, 24.1)
Esophageal-stomach	23 / 2212	1.5 (0.5, 4.5)	18.8 (-31.7, 69.2)	1.2 (1.0, 1.4)	19.5 (-6.9, 45.8)
Liver	19 / 2216	0.6 (0.2, 1.9)	-15.1 (-69.8, 39.5)	0.7 (0.4, 1.2)	-6.7 (-14.2, 0.9)

Abbreviations: HR, Hazard ratios from Cox proportional hazards models; RD, rate differences from additive hazards models.

Models were adjusted for age, sex, BMI and center.

[a] Cumulative smoking models per 10 cigarette pack-years increase were additionally adjusted for current smoking status (yes/no).

Mediation Analysis. The Cox ISIS model selected 62 and 69 DMPs associated with lung cancer (Appendix C, Table C1) and smoking-related cancers (Appendix C, Table C2), respectively. In lung cancer models, 29 (out of 62) CpGs had statistically significant indirect effects in the SHS for current versus never smoking. Among those, 20 were also measured in the FHS, of which 14 were replicated in the FHS (Table 4.18). For cumulative smoking, 20 (out of 62) CpGs had statistically significant indirect effects in the SHS. Among those, 14 were also measured in the FHS, of which four were replicated in the FHS (Appendix C, Table C3).

In smoking-related cancer models, for current versus never smoking, 37 (out of 69) CpGs had statistically significant indirect effects in the SHS. Among those, 17 CpGs were measured in the FHS, of which five were replicated in the FHS (Appendix C, Table C4). For cumula-

tive smoking, 20 CpGs (out of 69) had statistically significant indirect effects in the SHS. Among those, 11 were measured in the FHS, of which six were replicated in the FHS (Appendix C, Table C5).

Table 4.18: Differences in lung cancer cases per 100,000 person-years comparing current to never smokers attributable to differences in DNA methylation for each CpG ('mediated effects') in the Strong Heart Study and replication in the Framingham Heart Study.

		Strong Heart Study			Framingham Heart Study		
		Mediated (i.e., indirect) effect of current vs never smoking through DNAm[b]		Direct effect of current vs never smoking[a]	Mediated (i.e., indirect) effect of current vs never smoking through DNAm[b]		Direct effect of current vs never smoking[a]
CpG	Gene	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases comparing current vs never smokers (95 % CI) per 100,000 person-years	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases comparing current vs never smokers (95 % CI) per 100,000 person-years
cg05575921	*AHRR*	253.9 (167.3, 342.3)	76.5 (50.9, 113.6)	78.0 (-34.5, 190.5)	207.7 (65.7, 350.1)	68.8 (23.2, 172.8)	94.1 (-107.9, 295.5)
cg21566642	*ALPG*	152.9 (85.9, 221.3)	45.5 (25.3, 73.7)	183.6 (66.5, 300.3)	172.2 (71.3, 273.7)	57.3 (24.2, 142.2)	128.5 (-58.1, 315.1)
cg14391737*	*PRSS23*	149.9 (92.2, 210.0)	42.7 (26.8, 64.4)	201.1 (94.4, 307.4)	-	-	-
cg03636183	*F2RL3*	136.9 (79.9, 195.5)	41.0 (24.3, 63.9)	196.8 (90.1, 303.2)	191.4 (57.5, 326.1)	63.9 (20.5, 162.1)	108.3 (-90.2, 306.7)
cg01940273	*ALPG*	107.0 (47.9, 167.2)	31.9 (14.0, 56.2)	228.9 (109.8, 347.8)	93.4 (12.5, 174.5)	31.7 (4.4, 89.4)	201.4 (12.3, 390.5)
cg24859433	*IER3*	91.4 (49.2, 136.4)	26.7 (14.7, 42.3)	251.1 (146.6, 355.5)	95.3 (21.8, 169.6)	32.1 (7.9, 84.7)	201.5 (18.3, 384.7)
cg03329539	*ALPG*	72.7 (32.5, 115.1)	21.6 (9.5, 38.1)	263.9 (152.6, 374.8)	76.0 (30.2, 122.6)	25.7 (10.9, 62.5)	219.9 (44.1, 395.7)
cg17739917*	*RARA*	69.9 (26.9, 114.1)	20.6 (8.1, 35.8)	270.4 (163.1, 377.4)	-	-	-
cg09842685*	*FGF23*	64.1 (36.2, 94.1)	18.8 (10.8, 29.6)	276.5 (173.9, 378.9)	-	-	-
cg01899089	*AHRR*	51.3 (24.9, 80.2)	15.1 (7.6, 24.9)	288.5 (184.8, 392.1)	55.2 (16.2, 95.5)	18.6 (6, 45.8)	242.1 (64.7, 419.4)
cg04885881	*SRM*	50.8 (15.4, 87.7)	14.8 (4.7, 27.1)	291.6 (185.4, 397.6)	89.8 (40.1, 141.2)	30.4 (13.8, 75.3)	205.4 (28.2, 382.5)
cg03707168	*PPP1R15A*	48.4 (16.9, 82.7)	14.2 (5.1, 25.6)	292.4 (186.4, 398.2)	61.2 (17.2, 106.1)	20.6 (6.5, 50.5)	235.5 (59.3, 411.6)
cg11902777	*AHRR*	42.6 (25.3, 62.2)	12.4 (7.4, 19.5)	301.2 (196.0, 406.3)	43.7 (11.2, 77.2)	15.0 (4.1, 39.6)	248.5 (68.6, 428.2)
cg14580211	*SMIM3*	39.3 (10.2, 69.8)	11.5 (3.1, 21.9)	301.6 (194.9, 408.0)	42.1 (-13.9, 98.9)	14.4 (-6.2, 43.9)	250.1 (70.4, 429.9)
cg14624207	*LRP5*	38.3 (16.6, 62.7)	11.4 (4.9, 20.1)	298.7 (193.1, 404.1)	40.0 (-6.1, 86.7)	13.6 (-2.5, 41.9)	254.2 (70.3, 437.9)
cg27241845	*ECEL1P2*	36.2 (11.3, 63.6)	10.8 (3.4, 20.5)	299.4 (192.3, 406.1)	45.3 (6.5, 84.8)	15.5 (2.6, 39.3)	246.8 (70.9, 422.6)
cg01513913	*FAM30A*	35.1 (12.0, 60.5)	10.4 (3.5, 19.7)	301.7 (193.9, 409.3)	33.1 (-7.6, 74.9)	11.3 (-2.7, 41.2)	259.6 (68.6, 450.4)
cg16207944*	*FAM30A*	33.9 (12.5, 57.4)	10.1 (3.7, 18.6)	302.9 (196.3, 409.2)	-	-	-
cg23916896	*AHRR*	33.9 (12.5, 57.5)	10.0 (3.8, 17.9)	305.3 (200.6, 409.8)	95.8 (47.0, 145.7)	32.3 (15.1, 82.1)	201.1 (20.2, 382.0)
cg07251887	*RECQL5*	29.3 (10.5, 50.8)	8.7 (3.2, 16.1)	307.8 (201.7, 413.6)	59.8 (21.2, 99.6)	20.4 (6.8, 57.8)	233.5 (47.8, 419.3)
cg02738868*	*ELMSAN1*	28.7 (5.3, 53.8)	8.5 (1.6, 17.2)	310.6 (202.7, 418.2)	-	-	-
cg06521527*	*NEDD9*	27.4 (8.9, 48.1)	8.0 (2.7, 14.6)	314.5 (209.8, 419.1)	-	-	-
cg24947681*	*THBS1*	26.6 (7.9, 47.4)	7.9 (2.4, 15.3)	310.5 (203.5, 417.3)	-	-	-
cg16201146	*SLC24A3*	26.0 (9.1, 45.5)	7.6 (2.8, 13.8)	315.7 (210.9, 420.3)	17.9 (-6.6, 43.5)	6.2 (-2.9, 19.1)	274.2 (93.4, 455.1)
cg18158149*	*NOS1AP*	25.9 (6.8, 47.3)	7.6 (2.1, 14.1)	317.7 (213.8, 421.4)	-	-	-
cg23025288*	*HS6ST1*	23.6 (5.8, 43.7)	7.0 (1.8, 13.6)	312.7 (207.6, 417.6)	-	-	-
cg23771366	*PRSS23*	23.7 (3.7, 45.7)	7.0 (1.1, 14.4)	316.1 (209.1, 422.7)	80.1 (22.8, 138.6)	27.2 (9.1, 64.4)	214.4 (42.8, 385.9)
cg24556382	*GALNT7*	19.5 (5.3, 36.0)	5.7 (1.6, 10.7)	321.9 (217.5, 425.9)	32.8 (-2.9, 69.4)	11.2 (-1.2, 35.3)	260.2 (75.6, 444.8)
cg25799109	*ARHGEF3*	19.4 (3.8, 37.2)	5.7 (1.1, 11.6)	319.1 (213.3, 424.6)	3.9 (-14.2, 22.5)	1.4 (-6.8, 9.4)	286.0 (103.1, 468.9)

Abbreviations: CI, confidence interval; DNAm, DNA methylation.

* CpGs not present in the 450K array, therefore not evaluated in the Framingham Heart Study.

Models were adjusted for age, sex, former smoking, BMI and cell counts (CD8T, CD4T, NK, B cells and monocytes). Additionally adjusted for study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs in the Strong Heart Study.

[a] Absolute changes in cancer incidence (per 100,000 person-years) for current versus never smokers were obtained from additive hazards models.

[b] Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for the current versus never smoking comparison from the mediator model by the absolute change in cancer incidence cases for the current versus never smoking comparison (difference in change reflecting the number of attributable cancer cases per 100,000 person-years), and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % confidence intervals in the table were derived by simulation from the estimated model coefficients and covariance matrices.

A descriptive table comparing blood DNA methylation proportions in the SHS and the FHS for the CpGs that were statistically significant in the mediation analysis in both the SHS and the FHS is shown in Appendix C, Table C6. DNA methylation proportions at the specific CpGs were highly consistent in the SHS and the FHS. DNA methylation proportions were generally lower in individuals that developed cancer as compared to those that did not.

Expression quantitative trait methylation (eQTM) and biological pathway enrichment. At a statistical significance p-value $< 10^{-4}$, 17 mediating DMPs of lung cancer in common for the SHS and FHS were associated with 12 cis-eQTMs and 2415 trans-eQTMs. The large majority of the eQTM-associated transcripts (75.7 % of transcripts in trans and 83.3 % of transcripts in cis) showed, overall, gene expression downregulation. The number of cis-eQTMs and tras-eQTMs, as well as the direction of association and the CpG location, are shown for the DMPs that were significant in the mediation analysis for both the SHS and the FHS or in the multimediator model in Appendix C, Table C7. Biological pathway enrichment analysis of target genes

annotated to eQTM-associated transcripts showed 54 enriched biological pathways (Figures 4.1 and 4.2). Figure 4.1 displays overlapping DMPs, eQTMs and KEGG biological pathways by the evaluated exposures and endpoints. The enriched pathways were largely related to cancer (Figure 4.2).

Figure 4.1: Summary of identified differentially methylated positions, expression quantitative trait methylation genes and enriched biological pathways by endpoint and smoking-related variables. A) Venn diagram of differentially methylated positions with significant mediated effects both in the SHS and FHS by combinations of evaluated endpoints and smoking variables. B) Venn diagram of genes annotated to the differentially expressed transcripts in trans in the Framigham Heart Study by combinations of evaluated endpoints and smoking variables. C) Upset plot of the overlapping enriched KEGG pathways.

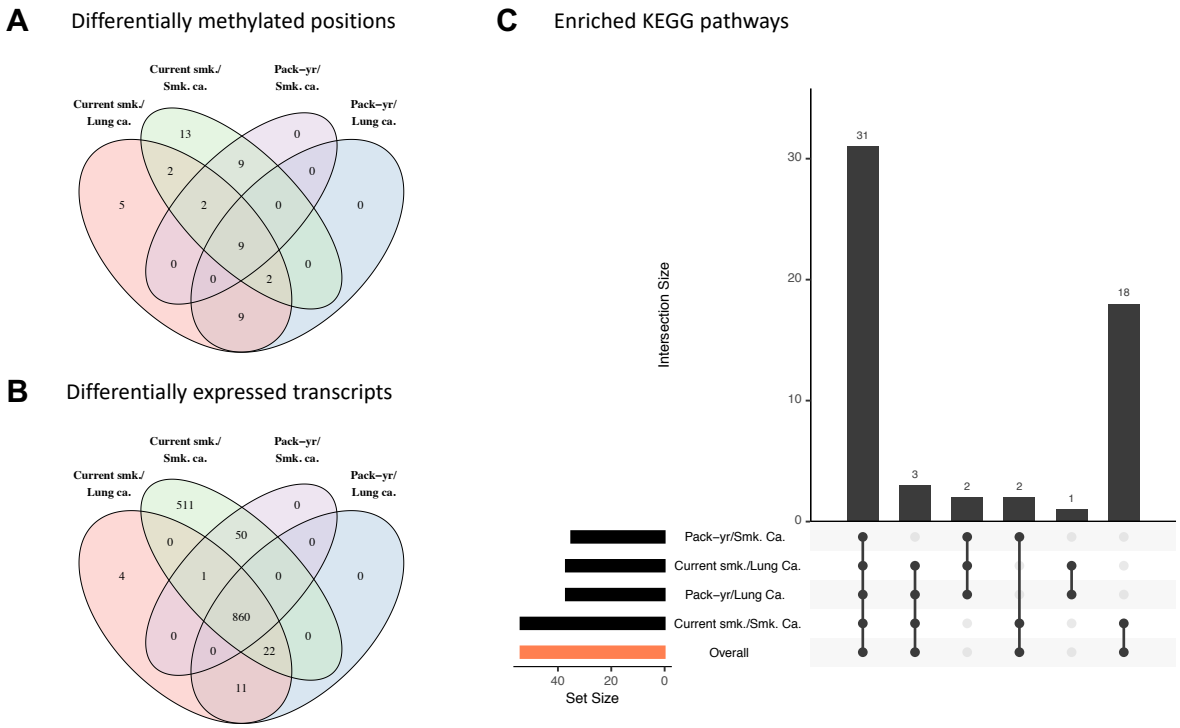
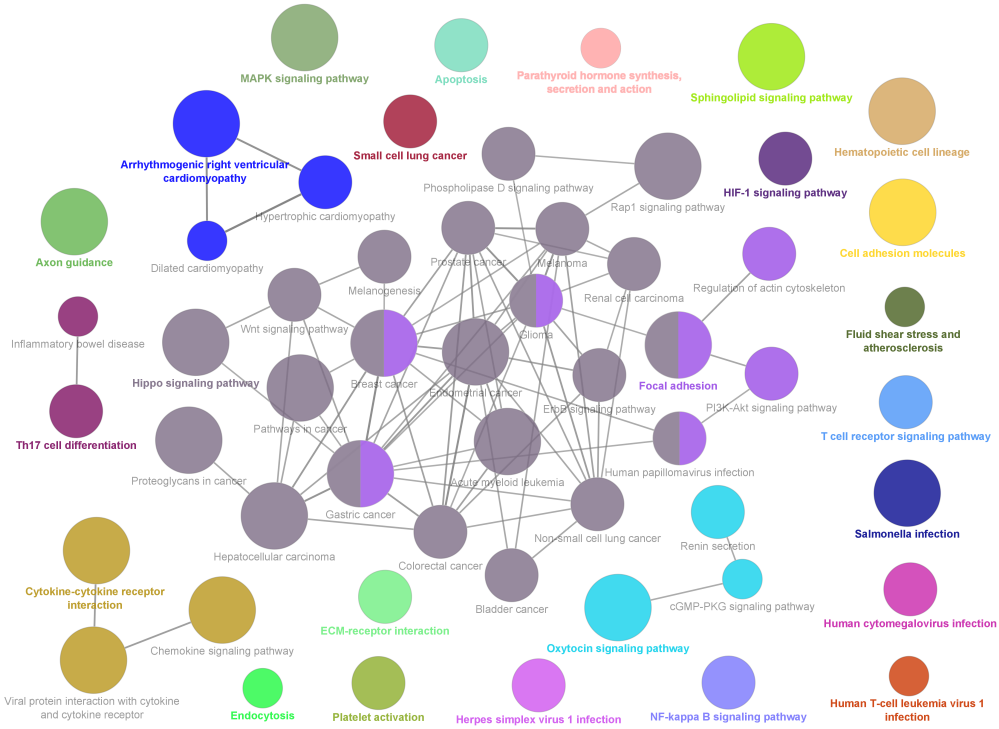


Figure 4.2: Network of significantly enriched pathways for annotated trans expression quantitative trait methylation genes from CpGs with significant mediated effects in the Strong Heart Study and the Framingham Heart Study.



KEGG pathways are represented as nodes and the node size represents the term enrichment significance (increasing size of nodes reflect smaller p-values). Nodes with the same colors reflect they belong to the same cluster based on a Kappa clustering statistic cut-off of 0.4. The nodes with colored letters represent the most significant pathway within a clustering group.

Multimediator analysis. In multi-mediator models, in absolute terms, of the 385.7 (95 % CI 265.9, 509.8) incident lung cancer cases per 100,000 person-years attributable to current smoking, 223.6 (95 % 126.1, 324.5), 62.6 (95 % CI 16.8, 110.2) and 28.3 (95 % CI 11.5, 46.5) lung cancer cases were attributable to differences in DNA methylation in cg05575921 (*AHRR*), cg24859433 (*IER3*) and cg11902777 (*AHRR*), respectively (Table 4.19). For incident smoking-related cancer, in ab-

solute terms, of the 506.7 (95 % CI 315.1, 698.6) smoking-related cancer cases per 100,000 person-years attributable to current smoking, 148.5 (95 % CI 59.7, 240.5), 90.9 (95 % CI 47.9, 137.9), 59.6 (95 % CI 26.2, 98.6) and 28.1 (95 % CI 9.5, 52.6) cases were attributable to DNA methylation differences in cg19859270 (*GPR15*), cg01513913 (*FAM30A*), cg16201146 (*SLC24A3*) and cg01002722 (*FSCN1*), respectively (Table 4.19). The joint mediated effects estimated using the "difference of coefficients method" were similar to the sum of individual mediated effects calculated using the multimediator model (Appendix C, Tables C8 and C9).

Table 4.19: Differences in cancer cases per 100,000 person-years comparing current to never smokers attributable to differences in DNA methylation for each CpG ('mediated effects') from a multimediator model in the Strong Heart Study.

CpG	Gene	Mediated (i.e. indirect) effect of current vs never smoking through DNAm (95 % CI)[a]	Percentage of difference in cancer cases attributable to DNAm (95 % CI)[b]
Lung cancer			
cg05575921	*AHRR*	223.6 (126.1, 324.5)	58.1 (30.8, 98.4)
cg24859433	*IER3*	62.6 (16.8, 110.2)	16.2 (4.2, 32.1)
cg11902777	*AHRR*	28.3 (11.5, 46.5)	7.3 (2.9, 13.8)
cg05575921 + cg24859433 + cg11902777	Joint effect	314.6 (210.4, 419.5)	81.3 (55.4, 120.4)
Smoking-related cancer			
cg19859270	*GPR15*	148.5 (59.7, 240.5)	29.2 (11.3, 57.3)
cg01513913	*FAM30A*	90.9 (47.9, 137.9)	17.9 (8.9, 33.0)
cg16201146	*SLC24A3*	59.6 (26.2, 98.6)	11.7 (4.9, 23.3)
cg01002722	*FSCN1*	28.1 (9.5, 52.6)	5.3 (1.8, 11.9)
cg19859270 + cg01513913 + cg16201146 + cg01002722	Joint effect	327.2 (211.8, 446.2)	64.4 (40.7, 103.7)

Abbreviations: DNAm, DNA methylation; CI; confidence interval.

Direct effect of smoking in lung cancer: 71.1 (-60.7, 200.0), total effect: 385.7 (265.9, 509.8).

Direct effect of smoking in smoking-related cancer: 179.6 (-13.1, 315.1), total effect: 506.7 (315.1, 698.6).

[a] Mediated effects are calculated based on the counterfactual framework, i.e. leaving the exposure constant and substracting the number of cancer cases per 100,000 person-years with DNA methylation fixed to the value it would take in presence of current smoking to the number of cancer cases per 100,000 person-years with DNA methylation fixed to the value it would take in absence of smoking: $Y(E, M(1))$–$Y(E, M(0))$, being 1 current smoking and 0 never smoking. For individual mediated effects, DNA methylation levels of all CpGs except the CpG of interest are fixed to the value of the exposure (i.e., only the CpG of interest is variable). For the joint mediated effects, all CpGs are variable.

[b] Mediated percentages are calculated dividing the mediated effect by the total effect. The total effect is calculated based on the counterfactual framework, i.e. substracting the number of cancer cases per 100,000 person-years with the exposure fixed to current smoking and DNA methylation fixed to the value it would take in presence of current smoking to the number of cancer cases per 100,000 person-years with the exposure fixed to never smoking and DNA methylation fixed to the value it would take in absence of smoking $Y(1, M(1))$–$Y(0, M(0))$, being 1 current smoking and 0 never smoking. Model adapted from [9].

Models were adjusted for age, sex, former smoking, BMI, study center (Arizona, Oklahoma or North and South Dakota) cell counts (CD8T, CD4T, NK, B cells and monocytes) and five genetic PCs.

Sensitivity analysis. The mediation models excluding cancer cases diagnosed during the first 5 follow-up years yielded similar results as compared to the main analyses (Appendix, Tables C10 to C13). The results of the mediation analysis excluding liver cancer from the smoking-related cancer endpoint yielded highly similar results (Appendix, Table C14).

Discussion

In our study, we conducted a formal mediation analysis (including multiple mediators evaluated simultaneously) using time-to-newly diagnosed cancer data, and found that a substantial extent of the prospective association of smoking with lung and smoking-related cancers was explained by differences in blood DNA methylation. Results were largely consistent in the FHS, including additional validation of findings with gene expression data, which mostly showed methylation-related downregulation of distant genes that have a plausible role on

cancer biological pathways. In the multimediator model, a joint mediated effect of 81.3 % was driven by three DMPs (annotated to *AHRR* and *IER3*) for lung cancer, and a joint mediated effect of 64.4 % was driven by four DMPs (annotated to *GPR15*, *FAM30A*, *SLC24A3* and *FSCN1*) for smoking-related cancers.

Of note, our novel multimediate algorithm enabled us to explore the joint mediated effects of DMPs. Although many DMPs showed individual mediated effects in the single mediation analysis, the multimediate algorithm identified that the mediated effect was only driven by three and four DMPs for lung and smoking-related cancers, respectively. This means that many DMPs were identified as mediators by the single mediation analysis just because of having high correlations with actual mediators, but when considering them jointly in the same model, their contribution to the mediated effect was not significant. This fact highlights the importance of considering a multiple mediation approach as opposed to a simple mediation one.

The fact that *AHRR* and *F2RL3* genes showed significant mediated effects in our single mediator analysis for both endpoints is widely consistent with findings from numerous study populations [18]. However, previous studies lack formal mediation analysis, except for a case-control study which was part of a Norwegian cohort [204]. This study reported that *AHRR* and *F2RL3* genes explained ~ 37 % of the total effect of smoking in lung cancer. Nevertheless, only single mediation analysis was conducted, and the study lacked follow-up. Also, a study used data from The Cancer Genome Atlas to assess mediation of the association between smoking and lung cancer mortality by blood DNA methylation [222] with inconsistent findings compared to our study. However, this study had a smaller sample size (N=907) and used Cox proportional hazards models in mediation analysis, which is not advisable due to the non-collapsibility of the hazard ratios, as explained in the introduction of this section (section 4). A recent study conducted a Mendelian randomization analysis to assess the potential causal association of DNA methylation in several smoking-related genes including *AHRR* and *F2RL3* and lung cancer with conflicting results [206], possibly given some of the limitations reported by the au-

thors. Additional Mendelian Randomization studies with sufficiently valid genetic instruments and methods to accomodate the multiple correlated DNA methylation mediators are needed.

Interestingly, we mostly found inverse associations between blood DNA methylation at sites identified in the mediation analysis and gene expression. Of especial interest is *GPR15*, as it was identified both as a closest annotated gene to a relevant DMP from the multimediator analysis, and as a trans target gene of other DMPs in the eQTM analysis. DNA methylation in this gene was identified as a potential mediator on the association between smoking and lung cancer in a previous study [207]. Upregulation of *GPR15* was proposed as a biological mechanism involved in smoking-related chronic inflammatory diseases [223]. Subsequent biological pathway enrichment analysis among target genes annotated to eQMTs pointed to relevant pathways in cancer [224, 225, 226]. The association of DNA methylation with gene expression in our cross-sectional analysis, however, is not definitive proof that changes on DNA methylation result in changes on gene expression. Research is needed to confirm the influence of smoking-related DNA methylation on gene expression.

This study has several limitations. First, although the replication in the FHS was high for lung cancer in the current versus never smoking model, it was smaller for lung cancer in the cumulative smoking model and for smoking-related cancers. Differences in smoking intensity and cessation across the SHS and FHS could explain some of the non-replicated DMPs. The somewhat lower replication for the combined smoking-related cancer compared to lung cancer may be due to the fact that the smoking-related cancer endpoint could not be defined homogeneously in the SHS and the FHS, as the FHS lacks data on esophagus-stomach, colorectal, kidney, pancreatic and liver cancer, and the SHS lacks data on bladder and prostate cancer. Also, non-fatal cancer data might be incomplete in the SHS as non-fatal cancers were not confirmed with chart review and no linkage with the cancer registry is available. Despite these limitations, however, we still found substantial replication of findings between the SHS and the FHS for smoking-related cancers.

Second, mediation analysis provides valid estimates only if the mediation assumptions such as absence of unmeasured confounding, which cannot be fully verified in practice, hold [227]. In addition, the multimediate algorithm is only valid in settings of non-causal correlations [9]. Our results need to be interpreted with caution, especially for probes that could not be replicated because were not available by design in the replication microarray. Experimental studies are needed to confirm the role of the identified blood DNA methylation signature of smoking in the association between smoking and smoking-related cancers.

Strengths of our study include replication in an independent cohort, the large sample size with methylation data from one of the largest microarrays nowadays available, the availability of information to account for numerous potential confounders and the additional validation of the results using gene expression data. In addition, we used state-of-the-art statistical methods including the multimediate algorithm for time-to-event data, which enabled the evaluation of correlated methylation sites jointly.

In conclusion, the prospective association of smoking with lung cancer in this study was largely explained by differences in few specific blood DNA methylation sites. These findings contribute to the identification of potentially novel mechanisms of lung cancer, and provide evidence in favor of DNA methylation as a potential biological intermediary in the association between smoking and smoking-related cancers. Additional experimental and translational research targeting the identified methylation sites is needed to assess the relevance of these epigenetic signatures for the prevention and control of smoking-related cancer and lung cancer.

CHAPTER 5

Prospects for future research: transcriptomics from single cell RNA sequencing

As explained in section 1.3, scRNAseq is able to identify cellular heterogeneity in a more precise way as compared to bulk RNAseq. Many bioinformatic tools have been developed in the last years for the assessment of transcriptional differences across genes using scRNAseq. However, most of those tools focus on differences in mean, and do not explore differences in variability. In fact, to our knowledge, no specific method for differential variability testing has been developed for scRNAseq data. Increased cell-to-cell transcriptional and epigenetic variability has been proposed to be a major biomarker of ageing [228]. In addition, transcriptional variability has been proposed to contribute to early cancer evolution [229]. Evaluating differential transcriptional variability at a single cell level is relevant to identify biological features that might not be captured by bulk RNAseq.

However, scRNAseq data pose statistical challenges beyond those present in regular bulk RNAseq data given the very high amount (over 90 % for certain genes) of zeros present in the data [230]. Often, zeros in the gene expression matrix correspond to genes actually expressed

in a given cell, but incorrectly measured as unexpressed. To overcome this limitation, several solutions have been proposed including imputation [231], aggregation of cells within biological replicates (pseudo-bulk scRNAseq) [232], or aggregation of transcriptionally similar cells (SuperCell R package) [233].

The aim of this work was to explore the performance of statistical tools for differential variability developed for other omics data types, in scRNAseq data. Our main focus was the diffVar algorithm [234], which is a statistical tool for identification of differential variability originally developed for DNA methylation data, and later adapted to bulk RNAseq data. We also consider two additional methods beyond diffVar. Distinct [235], which captures differences in distribution (including, but not limited to differences in variability) and scDD [236], which, in addition to differences in mean, captures differences in modalities and proportions. We conducted a simulation study to evaluate the performance of those methods at the single cell level in presence of different proportions of zeros, as well as using imputations, bulk scRNAseq and SuperCell.

Methods

diffVar. diffVar is a statistical tool implemented in the R package *missMethyl* [234]. It aims to test for differential variability using the Levene's z test, which can be thought of as the distance of each point within a group from the group mean. In addition, it applies an empirical Bayes framework to stabilize the t-statistics and avoid high rates of false positives [104]. It was first developed for DNA methylation data, and was later extended to bulk RNAseq data under the limma framework. Good control of the FDR has been documented for this tool using DNA methylation data [234]. However, to date, its performance has not been tested in scRNAseq data.

SuperCell. The SuperCell tool implements the walktrap algorithm, a network-based coarse-graining framework, to merge transcriptionally similar cells into a single feature, called supercell [233]. The graining level (gamma parameter) represents the number of cells that are

encompassed into each supercell. The number of k nearest neighbors (kNN) for the walktrap algorithm is also user-specified. Rather than identifying populations of cells that can be mapped to biological cell types (which is the goal of standard clustering), the goal of SuperCell is to put together cells with similar transcriptomic information, in order to synthesize the information they provide. The SuperCell framework has shown to efficiently preserve the structure of scRNA-seq data while reducing the dimensionality of the matrix to simplify and accelerate the process, reduce the noise and enable efficient downstream analysis.

scRNAseq imputation: SAVER (Single-cell Analyses Via Expression Recovery). The SAVER tool borrows information across genes and cells to recover real expression levels for the zeros present in the gene expression matrix in scRNAseq data [237]. SAVER assumes that the gene expression level of each gene in each cell follows a negative binomial distribution. The prior parameters are estimated using an empirical Bayes approach with a Poisson Lasso regression, using the expression of other genes as predictors. The posterior mean of the distribution is used as the imputed expression value.

Distinct. The distinct tool, implemented in the *distinct* Bioconductor package, aims to test for differences in full distribution, including, but not limited to differential variability [235]. Differences in distribution are quantified using hierarchical non-parametric permutation tests on the cumulative distribution functions (CDFs) of each sample. P-values are then adjusted for multiple comparisons using the Benjamini and Hochberg approach [102]. Distinct has the advantage that it does not rely on asymptotic theory, and avoids parametric assumptions. This method showed good control of FDR and was able to detect more differential patterns as compared to other methods such as limma-voom or edgeR [235].

ScDD. ScDD uses flexible Dirichlet Process Bayesian mixture models to explicitly handle heterogeneity within cell populations in scRNAseq data [236]. It tests for differences in mean, differences in modality,

differences in proportions in multimodal genes, and differences in proportions of zeros. The log-transformed expression values are assumed to follow a Dirichlet Process Mixture of normal distribution, which characterizes expression distribution in terms of number of modes. A Bayes factor score compares the conditional likelihood under the equivalent distributions hypothesis (both conditions or groups are generated from the same clustering process), with the differential distributions hypothesis (each condition is generated from its own clustering process). P-values of statistical significance are obtained empirically via permutations.

Pseudo-bulk RNAseq. Pseudo-bulk analysis consists in summing scRNAseq counts accross cells to get grouped expression levels for each sample, similar to bulk RNAseq data, which does not provide data at the cellular level. This approach helps to avoid zero counts, but at the same time, the precision of the single cell level is lost.

Statistical methods

We simulated scRNAseq data using the *muscat* R package [238], which conducts simulations using a provided dataset as reference. This method assumes that gene expresion data follow a non-zero-inflated negative binomial distribution. We used data from a post-menopausal breast sample as described in Pal et al. [239] for reference for the simulations. We focused on only one sample and one cluster for the simulations for simplicity. From 14,370 genes available, we filtered out genes that had more than 5 % zero counts, as well as mitochondrial and ribosomal genes and genes with missing Entrez Gene IDs [240]. After filtering, we had 7241 genes for the simulation. We used the biological coefficient of variation (BCV) to simulate differences in variability between groups. We simulated two groups of observations in two different scenarios: one with non-differential variability between the two groups (i.e.: BCV=0.00001 in both groups), and the second one with 5 % of genes five times more variable in one of the groups (i.e. BCV=0.5).

The muscat workflow does not simulate zero counts in the count matrix. As our aim was to test whether differential variability methods work in presence of sparsity, we artificially introduced zero counts in the database so that 20% of genes had no zero counts, 20 % had 25 % zero counts, 20 % had 50 % zero counts, 20 % had 75 % zero counts and 20 % had 90 % zero counts. Having genes with > 90 % zero counts is common in scRNAseq data [230]. Zero counts were introduced by randomly selecting the genes and replacing the counts for that gene by zero in order, starting from the lowest count, until reaching the pre-specified zero counts percentage for that concrete gene.

Performance of the different methods was evaluated with and without the introduction of zero counts. FDRs were calculated as the number of false positives divided by the sum of false positives and true negatives, and true positive rates (TPRs) were calculated as the number of true positives divided by the sum of true positives and false negatives. Simulations were repeated 100 times, and descriptives of the FDR and TPR were calculated.

diffVar was applied with the default settings. We applied SuperCell with graining level 5 and number of kNN 5 to the simulated datasets with and without adding zeros. In addition, we attempted to impute the zeros we previously introduced in the simulated data to evaluate whether imputation helped better identification of differentially variable genes. We applied SAVER imputation on the simulated datasets after adding zeros. Distinct and scDD were applied to two randomly selected simulated databases (one for the non-differential variability between groups scenario, the other one for the differential variability scenario). They were not applied to the 100 simulated databases due to its intense computational cost. ScDD was applied using the Bayes factor permutation test with 100 iterations, as recommended in the reference handbook [236].

For pseudo-bulk scRNAseq analyses, we simulated data from five post-menopausal breast samples [239]. From 14,370 genes available, after conducting the same filtering conducted in the single cell level analyses, we kept 6744 genes. We aggregated data across each biological sample, thus having five samples with aggregated expression data

in each of the two groups. The two simulation settings were the same as the ones used at the single cell level.

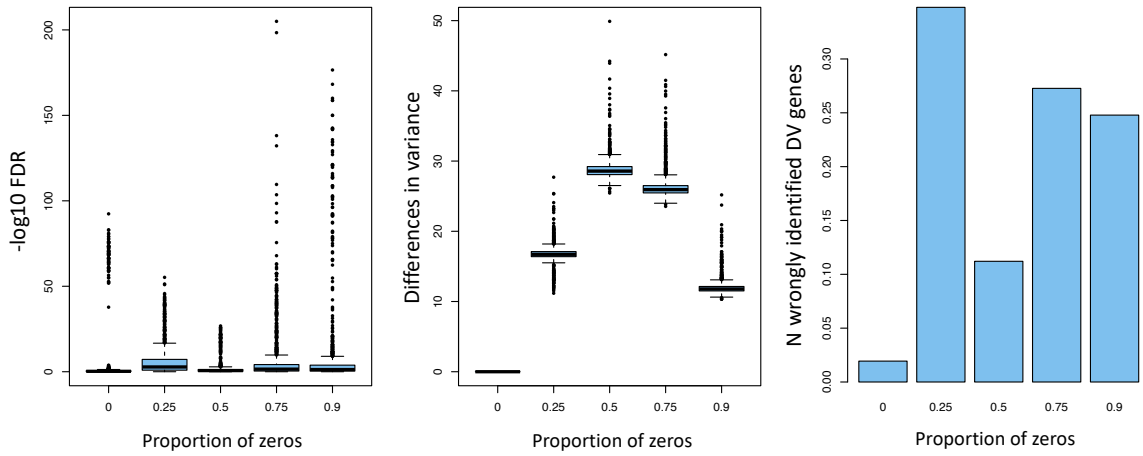
Results

Figure 5.1 shows the relationship between proportion of zeros and adjusted p-values from diffVar, differences in variance before and after inserting zeros, and number of wrongly identified differentially variable genes, in a randomly chosen database from the differential variability simulation setting after applying diffVar. Introducing zeros leads to more extreme p-values, more extreme differences in variances and to higher rates of wrongly identified differentially variable genes (false discoveries).

Figure 5.2 shows boxplots for the simulations in non-differentially variable groups settings (therefore, only FDR-s were computed). The boxplots show FDR-s for the raw simulated data, the simulated data after adding zeros, the simulated data after adding zeros and imputing zeros with SAVER, the simulated data after applying SuperCell and the simulated data after adding zeros and applying SuperCell.

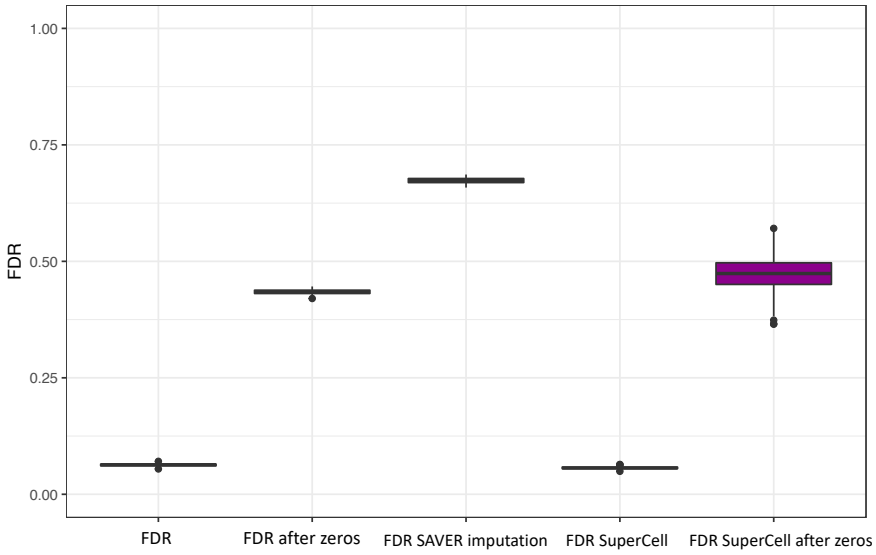
Figure 5.3 shows boxplots for the simulations in the setting in which 5 % of the genes are five times more differentially variable in one group as compared to the other. The boxplots show FDR-s and TPR-s. The distribution of the boxplots is quite flat in general, which shows all simulated datasets have similar behavior in terms of FDR-s and TPR-s.

Figure 5.1: Relationship between proportion of introduced zeros and A) adjusted p-values from diffVar, B) differences in variance before and after inserting zeros, and C) number of wrongly identified differentially variable genes by diffVar, in the differential variability simulation setting.



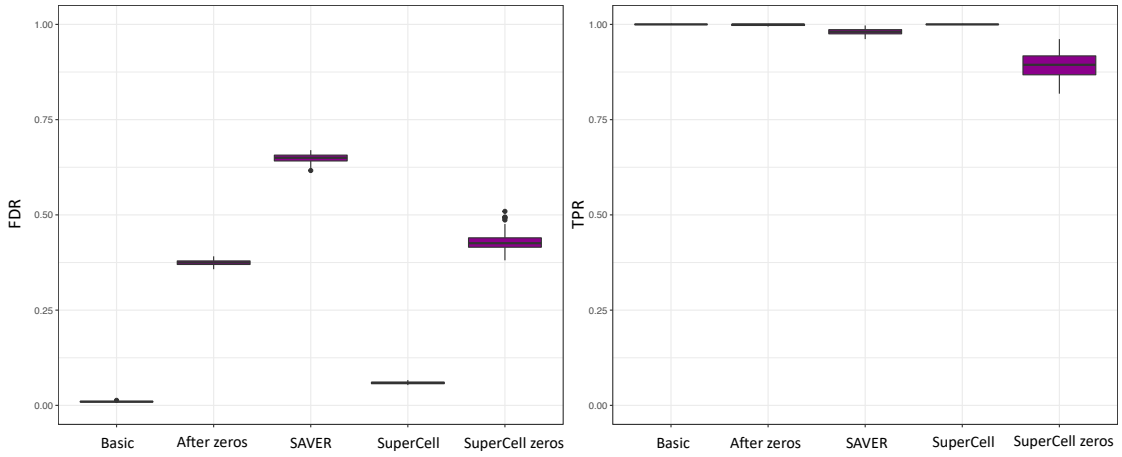
Abbreviations: N, number; DV, differentially variable; FDR, false discovery rate.

Figure 5.2: False discovery rates of diffVar for different simulation scenarios in the setting of non-differential variability between groups.



Abbreviations: FDR, false discovery rate.

Figure 5.3: False discovery and true positive rates of diffVar for different simulation scenarios in the setting of non-differential variability between groups.



Abbreviations: FDR, false discovery rate; TPR, true positive rate

Results for distinct and scDD are shown in Table 5.1. Distinct did not select any genes as differentially distributed in the non-differential variability between groups setting (FDR=0), and even when adding zeros, the FDR was much smaller as compared to diffVar, which shows a good control of FDR. In the differential variability between groups setting, the TPRs are as high as for diffVar, and the FDRs are much smaller than for diffVar. Conversely, scDD shows very high FDRs when adding zeros, as well as a decrease in TPR.

Table 5.1: False discovery and true positive rates for different simulation scenarios using the distinct and scDD algorithms.

	Non-differential variability setting	Differential variability setting (5 % of genes five times more variable)
distinct		
FDR (%)	0	11.1
FDR after adding zeros (%)	4.17	15.01
TPR (%)	-	100
TPR after adding zeros (%)	-	100
scDD		
FDR (%)	0.98	29.6
FDR after adding zeros (%)	46.1	54.7
TPR (%)	-	100
TPR after adding zeros (%)	-	74.4

Distinct was run with the default settings. Genes that had an adjusted p-value < 0.05 were considered significantly differentially distributed. ScDD was run with 100 permutations.

Abbreviations: FDR, false discovery rate; TPR, true positive rate.

When aggregating cells for the pseudo-bulk analysis, in the differential variability simulation setting, diffVar was not able to identify any differentially variable genes. Thus, both FDRs and TPRs were zero in all simulations (as there were no discoveries, neither false nor true). This supports that, even if differential variability between groups is present at the single cell level, when aggregating the data to the sample level, those differences cannot be identified. When adding zeros to the data, several differentially variable genes were identified (mean TPR=28.7 %, mean FDR=37.7 %). This reflects that, as noted before, adding zeros leads to inflated variances and to false positives.

Discussion

We studied the impact of increasingly introducing zeros when evaluating transcriptional variability between groups in scRNAseq data using the diffVar tool, and found that introducing zeros leads to inflated variances and p-values, as well as false discoveries. We tested several

alternative tools for differential variability analysis in scRNAseq data, and found that the distinct tool provides the best compromise between TPR and FDR for identification of differentially variable genes, even when adding zeros.

Controversy exists in the scientific community regarding high proportions of zeros in scRNAseq data. Some researchers see zeros as true biological signals representing no or low gene expression, whereas other scientists see zeros as missing data to be corrected [230]. Thus, there is no consensus on whether zeros in scRNAseq data should be used as valuable information, removed or imputed.

Another approach that has been used to handle zeros in scRNAseq data is to binarize gene expression data, which consists in considering as "1" all counts that are non-zero, and as "0" all zero counts. It has been proved that this approach can lead to reasonable cell clustering [230]. Nevertheless, it has the obvious limitation of the tremendous loss of information derived from treating highly and lowly expressed genes equally.

In our work, SuperCell, SAVER imputation and pseudo-bulk analysis did not provide any improvement in FDRs and TPRs as compared to the single cell level analyses. This shows that grouping expression levels by samples or into transcriptionally similar cells, which tends to be helpful for differential mean expression analysis, is not a desirable approach for identifying differences in variance, probably because the variability structure of the data gets lost when the data is grouped in smaller units. In fact, data imputation is controversial in the field of scRNAseq, as researchers have argued that imputation in scRNAseq data leads to decreased variability accross cells [241] and reduction of biological variation [231].

In summary, we found that the diffVar method, which showed good performance to identify differential variability in DNA methylation and bulk RNAseq data, does not perform well for scRNAseq data when the proportion of zeros is high. In contrast, the distinct tool, which is not specific for differential variability analyses, identified most of the truly differentially variable genes, and did not get inflated FDRs,

even after the addition of zeros. The non-parametric approach of this method might suit better for scRNAseq data. Specific methods for differential transcriptional variability assessment in scRNAseq data need to be developed. Assessing differential variability in scRNAseq might help the evaluation of functional implications of environment-induced DNA methylation changes.

CHAPTER 6

Conclusions and final remarks

This doctoral thesis constitutes a biostatistical toolkit for conducting statistical analysis to disentangle the role of DNA methylation data in environment-related chronic disease. We have addressed variable selection and effect estimation in ultra-high dimensional settings with high correlations, and we have also extended existing tools to evaluate the potential mediating role of multiple methylation markers on the association between exposures and outcomes to a time-to-event setting.

Our novel statistical tools have enabled us to identify mediated effects of DNA methylation on the association between arsenic and CVD, and between smoking and cancer. In addition, we have extended our research to other omics data types by exploring scRNAseq, which enables the discovery of the transcriptional heterogeneity between individual cells and can enable functional validation of epigenetic findings. In this line, we have proved that conventional statistical methods developed to identify differences in transcriptional variability in bulk RNAseq data do not work, in general, in presence of high proportions of zeros.

In our work regarding variable selection in the omics data setting (section 3), we paired the ISIS tool with Aenet, elastic-net and

MSAenet, and showed that ISIS-Aenet provides the best predictive ability for the continuous and dichotomous outcomes, while being consistent for effect estimation by fulfilling the oracle property. In addition, the bioinformatics analysis showed that ISIS-Aenet led to the most biologically meaningful selection of DMPs. This is evidence that the ISIS-Aenet tool is an improvement beyond existing methods for variable selection in ultra-high dimensional settings. In addition, our epidemiologic studies conducted in sections 3.2.2 and 4.4.2 showed that ISIS-Aenet and ISIS-enet are effective tools to select DMPs associated with CVD and cancer, respectively, as many of the DMPs that were selected by ISIS subsequently showed significant mediated effects in our mediation analysis.

However, the ISIS-Aenet model has some limitations. First, ISIS-Aenet notably outperformed the other methods in predictive ability for continuous and binary outcomes. Nevertheless, we were not able to fully explore the performance of ISIS-Aenet in survival and dichotomous outcomes, as bigger sample sizes are needed for those outcomes to obtain the same number of selected variables as for the continuous outcomes. Future work should include the development of an Aenet algorithm for Poisson data to evaluate whether better results can be obtained adapting time-to-event data to a Poisson model. On the other hand, computational cost is a major limitation of the ISIS-Aenet tool, indeed, a high performance computing cluster is needed to run these models. Future research should also focus on reducing the computational cost.

In the second part of this thesis, focused on extending the multimediate algorithm to time-to-event outcomes, we showed that multimediate leads to smaller MSEs and better CI coverage as compared to simple mediation, even in the setting of no correlations. Nevertheless, this model also presents some limitations beyond those mentioned in section 4.4.1 regarding convergency issues of the additive hazards models, assumption of multivariate normality and assumption of the correlations between mediators being independent of the exposure. In fact, this algorithm only handles the setting in which mediators are uncausally correlated, i.e., there are no causal associations between

mediators. Although this is a plausible setting for omics data, this algorithm should eventually be extended to the setting in which causal correlations between mediators exist.

Although DNA methylation shows good predictive ability for the health outcomes considered in this thesis, and shows evidence of a mediating role between environmental exposures and disease, establishing whether the association is causal or, conversely, DNA methylation is a biomarker of other disrupted biological processes, is challenging. The no unmeasured confounding assumption, which is essential to identify mediated effects, is impossible to verify in practice for observational studies [227]. Thus, sensitivity analyses are desirable to measure the impact of those potential unmeasured confounders in our mediated effects. Many sensitivity analysis techniques have been developed for mediation analysis, including for survival outcomes [242, 194, 227, 243, 138]. Relevant future work should include the adaptation of these sensitivity analysis techniques to the multimediate algorithm setting. In addition, experimental studies are needed to investigate whether the identified effects of DNA methylation on health outcomes are causal.

The last part of this thesis is focused on scRNAseq. This technology has opened a whole new horizon for the omics data research community, and could lead to unprecedented scientific discovery [87]. However, no consensus has been reached regarding the statistical analysis pipeline for the analysis of these data. For example, some researchers tend to assume that scRNAseq data follow a Poisson distribution, while others argue that a negative binomial distribution fits better [244]. In this work, we evidence the challenges of existing statistical tools to detect differences in variability in scRNAseq data. Although the distinct algorithm showed good performance for differential variability detection, specific tools for differential variability in the scRNAseq setting need to be developed.

Even though most of this thesis has been developed in the setting of DNA methylation data, with a small part focusing on gene expression data, these statistical tools could easily be adapted to other omics data. Indeed, the ultimate goal of the omics data field researchers

would be to be able to integrate all omics data together in statistical models, in order to maximize the information and to be able to use it for early detection and treatment development. Several efforts have been made to integrate different omics data, such as the Signature Regulatory Clustering (SiRCle) tool [245], which aims to integrate DNA methylation, RNA-seq and proteomics data. The integration of proteomics data, however, constitutes another statistical challenge, as proteomics data generally present a huge number of missing data, sometimes above 90 % [246]. Future research should focus on disentangling how each omics layer influences the subsequent layer and how to integrate all layers in statistical models.

In conclusion, our work has brought improvements in the statistical pipeline for DNA methylation data analysis, also extendable to other omics data types. Thus, this work is a contribution to the community of omics data research, both by providing novel statistical methods for DNA methylation data analysis and by contributing to the body of epidemiological evidence that supports the relevance of environmental epigenetics on chronic diseases.

CHAPTER 7

Scientific production during the PhD program

Coauthorship of R packages

1. *SIS: Sure Independence Screening* (CRAN repository).
2. *multimediate: https://github.com/AllanJe/multimediate.*

Peer-reviewed publications directly related with the subject of this doctoral thesis

1. **Domingo-Relloso A**, Gribble MO, Riffo-Campos AL, Haack K, Cole SA, Tellez-Plaza M, Umans JG, Fretts AM, Zhang Y, Fallin MD, Navas-Acien A, Everson TM. Epigenetics of type 2 diabetes and diabetes-related outcomes in the Strong Heart Study. *Clinical Epigenetics* 2022:14,177.
2. Lieberman-Cribbin W, **Domingo-Relloso A**, Navas-Acien A, Cole SA, Haack K, Umans J, Tellez-Plaza M, Colicino E, Baccarelli AA, Gao X, Kupsco A. Epigenetic biomarkers of lead exposure and cardiovascular disease: prospective evidence in the

Strong Heart Study. *Journal of the American Heart Association* 2022;11:e026934.

3. **Domingo-Relloso A**, Makhani K, Riffo-Campos AL, Tellez-Plaza M, Klein KO, Subedi P,..., Navas-Acien A. Arsenic Exposure, Blood DNA Methylation, and Cardiovascular Disease. *Circulation Research* 2022;131:e51–e69.

4. **Domingo-Relloso A**, Riffo-Campos AL, Powers M, Tellez-Plaza M, Haack K, Brown RH, Umans JG, Fallin MD, Cole SA, Navas-Acien A, Sanchez TR. An epigenome wide study of DNA methylation profles and lung function among American Indians in the Strong Heart Study. *Clinical Epigenetics* 2022, 14:75.

5. **Domingo-Relloso A**, Bozack AK, Kiihl S, Rodriguez-Hernandez Z, Rentero-Garrido P, Casasnovas JA, Leon-Latre M, Garcia-Barrera T, Gomez-Ariza JL, Moreno B, Cenarro C, de Marco G, Parvez F, Siddique AB, Shahriar H, Udin MN, Islam T, Navas-Acien A, Gamble M. Arsenic Exposure and Human Blood DNA Methylation and Hydroxymethylation Profiles in Two Diverse Populations from Bangladesh and Spain. *Environmental Research* 2022, 204 (B).

6. Navas-Acien A, **Domingo-Relloso A**, Subedi P, Riffo-Campos AL, Xia R, Gomez L, Haack K, Goldsmith J, Howard BV, Best LG, Devereux R, Tauqeer A, Zhang Y, Fretts AM, Pichler G, Daniel Levy D, Vasan RS, Baccarelli AA, MD, Herreros-Martinez M, Tang WY, Bressler J, Fornage M, Umans JG, Tellez-Plaza M, Fallin MD, Zhao J, Cole SA. Blood DNA Methylation and Incident Coronary Heart Disease. *JAMA Cardiology* 2021, 6 (11) 1237-1246.

7. **Domingo-Relloso A**, Haack K, Fallin DM, Terry MB, Rhoades DA, Tang WY, Herreros-Martinez M, Garcia-Esquinas E, Bozack AK, Cole SA, Tellez-Plaza M, Navas-Acien A. DNA methylation and cancer incidence: signals for lymphatic-hematopoietic vs. solid cancers from the Strong Heart Study. *Clinical Epigenetics* 2021, 13 (43).

8. Christiansen C, Castillo-Fernandez JE, **Domingo-Relloso A**, Zhao W, El-Sayed Moustafa JS, Tsai PC, Maddock J, Haack K, Cole SA, Kardia SLR, Molokhia M, Suderman M, Power C, Relton C, Wong A, Kuh D, Goodman A, Small KS, Smith JA, Tellez-Plaza M, Navas-Acien A, Ploubidis GB, Hardy R and Bell JT. Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. *Clinical Epigenetics* 2021, 13 (36).

9. Crocker KC, **Domingo-Relloso A**, Tellez-Plaza M, Haack K, Fretts M, Tang WY, Fallin DM, Cole SA, Navas-Acien A. DNA methylation and adiposity phenotypes: an epigenome-wide association study among adults in the Strong Heart Study. *International Journal of Obesity* 2020(11):2313-2322.

10. Bozack AK, **Domingo-Relloso A**, Tellez-Plaza M, Haack K, Gamble M, Umans JG, Best LG, Yracheta J, Gribble MO, Cardenas A, Francesconi KA, Goessler W, Tang WY, Fallin DM, Cole SA, Navas-Acien A. Loci-specific differential DNA methylation and urinary arsenic: An epigenome-wide association study among adults with low-to-moderate arsenic exposure. *Environmental Health Perspectives* 2020;128:6705.

11. **Domingo-Relloso A**, Riffo-Campos A, Haack K, Rentero-Garrido P, Ladd-Acosta C, Fallin D, Tang WY, Herreros-Martinez M, Gonzalez JR, Bozack A, Cole S, Navas-Acien A, Tellez-Plaza M. Cadmium, smoking, and human blood DNA methylation profiles in adults from the Strong Heart Study. *Environmental Health Perspectives* 2020;128:67005.

Peer-reviewed publications not directly related with the subject of this doctoral thesis

1. Galvez-Fernandez M, Powers M, Grau-Perez M, **Domingo-Relloso A**, Lolacono N, Goessler W, Zhang Y, Fretts A, Umans JG, Maruthur N, Navas-Acien A. Urinary Zinc and Incident Type 2 Diabetes: Prospective Evidence From the Strong Heart Study. *Diabetes Care* 2022,45(11):2561-2569.

2. Martinez-Ales G,[*] **Domingo-Relloso A**,[*] Quintana-Diaz M, Fernandez-Capitan C, Hernan MA. Thromboprophylaxis with standard-dose vs. flexible-dose heparin for hospitalized COVID-19 patients: a target trial emulation. *Journal of Clin Epi* 2022, 96-103 ([*] co-first authors).

3. Huan T, Nguyen S, Colicino E Ochoa-Rosales C, Hill WD, Brody JA, Soerensen M, Zhang Y, Baldassari A, Elhadad MA, Toshiko T, Zheng Y, **Domingo-Relloso A**, Lee DH, Ma J, Yao C, Liu C, Hwang SJ, Joehanes R, Fornage M, Bressler J, van Meurs JBJ, Debrabant B, Mengel-From J, Hjelmborg J, Christensen K, Vokonas P, Schwartz J, Gahrib SA, Sotoodehnia N, Sitlani CM, Kunze S, Gieger C, Peters A, Waldenberger M, Deary IJ, Ferrucci L, Qu Y, Greenland P, Lloyd-Jones DM, Hou L, Bandinelli S, Voortman T, Hermann B, Baccarelli A, Whitsel E, Pankow JS, Levy D. Integrative analysis of clinical and epigenetic biomarkers of mortality. *Aging Cell* 2022 (e13608).

4. Galvez-Fernandez M, Sanchez-Saez F, **Domingo-Relloso A**, Rodriguez-Hernandez Z, Tarazona S, Gonzalez-Marrachelli V, Grau-Perez M, Morales-Tatay JM, Amigo N, Garcia-Barrera T, Gomez-Ariza JL, Chaves FJ, Garcia-Garcia AB, Melero R, Tellez-Plaza M, Martin-Escudero JC, Redon J, Monleon D. Gene-environment interaction analysis of redox-related metals and genetic variants with plasma metabolic patterns in a general population from Spain: The Hortega Study. *Redox Biology* 2022:52(102314).

5. Zhao D, **Domingo-Relloso A**, Tellez-Plaza M, Nigra AE, Valeri L, Moon KA, Goessler W, Best LG, Ali T, Umans JG, Fretts A, Cole SA, Navas-Acien A. High level of selenium exposure in the Strong Heart Study: a cause for incident cardiovascular disease? *Antioxid Redox Signal* 2022;37(13-15):990-997.

6. Delgado-Velandia M, Gonzalez-Marrachelli V, **Domingo-Relloso A**, Galvez-Fernandez M, Grau-Perez M, Olmedo P, Galan I, Rodriguez-Artalejo F, Amigo N, Briongos-Figuero L, Redon J, Martin-Escudero JC, Monleon-Salvado D, Tellez-Plaza M, Sotos-

Prieto M. Healthy lifestyle, metabolomics and incident type 2 diabetes in a population-based cohort from Spain. *International Journal of Behavioral Nutrition and Physical Activity* 2022:19(8).

7. Grau-Perez M, Caballero-Mateos MJ, **Domingo-Relloso A**, Navas-Acien A, Gomez-Ariza JL, Garcia-Barrera T, Leon-Latre M, Soriano-Gil Z, Jarauta E, Cenarro A, Moreno-Franco B, Laclaustra M, Civeira F, Casasnovas JA, Guallar E, and Tellez-Plaza M. Toxic Metals and Subclinical Atherosclerosis in Carotid, Femoral, and Coronary Vascular Territories: The Aragon Workers Health Study. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2022, (42) 87–99.

8. Schlosser, P., Tin, A., Matias-Garcia, P.R., Thio C, Joehanes R, Liu R, Weihs A, Yu Z, Hoppmann A, Grundner-Culemann F, Min J, Adeyemo A, Agyemang C, Ärnlöv J, N Aziz NA, Baccarelli A, Bochud M, Brenner H, Breteler M, Carmeli C, Chaker L, Chambers J, Cole S, Coresh J, Corre T, Correa A, Cox S, de Klein N, Delgado G, **Domingo-Relloso A**, Eckardt KU, Ekici A, Endlich K, Evans K, Floyd J, Fornage M, Franke L, Fraszczyk E, Gao X, Gào X, Ghanbari M, Ghasemi S, Gieger G, Greenland P, Grove M, Harris S, Hemani G, Henneman P, Herder C, Horvath S, Hou L, Hurme M, Hwang SJ, Jarvelin MR, Kardia S, Kasela S, Kleber M, Koenig W, Kooner J, Kramer H, Kronenberg F, Kühnel B, Lehtimäki T, Lind L, Liu D, Liu Y, Lloyd-Jones D, Lohman K, Lorkowski S, Lu A, Marioni R, März W, McCartney D, Meeks K, Milani L, Mishra P, Nauck M, Navas-Acien A, Nowak C, Peters A, Prokisch H, Psaty B, Raitakari O, Ratliff S, Reiner A, Rosas S, Schottker B, Schwartz J, Sedaghat S, Smith J, Sotoodehnia N, Stocker H, Stringhini S, Sundstrom J, Swenson B, Tellez-Plaza M, van Meurs J, Van Vliet-Ostaptchouk J, Venema A, Verweij N, Walker R, Wielscher M, Winkelmann J, Wolffenbuttel B, Zhao W, Zheng Y, Estonian Biobank Research Team, Genetics of DNA Methylation Consortium, Loh M, Snieder H, Levy D, Waldenberger M, Susztak K, Kottgen A, Teumer A. Meta-analyses identify DNA methylation associated with kidney function and damage. *Nature Communications* 2021, 12 (7174).

9. Zhao D, Ilievski V, Slavkovich V, Olmedo P, **Domingo-Relloso A**, Rule AM, Kleiman NJ, Navas-Acien A, Hilpert M. Effects of e-liquid flavor, nicotine content, and puff duration on metal emissions from electronic cigarettes. *Environmental Research* 2022 204 (C).

10. Martinez-Ales G,* **Domingo-Relloso A**,* Arribas JR, Quintana-Diaz M, Hernan MA and the COVID@HULP Group. Critical Care Requirements Under Uncontrolled Transmission of SARS-CoV-2. *American Journal of Public Health* 2021 (111) 923:926 (* co-first authors).

11. Martinez-Ales G, Cruz Rodriguez JB, Lazaro P, **Domingo-Relloso A**, Barrigon ML, Angora R, Rodriguez-Vega B, Jimenez-Sola E, Sanchez-Castro P, Roman-Mazuecos E, Villoria L, Ortega A, Navio M, Stanley B, Rosenheck R, Baca-Garcia E, Bravo-Ortiz MF. Cost-effectiveness of a Contact Intervention and a Psychotherapeutic Program for Post-discharge Suicide Prevention. *Can J Psychiatry* 2021;66(8):737-746.

12. Shearston JA, Johnson AM, **Domingo-Relloso A**, Kioumourtzoglou MA, Hernandez D, Ross J, Chillrud SN, Hilpert M. Opening a Large Delivery Service Warehouse in the South Bronx: Impacts on Traffic, Air Pollution, and Noise. *Int. J. Environ. Res. Public Health* 2020, 17(9), 3208.

13. Tinkelman N, Spratlen M, **Domingo-Relloso A**, Tellez-Plaza M, Grau-Perez M, Francesconi K, Goessler W, Howard B, MacCluer J, North K, Umans J, Factor-Litvak P, Cole S, Navas-Acien A. Associations of maternal arsenic exposure with adult fasting glucose and insulin resistance in the Strong Heart Study and Strong Heart Family Study. *Environmental International* 2020, 137: 105531.

Bibliography

[1] Thomas Battram, Paul Yousefi, Gemma Crawford, Claire Prince, Mahsa Sheikhali Babaei, Gemma Sharp, Charlie Hatcher, María Jesús Vega-Salas, Sahar Khodabakhsh, Oliver Whitehurst, Ryan Langdon, Luke Mahoney, Hannah R. Elliott, Giulia Mancano, Matthew A. Lee, Sarah H. Watkins, Abigail C. Lay, Gibran Hemani, Tom R. Gaunt, Caroline L. Relton, James R. Staley, and Matthew Suderman. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Research*, 7(31):41, 2022.

[2] Ryo Yamada, Daigo Okada, Juan Wang, Tapati Basak, and Satoshi Koyama. Interpretation of omics data analyses. *Journal of Human Genetics*, 66(1):93–102, 2020.

[3] Estela G Toraño, María G García, Juan Luis Fernández-Morera, Pilar Niño-García, and Agustín F Fernández. The Impact of External Factors on the Epigenome: In Utero and over Lifetime. *Biomed Res Int.*, 2016:2568635, 2016.

[4] Stephen B Baylin and Peter A Jones. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol*, 1(8):9(019505), 2016.

[5] Eleni Stylianou. Epigenetics of chronic inflammatory diseases. *Journal of Inflammation Research*, 12(1):14, 2019.

[6] Shuk Mei Ho, Abby Johnson, Pheruza Tarapore, Vinothini Janakiram, Xiang Zhang, and Yuet Kin Leung. Environmental epigenetics and its implication on disease risk and health outcomes. *ILAR journal / National Research Council, Institute of Laboratory Animal Resources*, 53(3-4):289–305, 2012.

[7] Patrik Waldmann, Gábor Mészáros, Birgit Gredler, Christian Fuerst, and Johann Sölkner. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, 4:270, 2013.

[8] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[9] Allan Jérolon, Laura Baglietto, Etienne Birmelé, Flora Alarcon, and Vittorio Perduca. Causal mediation analysis in presence of multiple mediators uncausally related. *The International Journal of Biostatistics*, 17(2):191–221, 2020.

[10] C. H. Waddington. The epigenotype. 1942. *International journal of epidemiology*, 41(1):10–13, feb 2012.

[11] Huda Y. Zoghbi and Arthur L. Beaudet. Epigenetics and Human Disease. *Cold Spring Harbor Perspectives in Biology*, 8(2):1–28, feb 2016.

[12] Sorayya Ghasemi. Cancer's epigenetic drugs: where are they in the cancer medicines? *The Pharmacogenomics Journal*, 20(3):367–379, dec 2019.

[13] Lisa D Moore, Thuc Le, and Guoping Fan. DNA Methylation and Its Basic Function. *Neuropsychopharmacol*, 38:23–38, 2013.

[14] Pui Pik Law and Michelle L. Holland. DNA methylation at the crossroads of gene and environment interactions. *Essays in Biochemistry*, 63(6):717, 2019.

[15] Benjamin Lebecque, Céline Bourgne, Véronique Vidal, and Marc G. Berger. DNA Methylation and Intra-Clonal Het-

erogeneity: The Chronic Myeloid Leukemia Model. *Cancers*, 13(14):3587, jul 2021.

[16] Elizabeth M Martin and Rebecca C Fry. Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. *Annu. Rev. Public Health*, 39:309–333, 2018.

[17] Adrian Ruiz-Hernandez, Chin-Chi Kuo, Pilar Rentero-Garrido, Wan-Yee Tang, Josep Redon, Jose M Ordovas, Ana Navas-Acien, and Maria Tellez-Plaza. Environmental chemicals and DNA methylation in adults: a systematic review of the epidemiologic evidence. *Clinical Epigenetics*, 7(1):55, dec 2015.

[18] Roby Joehanes, Allan C Just, Riccardo E Marioni, Luke C Pilling, Lindsay M Reynolds, Pooja R Mandaviya, Weihua Guan, Tao Xu, Cathy E Elks, Stella Aslibekyan, Hortensia Moreno-Macias, Jennifer A Smith, Jennifer A Brody, Radhika Dhingra, Paul Yousefi, James S Pankow, Sonja Kunze, Sonia H Shah, Allan F McRae, Kurt Lohman, Jin Sha, Devin M Absher, Luigi Ferrucci, Wei Zhao, Ellen W Demerath, Jan Bressler, Megan L Grove, Tianxiao Huan, Chunyu Liu, Michael M Mendelson, Chen Yao, Douglas P Kiel, Annette Peters, Rui Wang-Sattler, Peter M Visscher, Naomi R Wray, John M Starr, Jingzhong Ding, Carlos J Rodriguez, Nicholas J Wareham, Marguerite R Irvin, Degui Zhi, Myrto Barrdahl, Paolo Vineis, Srikant Ambatipudi, André G Uitterlinden, Albert Hofman, Joel Schwartz, Elena Colicino, Lifang Hou, Pantel S Vokonas, Dena G Hernandez, Andrew B Singleton, Stefania Bandinelli, Stephen T Turner, Erin B Ware, Alicia K Smith, Torsten Klengel, Elisabeth B Binder, Bruce M Psaty, Kent D Taylor, Sina A Gharib, Brenton R Swenson, Liming Liang, Dawn L DeMeo, George T O'Connor, Zdenko Herceg, Kerry J Ressler, Karen N Conneely, Nona Sotoodehnia, Sharon L R Kardia, David Melzer, Andrea A Baccarelli, Joyce B J van Meurs, Isabelle Romieu, Donna K Arnett, Ken K Ong, Yongmei Liu, Melanie Waldenberger, Ian J Deary, Myriam Fornage, Daniel Levy, and Stephanie J London.

Epigenetic Signatures of Cigarette Smoking. *Circulation. Cardiovascular genetics*, 9(5):436–447, oct 2016.

[19] Stig E. Bojesen, Nicholas Timpson, Caroline Relton, George Davey Smith, and Børge G. Nordestgaard. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax*, 72(7):646–653, jan 2017.

[20] Xu Gao, Min Jia, Yan Zhang, Lutz Philipp Breitling, and Hermann Brenner. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clinical Epigenetics*, 7(1):113, dec 2015.

[21] Lindsay M. Reynolds, Ma Wan, Jingzhong Ding, Jackson R. Taylor, Kurt Lohman, Dan Su, Brian D. Bennett, Devin K. Porter, Ryan Gimple, Gary S. Pittman, Xuting Wang, Timothy D. Howard, David Siscovick, Bruce M. Psaty, Steven Shea, Gregory L. Burke, David R. Jacobs, Stephen S. Rich, James E. Hixson, James H. Stein, Hendrik Stunnenberg, R. Graham Barr, Joel D. Kaufman, Wendy S. Post, Ina Hoeschele, David M. Herrington, Douglas A. Bell, and Yongmei Liu. DNA Methylation of the Aryl Hydrocarbon Receptor Repressor Associations With Cigarette Smoking and Subclinical Atherosclerosis. *Circulation: Cardiovascular Genetics*, 8(5):707–716, oct 2015.

[22] IARC Working Group. Agents Classified by the IARC Monographs, Volumes 1–132. `https://monographs.iarc.who.int/agents-classified-by-the-iarc/`, 2022. [Online; accessed 25-December-2022].

[23] Thomas Haarmann-Stemmann and Josef Abel. The arylhydrocarbon receptor repressor (AhRR): structure, expression, and function. *Biological chemistry*, 387(9):1195–1199, sep 2006.

[24] Dorothea M. Heuberger and Reto A. Schuepbach. Protease-activated receptors (PARs): Mechanisms of action and potential therapeutic modulators in PAR-driven inflammatory diseases. *Thrombosis Journal*, 17(1):1–24, mar 2019.

[25] IARC Working Group. Arsenic, Metals, Fibres, and Dusts. `https://monographs.iarc.who.int/wp-content/uploads/2018/06/mono100C.pdf`, 2012. [Online; accessed 25-December-2022].

[26] Katherine A. Moon, Eliseo Guallar, Jason G. Umans, Richard B. Devereux, Lyle G. Best, Kevin A. Francesconi, Walter Goessler, Jonathan Pollak, Ellen K. Silbergeld, Barbara V. Howard, and Ana Navas-Acien. Association Between Exposure to Low to Moderate Arsenic Levels and Incident Cardiovascular Disease. *Annals of Internal Medicine*, 159(10):649, sep 2013.

[27] National Research Council. *Critical aspects of EPA's IRIS assessment of inorganic arsenic: Interim report.* National Academies Press, dec 2013.

[28] Chien-Jen Chen, Hung-Yi Chiou, Ming-Hsi Chiang, Li-Ju Lin, and Tong-Yuan Tai. Dose-Response Relationship Between Ischemic Heart Disease Mortality and Long-term Arsenic Exposure. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 16(4):504–510, apr 1996.

[29] Yu Mei Hsueh, Wen Lin Wu, Ya Li Huang, Hung Yi Chiou, Chin Hsiao Tseng, and Chien Jen Chen. Low serum carotene level and increased risk of ischemic heart disease related to long-term arsenic exposure. *Atherosclerosis*, 141(2):249–257, dec 1998.

[30] Maria Monrad, Annette Kjær Ersbøll, Mette Sørensen, Rikke Baastrup, Birgitte Hansen, Anders Gammelmark, Anne Tjønneland, Kim Overvad, and Ole Raaschou-Nielsen. Low-level arsenic in drinking water and risk of incident myocardial infarction: A cohort study. *Environmental Research*, 154:318–324, 2017.

[31] Lingqian Xu, Debapriya Mondal, and David A. Polya. Positive association of cardiovascular disease (CVD) with chronic exposure to drinking water arsenic (As) at concentrations below the WHO provisional guideline value: A systematic review and

meta-analysis. *International Journal of Environmental Research and Public Health*, 17(7):2536, apr 2020.

[32] Jonathan D. Newman, Ana Navas-Acien, Chin Chi Kuo, Eliseo Guallar, Barbara V. Howard, Richard R. Fabsitz, Richard B. Devereux, Jason G. Umans, Kevin A. Francesconi, Walter Goessler, Lyle T. Best, and Maria Tellez-Plaza. Peripheral arterial disease and its association with arsenic exposure and metabolism in the strong heart study. *American Journal of Epidemiology*, 184(11):806–817, dec 2016.

[33] Yu Chen, Joseph H. Graziano, Faruque Parvez, Mengling Liu, Vesna Slavkovich, Tara Kalra, Maria Argos, Tariqul Islam, Alauddin Ahmed, Muhammad Rakibuz-Zaman, Rabiul Hasan, Golam Sarwar, Diane Levy, Alexander Van Geen, and Habibul Ahsan. Arsenic exposure from drinking water and mortality from cardiovascular disease in Bangladesh: Prospective cohort study. *BMJ*, 5(342):d2431, may 2011.

[34] Ma José Medrano, Raquel Boix, Roberto Pastor-Barriuso, Margarita Palau, Javier Damián, Rebeca Ramis, José Luis del Barrio, and Ana Navas-Acien. Arsenic in public water supplies and cardiovascular mortality in Spain. *Environmental Research*, 110(5):448–454, jul 2010.

[35] Lalita N. Abhyankar, Miranda R. Jones, Eliseo Guallar, and Ana Navas-Acien. Arsenic exposure and hypertension: a systematic review. *Environmental health perspectives*, 120(4):494–500, apr 2012.

[36] Farrah J. Mateen, Maria Grau-Perez, Jonathan S. Pollak, Katherine A. Moon, Barbara V. Howard, Jason G. Umans, Lyle G. Best, Kevin A. Francesconi, Walter Goessler, Ciprian Crainiceanu, Eliseo Guallar, Richard B. Devereux, Mary J. Roman, and Ana Navas-Acien. Chronic arsenic exposure and risk of carotid artery disease: The Strong Heart Study. *Environmental Research*, 157:127–134, 2017.

[37] Chih Hao Wang, Jiann Shing Jeng, Ping Keung Yip, Chi Ling Chen, Lin I. Hsu, Yu Mei Hsueh, Hung Yi Chiou, Meei Mann Wu, and Chien Jen Chen. Biological gradient between long-term arsenic exposure and carotid atherosclerosis. *Circulation*, 105(15):1804–1809, apr 2002.

[38] Kiran Makhani, Christopher Chiavatti, Dany Plourde, Luis Fernando Negro Silva, Maryse Lemaire, Catherine A. Lemarié, Stephanie Lehoux, and Koren K. Mann. Using the Apolipoprotein E Knock-Out Mouse Model to Define Atherosclerotic Plaque Changes Induced by Low Dose Arsenic. *Toxicological sciences : an official journal of the Society of Toxicology*, 166(1):213–218, nov 2018.

[39] Luis Fernando Negro Silva, Maryse Lemaire, Catherine A. Lemarié, Dany Plourde, Alicia M. Bolt, Christopher Chiavatti, D. Scott Bohle, Vesna Slavkovich, Joseph H. Graziano, Stéphanie Lehoux, and Koren K. Mann. Effects of inorganic arsenic, methylated arsenicals, and arsenobetaine on atherosclerosis in the apoE-/- mouse model and the role of as3mt-mediated methylation. *Environmental Health Perspectives*, 125(7):077001, jul 2017.

[40] Kathryn Demanelis, Maria Argos, Lin Tong, Justin Shinkle, Mekala Sabarinathan, Muhammad Rakibuz-Zaman, Golam Sarwar, Hasan Shahriar, Tariqul Islam, Mahfuzar Rahman, Mohammad Yunus, Joseph H. Graziano, Karin Broberg, Karin Engström, Farzana Jasmine, Habibul Ahsan, and Brandon L. Pierce. Association of Arsenic Exposure with Whole Blood DNA Methylation: An Epigenome-Wide Study of Bangladeshi Adults. *Environmental Health Perspectives*, 127(5):057011, may 2019.

[41] Maria Argos, Lin Chen, Farzana Jasmine, Lin Tong, Brandon L. Pierce, Shantanu Roy, Rachelle Paul-Brutus, Mary V. Gamble, Kristin N. Harper, Faruque Parvez, Mahfuzar Rahman, Muhammad Rakibuz-Zaman, Vesna Slavkovich, John A. Baron, Joseph H. Graziano, Muhammad G. Kibriya, and Habibul Ahsan. Gene-Specific Differential DNA Methylation and Chronic

Arsenic Exposure in an Epigenome-Wide Association Study of Adults in Bangladesh. *Environmental Health Perspectives*, 123(1):64–71, jan 2015.

[42] K. Broberg, S. Ahmed, K. Engström, M. B. Hossain, S. Jurkovic Mlakar, M. Bottai, M. Grandér, R. Raqib, and M. Vahter. Arsenic exposure in early pregnancy alters genome-wide DNA methylation in cord blood, particularly in boys. *Journal of Developmental Origins of Health and Disease*, 5(4):288–298, 2014.

[43] Anda R. Gliga, Karin Engström, Maria Kippler, Helena Skröder, Sultan Ahmed, Marie Vahter, Rubhana Raqib, and Karin Broberg. Prenatal arsenic exposure is associated with increased plasma IGFBP3 concentrations in 9-year-old children partly via changes in DNA methylation. *Archives of Toxicology*, 92(8):2487–2500, aug 2018.

[44] Molly L. Kile, E. Andres Houseman, Andrea A. Baccarelli, Quazi Quamruzzaman, Mahmuder Rahman, Golam Mostofa, Andres Cardenas, Robert O. Wright, and David C. Christiani. Effect of prenatal arsenic exposure on DNA methylation and leukocyte subpopulations in cord blood. *Epigenetics*, 9(5):774–782, feb 2014.

[45] Daniel Rojas, Julia E. Rager, Lisa Smeester, Kathryn A. Bailey, Zuzana Drobná, Marisela Rubio-Andrade, Miroslav Stýblo, Gonzalo García-Vargas, and Rebecca C. Fry. Prenatal arsenic exposure and the epigenome: identifying sites of 5-methylcytosine alterations that predict functional changes in gene expression in newborn cord blood and subsequent birth outcomes. *Toxicological sciences : an official journal of the Society of Toxicology*, 143(1):97–106, jan 2015.

[46] Syeda Shegufta Ameer, Karin Engström, Mohammad Bakhtiar Hossain, Gabriela Concha, Marie Vahter, and Karin Broberg. Arsenic exposure from drinking water is associated with decreased gene expression and increased DNA methylation in peripheral blood. *Toxicology and Applied Pharmacology*, 321:57–66, apr 2017.

[47] Akhilesh Kaushal, Hongmei Zhang, Wilfried J J Karmaus, Todd M Everson, Carmen J Marsit, Margaret R Karagas, Shih-Fen Tsai, Hui-Ju Wen, and Shu-Li Wang. Genome-wide DNA methylation at birth in relation to in utero arsenic exposure and the associated health in later life. 16(1):50, 2017.

[48] Xiaojuan Guo, Xushen Chen, Jie Wang, Zhiyue Liu, Daniel Gaile, Hongmei Wu, Guan Yu, Guangyun Mao, Zuopeng Yang, Zhen Di, Xiuqing Guo, Li Cao, Peiye Chang, Binxian Kang, Jinyu Chen, Wen Gao, and Xuefeng Ren. Multi-generational impacts of arsenic exposure on genome-wide DNA methylation and the implications for arsenic-induced skin lesions. *Environment International*, 119:250–263, oct 2018.

[49] Anne K. Bozack, Arce Domingo-Relloso, Karin Haack, Mary V. Gamble, Maria Tellez-Plaza, Jason G. Umans, Lyle G. Best, Joseph Yracheta, Matthew O. Gribble, Andres Cardenas, Kevin A. Francesconi, Walter Goessler, Wan-Yee Tang, M. Daniele Fallin, Shelley A. Cole, and Ana Navas-Acien. Locus-Specific Differential DNA Methylation and Urinary Arsenic: An Epigenome-Wide Association Study in Blood among Adults with Low-to-Moderate Arsenic Exposure. *Environmental Health Perspectives*, 128(6):067015, jun 2020.

[50] Andres Cardenas, Devin C. Koestler, E. Andres Houseman, Brian P. Jackson, Molly L. Kile, Margaret R. Karagas, and Carmen J. Marsit. Differential DNA methylation in umbilical cord blood of infants exposed to mercury and arsenic in utero. *Epigenetics*, 10(6):508–515, jan 2015.

[51] Benjamin B. Green, Margaret R. Karagas, Tracy Punshon, Brian P. Jackson, David J. Robbins, E. Andres Houseman, and Carmen J. Marsit. Epigenome-wide assessment of DNA methylation in the placenta and arsenic exposure in the New Hampshire Birth Cohort Study (USA). *Environmental Health Perspectives*, 124(8):1253–1260, 2016.

[52] John F. Reichard, Michael Schnekenburger, and Alvaro Puga. Long term low-dose arsenic exposure induces loss of DNA methy-

lation. *Biochemical and biophysical research communications*, 352(1):188, jan 2007.

[53] Partha M. Das and Rakesh Singal. DNA methylation and cancer. *Journal of Clinical Oncology*, 22(22):4632–4642, sep 2004.

[54] Matthias Wielscher, Pooja R. Mandaviya, Brigitte Kuehnel, Roby Joehanes, Rima Mustafa, Oliver Robinson, Yan Zhang, Barbara Bodinier, Esther Walton, Pashupati P. Mishra, Pascal Schlosser, Rory Wilson, Pei Chien Tsai, Saranya Palaniswamy, Riccardo E. Marioni, Giovanni Fiorito, Giovanni Cugliari, Ville Karhunen, Mohsen Ghanbari, Bruce M. Psaty, Marie Loh, Joshua C. Bis, Benjamin Lehne, Nona Sotoodehnia, Ian J. Deary, Marc Chadeau-Hyam, Jennifer A. Brody, Alexia Cardona, Elizabeth Selvin, Alicia K. Smith, Andrew H. Miller, Mylin A. Torres, Eirini Marouli, Xin Gào, Joyce B.J. van Meurs, Johanna Graf-Schindler, Wolfgang Rathmann, Wolfgang Koenig, Annette Peters, Wolfgang Weninger, Matthias Farlik, Tao Zhang, Wei Chen, Yujing Xia, Alexander Teumer, Matthias Nauck, Hans J. Grabe, Macus Doerr, Terho Lehtimäki, Weihua Guan, Lili Milani, Toshiko Tanaka, Krista Fisher, Lindsay L. Waite, Silva Kasela, Paolo Vineis, Niek Verweij, Pim van der Harst, Licia Iacoviello, Carlotta Sacerdote, Salvatore Panico, Vittorio Krogh, Rosario Tumino, Evangelia Tzala, Giuseppe Matullo, Mikko A. Hurme, Olli T. Raitakari, Elena Colicino, Andrea A. Baccarelli, Mika Kähönen, Karl Heinz Herzig, Shengxu Li, Karen N. Conneely, Jaspal S. Kooner, Anna Köttgen, Bastiaan T. Heijmans, Panos Deloukas, Caroline Relton, Ken K. Ong, Jordana T. Bell, Eric Boerwinkle, Paul Elliott, Hermann Brenner, Marian Beekman, Daniel Levy, Melanie Waldenberger, John C. Chambers, Abbas Dehghan, and Marjo Riitta Järvelin. DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. *Nature Communications*, 13(1):1–14, may 2022.

[55] Thomas Vaissière, Rayjean J. Hung, David Zaridze, Anush Moukeria, Cyrille Cuenin, Virginie Fasolo, Gilles Ferro, Anupam

Paliwal, Pierre Hainaut, Paul Brennan, Jörg Tost, Paolo Boffetta, and Zdenko Herceg. Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors. *Cancer Research*, 69(1):243–252, jan 2009.

[56] Rejane Hughes Carvalho, Jun Hou, Vanja Haberle, Joachim Aerts, Frank Grosveld, Boris Lenhard, and Sjaak Philipsen. Genomewide DNA methylation analysis identifies novel methylated genes in non-small-cell lung carcinomas. *Journal of Thoracic Oncology*, 8(5):562–573, may 2013.

[57] Els Wauters, Wim Janssens, Johan Vansteenkiste, Herbert Decaluwé, Nele Heulens, Bernard Thienpont, Hui Zhao, Dominiek Smeets, Xavier Sagaert, Johan Coolen, Marc Decramer, Adrian Liston, Paul De Leyn, Matthieu Moisse, and Diether Lambrechts. DNA methylation profiling of non-small cell lung cancer reveals a COPD-driven immune-related signature. *Thorax*, 70(12):1113–1122, dec 2015.

[58] Jeffrey A. Tsou, Linda Y.C. Shen, Kimberly D. Siegmund, Tiffany I. Long, Peter W. Laird, Chandrika K. Seneviratne, Michael N. Koss, Harvey I. Pass, Jeffrey A. Hagen, and Ite A. Laird-Offringa. Distinct DNA methylation profiles in malignant mesothelioma, lung adenocarcinoma, and non-tumor lung. *Lung Cancer*, 47(2):193–204, feb 2005.

[59] Maria Moksnes Bjaanæs, Thomas Fleischer, Ann Rita Halvorsen, Antoine Daunay, Florence Busato, Steinar Solberg, Lars Jørgensen, Elin Kure, Hege Edvardsen, Anne Lise Børresen-Dale, Odd Terje Brustugun, Jörg Tost, Vessela Kristensen, and Åslaug Helland. Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Molecular Oncology*, 10(2):330–343, feb 2016.

[60] Martha L Slattery, John D Potter, Gary D Friedman, Khe-Ni Ma, and Sandra Edwards. Tobacco use and colon cancer. *Int. J. Cancer*, 70(3):259–264, 1997.

[61] Inger T. Gram, Song Yi Park, Lynne R. Wilkens, Christopher A. Haiman, and Loïc Le Marchand. Smoking-Related Risks of Colorectal Cancer by Anatomical Subsite and Sex. *American Journal of Epidemiology*, 189(6):543, jun 2020.

[62] Jessica L. Petrick, Peter T. Campbell, Jill Koshiol, Jake E. Thistle, Gabriella Andreotti, Laura E. Beane-Freeman, Julie E. Buring, Andrew T. Chan, Dawn Q. Chong, Michele M. Doody, Susan M. Gapstur, John Michael Gaziano, Edward Giovannucci, Barry I. Graubard, I. Min Lee, Linda M. Liao, Martha S. Linet, Julie R. Palmer, Jenny N. Poynter, Mark P. Purdue, Kim Robien, Lynn Rosenberg, Catherine Schairer, Howard D. Sesso, Rashmi Sinha, Meir J. Stampfer, Marcia Stefanick, Jean Wactawski-Wende, Xuehong Zhang, Anne Zeleniuch-Jacquotte, Neal D. Freedman, and Katherine A. McGlynn. Tobacco, alcohol use and risk of hepatocellular carcinoma and intrahepatic cholangiocarcinoma: The Liver Cancer Pooling Project. *British Journal of Cancer*, 118(7):1005, apr 2018.

[63] American Cancer Society. Liver Cancer Risk Factors. `https://www.cancer.org/cancer/liver-cancer/causes-risks-prevention/risk-factors.html`, 2019. [Online; accessed 04-January-2023].

[64] Jay D. Hunt, Olga L. Van Der Hel, Garnett P. McMillan, Paolo Boffetta, and Paul Brennan. Renal cell carcinoma in relation to cigarette smoking: meta-analysis of 24 studies. *International journal of cancer*, 114(1):101–108, mar 2005.

[65] American Cancer Society. Kidney Cancer Risk Factors. `https://www.cancer.org/cancer/kidney-cancer/causes-risks-prevention/risk-factors.html`, 2020. [Online; accessed 04-January-2023].

[66] Stephen J. Pandol, Minoti V. Apte, Jeremy S. Wilson, Anna S. Gukovskaya, and Mouad Edderkaoui. The Burning Question: Why is Smoking a Risk Factor for Pancreatic Cancer? *Pancreatology*, 12(4):344, 2012.

[67] American Cancer Society. Pancreatic Cancer Risk Factors. `https://www.cancer.org/cancer/pancreatic-cancer/causes-risks-prevention/risk-factors.html`, 2020. [Online; accessed 04-January-2023].

[68] Delphine Praud, Matteo Rota, Claudio Pelucchi, Paola Bertuccio, Tiziana Rosso, Carlotta Galeone, Zuo Feng Zhang, Keitaro Matsuo, Hidemi Ito, Jinfu Hu, Kenneth C. Johnson, Guo Pei Yu, Domenico Palli, Monica Ferraroni, Joshua Muscat, Nuno Lunet, Bárbara Peleteiro, Reza Malekzadeh, Weimin Ye, Huan Song, David Zaridze, Dmitry Maximovitch, Nuria Aragonés, Gemma Castaño-Vinyals, Jesus Vioque, Eva M. Navarrete-Muñoz, Mohammadreza Pakseresht, Farhad Pourfarzi, Alicja Wolk, Nicola Orsini, Andrea Bellavia, Niclas Håkansson, Lina Mu, Roberta Pastorino, Robert C. Kurtz, Mohammad H. Derakhshan, Areti Lagiou, Pagona Lagiou, Paolo Boffetta, Stefania Boccia, Eva Negri, and Carlo La Vecchia. Cigarette smoking and gastric cancer in the Stomach Cancer Pooling (StoP) Project. *European journal of cancer prevention*, 27(2):124–133, 2018.

[69] American Cancer Society. Stomach Cancer Risk Factors. `https://www.cancer.org/cancer/stomach-cancer/causes-risks-prevention/risk-factors.html`, 2021. [Online; accessed 04-January-2023].

[70] Arce Domingo-Relloso, Tianxiao Huan, Karin Haack, Angela L. Riffo-Campos, Daniel Levy, M. Daniele Fallin, Mary Beth Terry, Ying Zhang, Dorothy A. Rhoades, Miguel Herreros-Martinez, Esther Garcia-Esquinas, Shelley A. Cole, Maria Tellez-Plaza, and Ana Navas-Acien. DNA methylation and cancer incidence: lymphatic–hematopoietic versus solid cancers in the Strong Heart Study. *Clinical Epigenetics*, 13(1):43, dec 2021.

[71] Yves J.R. Menezo, Erica Silvestris, Brian Dale, and Kay Elder. Oxidative stress and alterations in DNA methylation: two sides of the same coin in reproduction. *Reproductive biomedicine online*, 33(6):668–683, dec 2016.

[72] Gopal Gopisetty, Kavitha Ramachandran, and Rakesh Singal. DNA methylation and apoptosis. *Molecular Immunology*, 43(11):1729–1740, apr 2006.

[73] James G. Herman. Hypermethylation of tumor suppressor genes in cancer. *Seminars in cancer biology*, 9(5):359–367, 1999.

[74] Jinke Sui, Xianrui Wu, Chenyang Wang, Guoqiang Wang, Chengcheng Li, Jing Zhao, Yuzi Zhang, Jianxing Xiang, Yu Xu, Weiqi Nian, Fuao Cao, Guanyu Yu, Zheng Lou, Liqiang Hao, Lianjie Liu, Bingsi Li, Zhihong Zhang, Shangli Cai, Hao Liu, Ping Lan, and Wei Zhang. Discovery and validation of methylation signatures in blood-based circulating tumor cell-free DNA in early detection of colorectal carcinoma: a case–control study. *Clinical Epigenetics*, 13(1):1–10, dec 2021.

[75] Warwick J. Locke, Dominic Guanzon, Chenkai Ma, Yi Jin Liew, Konsta R. Duesing, Kim Y.C. Fung, and Jason P. Ross. DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Frontiers in Genetics*, 10:1150, nov 2019.

[76] Nature Research Custom Media. Epigenetic analysis for early cancer detection. `https://www.nature.com/articles/d42473-020-00273-y`. [Online; accessed 04-January-2023].

[77] Andrew E. Teschendorff and Caroline L. Relton. Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics*, 19(3):129–147, nov 2017.

[78] Dhruvajyoti Roy and Maarit Tiirikainen. Diagnostic Power of DNA Methylation Classifiers for Early Detection of Cancer. *Trends Cancer*, 6(2):78–81, 2020.

[79] Ana Navas-Acien, Arce Domingo-Relloso, Pooja Subedi, Angela L. Riffo-Campos, Rui Xia, Lizbeth Gomez, Karin Haack, Jeff Goldsmith, Barbara V. Howard, Lyle G. Best, Richard Devereux, Ali Tauqeer, Ying Zhang, Amanda M. Fretts, Gernot Pichler, Daniel Levy, Ramachandran S. Vasan, Andrea A. Baccarelli, Miguel Herreros-Martinez, Wan Yee Tang, Jan

Bressler, Myriam Fornage, Jason G. Umans, Maria Tellez-Plaza, M. Daniele Fallin, Jinying Zhao, and Shelley A. Cole. Blood DNA Methylation and Incident Coronary Heart Disease: Evidence From the Strong Heart Study. *JAMA Cardiology*, 6(11):1237–1246, nov 2021.

[80] Rohan Lowe, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. Transcriptomics technologies. *PLoS computational biology*, 13(5):e1005457, may 2017.

[81] Sergei Borukhov and Evgeny Nudler. RNA polymerase: the vehicle of transcription. *Trends in Microbiology*, 16(3):126–134, mar 2008.

[82] Ryan A. Irvine, Iping G. Lin, and Chih-Lin Hsieh. DNA Methylation Has a Local Effect on Transcription and Histone Acetylation. *Molecular and Cellular Biology*, 22(19):6689, oct 2002.

[83] C. Jake Harris, Marion Scheibe, Somsakul Pop Wongpalee, Wanlu Liu, Evan M. Cornett, Robert M. Vaughan, Xueqin Li, Wei Chen, Yan Xue, Zhenhui Zhong, Linda Yen, William D. Barshop, Shima Rayatpisheh, Javier Gallego-Bartolome, Martin Groth, Zonghua Wang, James A. Wohlschlegel, Jiamu Du, Scott B. Rothbart, Falk Butter, and Steven E. Jacobsen. A DNA methylation reader complex that enhances gene transcription. *Science (New York, N.Y.)*, 362(6419):1182, dec 2018.

[84] Yen Ching Lim, Jie Li, Yiyun Ni, Qi Liang, Junjiao Zhang, George S.H. Yeo, Jianxin Lyu, Shengnan Jin, and Chunming Ding. A complex association between DNA methylation and gene expression in human placenta at first and third trimesters. *PLoS ONE*, 12(7):e0181155, jul 2017.

[85] Shuxiang Li, Yunhui Peng, and Anna R. Panchenko. DNA methylation: Precise modulation of chromatin structure and dynamics. *Current Opinion in Structural Biology*, 75:102430, aug 2022.

[86] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57, jan 2009.

[87] Sarah Aldridge and Sarah A. Teichmann. Single cell transcriptomics comes of age. *Nature Communications*, 11(1):1–4, aug 2020.

[88] Assieh Saadatpour, Shujing Lai, Guoji Guo, and Guo Cheng Yuan. Single-cell analysis in cancer genomics. *Trends in genetics : TIG*, 31(10):576, oct 2015.

[89] Arce Domingo-Relloso, Kiran Makhani, Angela L. Riffo-Campos, Maria Tellez-Plaza, Kathleen Oros Klein, Pooja Subedi, Jinying Zhao, Katherine A. Moon, Anne K. Bozack, Karin Haack, Walter Goessler, Jason G. Umans, Lyle G. Best, Ying Zhang, Miguel Herreros-Martinez, Ronald A. Glabonjat, Kathrin Schilling, Marta Galvez-Fernandez, Jack W. Kent, Tiffany R. Sanchez, Kent D. Taylor, W. Craig Johnson, Peter Durda, Russell P. Tracy, Jerome I. Rotter, Stephen S. Rich, David Van Den Berg, Silva Kasela, Tuuli Lappalainen, Ramachandran S. Vasan, Roby Joehanes, Barbara V. Howard, Daniel Levy, Kurt Lohman, Yongmei Liu, M. Daniele Fallin, Shelley A. Cole, Koren K. Mann, and Ana Navas-Acien. Arsenic Exposure, Blood DNA Methylation, and Cardiovascular Disease. *Circulation Research*, 131(2):E51–E69, jul 2022.

[90] Yang Feng. SIS: Sure Independence Screening. `https://cran.r-project.org/web/packages/SIS/`, 2008. [Online; accessed 25-December-2022].

[91] Elisa T. Lee, Thomas K. Welty, Richard Fabsitz, Linda D. Cowan, Ngoc Anh Le, Arvo J. Oopik, Andrew J. Cucchiara, Peter J. Savage, and Barbara V. Howard. The Strong Heart Study. A study of cardiovascular disease in American Indians: design and methods. *American journal of epidemiology*, 132(6):1141–1155, 1990.

[92] Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363, may 2014.

[93] Timothy J. Triche, Daniel J. Weisenberger, David Van Den Berg, Peter W. Laird, and Kimberly D. Siegmund. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic acids research*, 41(7):e90, apr 2013.

[94] Jean Philippe Fortin, Timothy J. Triche, and Kasper D. Hansen. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics (Oxford, England)*, 33(4):558–560, feb 2017.

[95] Zongli Xu, Liang Niu, and Jack A. Taylor. The ENmix DNA methylation analysis pipeline for Illumina BeadChip and comparisons with seven other preprocessing pipelines. *Clinical Epigenetics*, 13(1):216, dec 2021.

[96] Daniel L. McCartney, Rosie M. Walker, Stewart W. Morris, Andrew M. McIntosh, David J. Porteous, and Kathryn L. Evans. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics Data*, 26(9):22–24, sep 2016.

[97] Eugene A. Houseman, William P. Accomando, Devin C. Koestler, Brock C. Christensen, Carmen J. Marsit, Heather H. Nelson, John K. Wiencke, and Karl T. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13:86, may 2012.

[98] Salas LA and Koestler DC. Flowsorted.blood.epic: Illumina epic data on immunomagnetic sorted peripheral adult blood cells, 2022. R package version 2.2.0.

[99] Richard T. Barfield, Lynn M. Almli, Varun Kilaru, Alicia K. Smith, Kristina B. Mercer, Richard Duncan, Torsten Klengel, Divya Mehta, Elisabeth B. Binder, Michael P. Epstein, Kerry J.

Ressler, and Karen N. Conneely. Accounting for Population Stratification in DNA Methylation Studies. *Genetic Epidemiology*, 38(3):231–241, apr 2014.

[100] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882, mar 2012.

[101] Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, and James W Lillard. A Study of Effects of MultiCollinearity in the Multivariable Analysis. *Int J Appl Sci Technol*, 4(5):9–19, 2014.

[102] Keegan Korthauer, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J. Alm, and Stephanie C. Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):1–21, jun 2019.

[103] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, apr 2015.

[104] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(Article3), feb 2004.

[105] Cecilia L Ovkvist, Ian B Dodd, Kim Sneppen, and Jan O Haerter. DNA methylation in human epigenomes depends on local topology of CpG sites This model can reproduce the effects of CpG cluster-ing on methylation and produces stable and heritable alternative methylation states of CpG clusters, thus providing a coherent m. *Nucleic Acids Research*, 44(11):5123–5132, 2016.

[106] Ornella Affinito, Domenico Palumbo, Annalisa Fierro, Mariella Cuomo, Giulia De Riso, Antonella Monticelli, Gennaro Miele, Lorenzo Chiariotti, and Sergio Cocozza. Nucleotide distance

influences co-methylation between nearby CpG sites. *Genomics*, 112(1):144–150, jan 2020.

[107] Amanda J. Lea, Christopher M. Vockley, Rachel A. Johnston, Christina A. Del Carpio, Luis B. Barreiro, Timothy E. Reddy, and Jenny Tung. Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife*, 7:e37513, dec 2018.

[108] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. 116(32):15849–15854, 2019.

[109] Miles C. Benton, Heidi G. Sutherland, Donia Macartney-Coxson, Larisa M. Haupt, Rodney A. Lea, and Lyn R. Griffiths. Methylome-wide association study of whole blood DNA in the Norfolk Island isolate identifies robust loci associated with age. *Aging*, 9(3):753–768, 2017.

[110] Gad Abraham, Adam Kowalczyk, Justin Zobel, and Michael Inouye. SparSNP: Fast and memory-efficient analysis of all SNPs for phenotype prediction. *BMC Bioinformatics*, 13(1):1–8, may 2012.

[111] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[112] Jerome Friedman, Trevor Hastie, Rob Tibshirani, and Balasubramanian Narasimhan. Package 'glmnet': Lasso and elastic-net regularized generalized linear models, 2022. R package version 4.1-6.

[113] Fadil Santosa and William W. Symes. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, jul 1986.

[114] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[115] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[116] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, 2005.

[117] Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist*, 37(4):1733–1751, aug 2009.

[118] Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to d.c. programming: Theory, Algorithm and Applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 2018.

[119] Yongdai Kim, Hosik Choi, and Hee Seok Oh. Smoothly Clipped Absolute Deviation on High Dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, dec 2007.

[120] Patrick Breheny and Jian Huang. Regularization Paths for SCAD and MCP Penalized Regression Models [R package ncvreg version 3.13.0]. *Annals of Applied Statistics*, 5(1):232–253, mar 2021.

[121] Cun Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, apr 2010.

[122] Yi Yang, Yuwen Gu, and Hui Zou. gcdnet: The (Adaptive) LASSO and Elastic Net Penalized Least Squares, Logistic Regression, Hybrid Huberized Support Vector Machines, Squared Hinge Loss Support Vector Machines and Expectile Regression using a Fast Generalized Coordinate Descent Algorithm. `https://cran.rstudio.com/web/packages/gcdnet/index.html`, 2022. [Online; accessed 04-January-2023].

[123] Xiang Li, Donglin Zeng, and Yuanjia Wang. Coxnet: Regularized Cox Model. `https://github.com/cran/Coxnet/`, 2015. [Online; accessed 04-January-2023].

[124] Nan Xiao and Qing Song Xu. Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection.

Journal of Statistical Computation and Simulation, 85(18):3755–3765, dec 2015.

[125] Nan Xiao and Qing Song Xu. Multi-Step Adaptive Estimation Methods for Sparse Regressions [R package msaenet version 3.1]. *Journal of Statistical Computation and Simulation*, 85(18):3755–3765, may 2019.

[126] Nicholas G. Polson, James G. Scott, Bertrand Clarke, and C. Severinski. *Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction*, volume 9780199694. Oxford University Press, oct 2011.

[127] A. Huang, S. Xu, and X. Cai. Empirical Bayesian elastic net for multiple quantitative trait locus mapping. *Heredity*, 114(1):107–115, jan 2015.

[128] Venkatesh Mallikarjun, Venkatesh Mallikarjun, Stephen M. Richardson, Joe Swift, and Joe Swift. BayesENproteomics: Bayesian Elastic Nets for Quantification of Peptidoforms in Complex Samples. *Journal of proteome research*, 19(6):2167–2184, jun 2020.

[129] Jiahan Li, Kiranmoy Das, Guifang Fu, Runze Li, and Rongling Wu. The Bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516, feb 2011.

[130] Enes Makalic and Daniel F. Schmidt. High-dimensional bayesian regularised regression with the bayesreg package, 2016. arXiv.

[131] Kyu Ha Lee, Sounak Chakraborty, and Jianguo Sun. Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data. *The International Journal of Biostatistics*, 7(1):1–32, 2011.

[132] D. F. Andrews and C. L. Mallows. Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, sep 1974.

[133] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya-Gamma la-

tent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

[134] VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. *Annual review of public health*, 37:17–32, mar 2016.

[135] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, 59(5):1–38, 2014.

[136] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[137] Kosuke Imai, Booil Jo, and Elizabeth A Stuart. Commentary: Using Potential Outcomes to Understand Causal Mediation Analysis. *Multivariate Behavioral Research*, 46:842–854, 2011.

[138] Jeffrey M. Albert and Wei Wang. Sensitivity analyses for parametric causal mediation effect estimation. *Biostatistics*, 16(2):339–351, apr 2015.

[139] Theis Lange and Jørgen V. Hansen. Direct and Indirect Effects in a Survival Context. *Epidemiology*, 22(4):575–581, jul 2011.

[140] Kosuke Imai, Luke Keele, and Dustin Tingley. A General Approach to Causal Mediation Analysis. *Psychol Methods*, 15(4):309–34, 2010.

[141] Theis Lange, Kim Wadt Hansen, Rikke Sørensen, and Søren Galatius. Applied mediation analyses: a review and tutorial. *Epidemiology and Health*, 39:e2017035, 2017.

[142] Rhian Daniel, Jingjing Zhang, and Daniel Farewell. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal. Biometrische Zeitschrift*, 63(3):528, mar 2021.

[143] Arvid Sjolander, Elisabeth Dahlqwist, and Johan Zetterqvist. A Note on the Noncollapsibility of Rate Differences and Rate Ratios. *Epidemiology*, 27(3):356–359, 2016.

[144] Torben Martinussen and Stijn Vansteelandt. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime data analysis*, 19(3):279–296, jul 2013.

[145] Tyler J Vanderweele. Causal mediation analysis with survival data. *Epidemiology*, 22(4):582–585, 2011.

[146] Torben Martinussen and Thomas H Scheike. *Dynamic Regression Models for Survival Data*. Springer New York, 2006.

[147] D. Y. Lin and Zhiliang Ying. Semiparametric Analysis of the Additive Risk Model. *Biometrika*, 81(1):61, mar 1994.

[148] Diego Franco Saldana and Yang Feng. SIS: An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models. *Journal of Statistical Software*, 83:1–25, feb 2018.

[149] Yichao Wu. Ultrahigh Dimensional Feature Selection: Beyond The Linear Model Jianqing Fan Richard Samworth. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

[150] Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality 1. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[151] Tao Wang, Pei Rong Xu, and Li Xing Zhu. Non-convex penalized estimation in high-dimensional models with single-index structure. *Journal of Multivariate Analysis*, 109:221–235, aug 2012.

[152] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[153] A. Ralph Henderson. The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clinica Chimica Acta*, 359(1-2):1–26, sep 2005.

[154] Samantha V. Abram, Nathaniel E. Helwig, Craig A. Moodie, Colin G. DeYoung, Angus W. MacDonald, and Niels G. Waller.

Bootstrap Enhanced Penalized Regression for Variable Selection with Neuroimaging Data. *Frontiers in neuroscience*, 10:344, 2016.

[155] Charles Laurin, Dorret Boomsma, and Gitta Lubke. The use of vector bootstrapping to improve variable selection precision in Lasso models. *Statistical applications in genetics and molecular biology*, 15(4):305, aug 2016.

[156] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[157] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, nov 2003.

[158] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091, apr 2009.

[159] Gabriela Bindea, Jé Rô Me Galon, and Bernhard Mlecnik. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics Applications*, 29(5):661–663, 2013.

[160] Michael P. Czech. Insulin action and resistance in obesity and type 2 diabetes. *Nature medicine*, 23(7):804, jul 2017.

[161] Béatrice Lauby-Secretan, Chiara Scoccianti, Dana Loomis, Yann Grosse, Franca Bianchini, and Kurt Straif. Body Fatness and Cancer — Viewpoint of the IARC Working Group. *New England Journal of Medicine*, 375(8):794–798, aug 2016.

[162] Paul D. Loprinzi, Carlos J. Crespo, Ross E. Andersen, and Ellen Smit. Association of body mass index with cardiovascular

disease biomarkers. *American Journal of Preventive Medicine*, 48(3):338–344, mar 2015.

[163] W. David Ashton, K. Nanchahal, and D. A. Wood. Body mass index and metabolic risk factors for coronary heart disease in women. *European Heart Journal*, 22(1):46–55, jan 2001.

[164] Stefania Lamon-Fava, Peter WF Wilson, Ernst J Schaefer, and Lipid Metab. Impact of Body Mass Index on Coronary Heart Disease Risk Factors in Men and Women The Framingham Offspring Study. *Arterioscler Thromb Vasc Biol*, 16(12):1509–15, 1996.

[165] Jürgen Scheer, Silvia Findenig, Walter Goessler, Kevin A. Francesconi, Barbara Howard, Jason G. Umans, Jonathan Pollak, Maria Tellez-Plaza, Ellen K. Silbergeld, Eliseo Guallar, and Ana Navas-Acien. Arsenic species and selected metals in human urine: validation of HPLC/ICPMS and ICPMS procedures for a long-term population-based epidemiological study. *Analytical methods : advancing methods and applications*, 4(2):406–413, feb 2012.

[166] Ana Navas-Acien, Jason G. Umans, Barbara V. Howard, Walter Goessler, Kevin A. Francesconi, Ciprian M. Crainiceanu, Ellen K. Silbergeld, and Eliseo Guallar. Urine arsenic concentrations and species excretion patterns in American Indian communities over a 10-year period: the Strong Heart Study. *Environmental health perspectives*, 117(9):1428–1433, 2009.

[167] Anne E. Nigra, Tiffany R. Sanchez, Keeve E. Nachman, David E. Harvey, Steven N. Chillrud, Joseph H. Graziano, and Ana Navas-Acien. The effect of the Environmental Protection Agency maximum contaminant level on arsenic exposure in the USA from 2003 to 2014: an analysis of the National Health and Nutrition Examination Survey (NHANES). *The Lancet Public Health*, 2(11):e513–e521, nov 2017.

[168] Anne E. Nigra, Qixuan Chen, Steven N. Chillrud, Lili Wang, David Harvey, Brian Mailloux, Pam Factor-Litvak, and Ana

Navas-Acien. Inequalities in Public Water Arsenic Concentrations in Counties and Community Water Systems across the United States, 2006-2011. *Environmental health perspectives*, 128(12):1–13, 2020.

[169] Kenneth Westerman, Paola Sebastiani, Paul Jacques, Simin Liu, Dawn Demeo, and José M. Ordovás. DNA methylation modules associate with incident cardiovascular disease and cumulative risk factor exposure. *Clinical Epigenetics*, 11(1):142, 2019.

[170] Esther García-Esquinas, Marina Pollan, Maria Tellez-Plaza, Kevin A. Francesconi, Walter Goessler, Eliseo Guallar, Jason G. Umans, Jeunliang Yeh, Lyle G. Best, and Ana Navas-Acien. Cadmium Exposure and Cancer Mortality in a Prospective Cohort: The Strong Heart Study. *Environmental Health Perspectives*, 122(4):363, 2014.

[171] Yen Tsung Huanga and Hwai I. Yangc. Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(3):370, may 2017.

[172] Michael Höfler, Hildegard Pfister, Roselind Lieb, and Hans Ulrich Wittchen. The use of weights to account for non-response and drop-out. *Social psychiatry and psychiatric epidemiology*, 40(4):291–299, apr 2005.

[173] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Lars J Jensen, and Christian Von Mering. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):607–613, 2018.

[174] Belinda Phipson, Jovana Maksimovic, and Alicia Oshlack. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*, 32(2):286–288, jan 2016.

[175] Godfrey S Getz and Catherine A Reardon. Diet and Murine Atherosclerosis. *Arterioscler Thromb Vasc Biol*, 26(2):242–9, 2006.

[176] Amanda M. Fretts, Barbara V. Howard, Barbara McKnight, Glen E. Duncan, Shirley A.A. Beresford, Mihriye Mete, Sigal Eilat-Adar, Ying Zhang, and David S. Siscovick. Associations of processed meat and unprocessed red meat intake with incident diabetes: the Strong Heart Family Study. *The American journal of clinical nutrition*, 95(3):752–758, mar 2012.

[177] Amanda M. Fretts, Barbara V. Howard, David S. Siscovick, Lyle G. Best, Shirley A.A. Beresford, Mihriye Mete, Sigal Eilat-Adar, Nona Sotoodehnia, and Jinying Zhao. Processed Meat, but Not Unprocessed Red Meat, Is Inversely Associated with Leukocyte Telomere Length in the Strong Heart Family Study. *The Journal of nutrition*, 146(10):2013–2018, 2016.

[178] Angelika Merkel, Marcos Ferná Ndez-Callejo, Eloi Casals, Santiago Marco-Sola, Ronald Schuyler, Ivo G Gut, and Simon C Heath. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 35(5):737–742, 2019.

[179] Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):1–10, oct 2012.

[180] Qianwen Wang, Ming Li, Tianzhi Wu, Li Zhan, Lin Li, Meijun Chen, Wenqin Xie, Zijing Xie, Erqiang Hu, Shuangbin Xu, and Guangchuang Yu. Exploring Epigenomic Datasets by ChIPseeker. *Current Protocols*, 2(10):e585, oct 2022.

[181] Raymond G. Cavalcante and Maureen A. Sartor. annotatr: genomic regions in context. *Bioinformatics*, 33(15):2381, aug 2017.

[182] Konstansa Lazarević, Dejan Nikolić, Ljiljana Stosić, Suzana Milutinović, Jelena Videnović, and Dragan Bogdanović. Determination of lead and arsenic in tobacco and cigarettes: an impor-

tant issue of public health. *Central European journal of public health*, 20(1):62–66, 2012.

[183] J. H. Huang, K. N. Hu, J. Ilgen, and G. Ilgen. Occurrence and stability of inorganic and organic arsenic species in wines, rice wines and beers from Central European market. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess*, 29(1):85–93, 2011.

[184] Guillaume Jondeau, Jacques Ropers, Ellen Regalado, Alan Braverman, Arturo Evangelista, Guisela Teixedo, Julie De Backer, Laura Muiño-Mosquera, Sophie Naudion, Cecile Zordan, Takayuki Morisaki, Hiroto Morisaki, Yskert Von Kodolitsch, Sophie Dupuis-Girod, Shaine A. Morris, Richmond Jeremy, Sylvie Odent, Leslie C. Adès, Madhura Bakshi, Katherine Holman, Scott Lemaire, Olivier Milleron, Maud Langeois, Myrtille Spentchian, Melodie Aubart, Catherine Boileau, Reed Pyeritz, and Dianna M. Milewicz. International Registry of Patients Carrying TGFBR1 or TGFBR2 Mutations: Results of the Montalcino Aortic Consortium. *Circulation. Cardiovascular genetics*, 9(6):548, dec 2016.

[185] Rosina De Cario, Elena Sticchi, Laura Lucarini, Monica Attanasio, Stefano Nistri, Rossella Marcucci, Guglielmina Pepe, and Betti Giusti. Role of TGFBR1 and TGFBR2 genetic variants in Marfan syndrome. *Journal of vascular surgery*, 68(1):225–233.e5, jul 2018.

[186] Bart L. Loeys, Junji Chen, Enid R. Neptune, Daniel P. Judge, Megan Podowski, Tammy Holm, Jennifer Meyers, Carmen C. Leitch, Nicholas Katsanis, Neda Sharifi, F. Lauren Xu, Loretha A. Myers, Philip J. Spevak, Duke E. Cameron, Julie De Backer, Jan Hellemans, Yan Chen, Elaine C. Davis, Catherine L. Webb, Wolfram Kress, Paul Coucke, Daniel B. Rifkin, Anne M. De Paepe, and Harry C. Dietz. A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2. *Nature Genetics*, 37(3):275–281, jan 2005.

[187] Yvan Devaux, Melanie Bousquenaud, Sophie Rodius, Pierre Yves Marie, Fatiha Maskali, Lu Zhang, Francisco Azuaje, and Daniel R. Wagner. Transforming growth factor β receptor 1 is a new candidate prognostic biomarker after acute myocardial infarction. *BMC Medical Genomics*, 4(1):1–13, dec 2011.

[188] Wei Li and Hong Yue. Thymidine phosphorylase: a potential new target for treating cardiovascular disease. *Trends in cardiovascular medicine*, 28(3):157, apr 2018.

[189] Faiza Altaf, Cornelia Vesely, Abdul Malik Sheikh, Rubab Munir, Syed Tahir Abbas Shah, and Aamira Tariq. Modulation of ADAR mRNA expression in patients with congenital heart defects. *PLoS ONE*, 14(4):e0200968, apr 2019.

[190] Luis Fernando Negro Silva, Kiran Makhani, Maryse Lemaire, Catherine A. Lemarié, Dany Plourde, Alicia M. Bolt, Christopher Chiavatti, D. Scott Bohle, Stéphanie Lehoux, Mark S. Goldberg, and Koren K. Mann. Sex-Specific Effects of Prenatal and Early Life Inorganic and Methylated Arsenic Exposure on Atherosclerotic Plaque Development and Composition in Adult ApoE-/- Mice. *Environmental Health Perspectives*, 129(5):57008–57008, may 2021.

[191] Yan Yuan, Guillermo Marshall, Catterina Ferreccio, Craig Steinmaus, Steve Selvin, Jane Liaw, Michael N Bates, and Allan H Smith. Acute Myocardial Infarction Mortality in Comparison with Lung and Bladder Cancer Mortality in Arsenic-exposed Region II of Chile from 1950 to 2000. *Am J Epidemiol*, 166(12):1381–91, 2007.

[192] Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Flexible Mediation Analysis With Multiple Mediators. *American Journal of Epidemiology*, 186(2):184–193, jul 2017.

[193] Tyler Van Der weele and Stijn Vansteelandt. Mediation Analysis with Multiple Mediators. *Epidemiologic methods*, 2(1):95, dec 2014.

[194] Kosuke Imai and Teppei Yamamoto. Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments. *Political Analysis*, 21:141–171, 2013.

[195] R. M. Daniel, B. L. De Stavola, S. N. Cousens, and S. Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, mar 2015.

[196] Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42(5):1511–1519, oct 2013.

[197] David P. MacKinnon, Chondra M. Lockwood, Jeanne M. Hoffman, Stephen G. West, and Virgil Sheets. A Comparison of Methods to Test Mediation and Other Intervening Variable Effects. *Psychological methods*, 7(1):83, 2002.

[198] Linda Valeri and Tyler J. VanderWeele. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological methods*, 18(2):137, jun 2013.

[199] Odd Aalen. A Model for Nonparametric Regression Analysis of Counting Processes. *Lecture Notes In Statistics*, 2:1–25, 1980.

[200] Luc Devroye. *Non-Uniform Random Variate Generation.* Springer New York, 1 edition, 1986.

[201] Mohammad Mohammadianpanah Farzan Madadizadeh, Amin Ghanbarnejad, Vahid Ghavami, Mohammad Zare Bandamiri. Applying Additive Hazards Models for Analyzing Survival in Patients with Colorectal Cancer in Fars Province, Southern Iran. *Asian Pac J Cancer Prev*, 18(4):1077–1083, 2011.

[202] Michael Klutstein, Deborah Nejman, Razi Greenfield, and Howard Cedar. DNA Methylation in Cancer and Aging. *Cancer Res*, 76(12):3446–50, 2016.

[203] Zdenko Herceg and Srikant Ambatipudi. Smoking-associated DNA methylation changes: no smoke without fire. *Epigenomics*, 11(10):1117–1119, 2019.

[204] Francesca Fasanelli, Laura Baglietto, Erica Ponzi, Florence Guida, Gianluca Campanella, Mattias Johansson, Kjell Grankvist, Mikael Johansson, Manuela Bianca Assumma, Alessio Naccarati, Marc Chadeau-Hyam, Ugo Ala, Christian Faltus, Rudolf Kaaks, Angela Risch, Bianca De Stavola, Allison Hodge, Graham G Giles, Melissa C Southey, Caroline L Relton, Philip C Haycock, Eiliv Lund, Silvia Polidoro, Torkjel M Sandanger, Gianluca Severi, and Paolo Vineis. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nature Communications*, 6:10192, 2015.

[205] Michelle Vaz, Stephen Y. Hwang, Ioannis Kagiampakis, Jillian Phallen, Ashwini Patil, Heather M. O'Hagan, Lauren Murphy, Cynthia A. Zahnow, Edward Gabrielson, Victor E. Velculescu, Hariharan P. Easwaran, and Stephen B. Baylin. Chronic Cigarette Smoke-Induced Epigenomic Changes Precede Sensitization of Bronchial Epithelial Cells to Single Step Transformation by KRAS Mutations. *Cancer cell*, 32(3):360, sep 2017.

[206] Thomas Battram, Rebecca C. Richmond, Laura Baglietto, Philip C. Haycock, Vittorio Perduca, Stig E. Bojesen, Tom R. Gaunt, Gibran Hemani, Florence Guida, Robert Carreras-Torres, Rayjean Hung, Christopher I. Amos, Joshua R. Freeman, Torkjel M. Sandanger, Therese H. Nøst, Børge G. Nordestgaard, Andrew E. Teschendorff, Silvia Polidoro, Paolo Vineis, Gianluca Severi, Allison M. Hodge, Graham G. Giles, Kjell Grankvist, Mikael B. Johansson, Mattias Johansson, George Davey Smith, and Caroline L. Relton. Appraising the causal relevance of DNA methylation for risk of lung cancer. *International Journal of Epidemiology*, 48(5):1493, oct 2019.

[207] Yi Qian Sun, Rebecca C. Richmond, Matthew Suderman, Josine L. Min, Thomas Battram, Arnar Flatberg, Vidar Beis-

vag, Therese Haugdahl Nøst, Florence Guida, Lin Jiang, Sissel Gyrid Freim Wahl, Arnulf Langhammer, Frank Skorpen, Rosie M. Walker, Andrew D. Bretherick, Yanni Zeng, Yue Chen, Mattias Johansson, Torkjel M. Sandanger, Caroline L. Relton, and Xiao Mei Mai. Assessing the role of genome-wide DNA methylation between smoking and risk of lung cancer using repeated measurements: the HUNT study. *International journal of epidemiology*, 50(5):1482–1497, oct 2021.

[208] Centers for Disease Control and Prevention. Smoking and Cancer. `https://www.cdc.gov/tobacco/campaign/tips/diseases/cancer.html#:~:text=Quitting%20smoking%20lowers%20the%20risk,acute%20myeloid%20leukemia%20(AML).`, 2022. [Online; accessed 15-January-2023].

[209] Yunhua Fan, Jian-Min Yuan, Renwei Wang, Yu-Tang Gao, and Mimi C Yu. Alcohol, Tobacco and Diet in Relation to Esophageal Cancer: The Shanghai Cohort Study. *Nutr Cancer*, 60(3):354–63, 2008.

[210] National Institutes of Health. Smoking and Bladder Cancer. `https://www.nih.gov/news-events/nih-research-matters/smoking-bladder-cancer`, 2011. [Online; accessed 21-January-2023].

[211] Neal D. Freedman, Debra T. Silverman, Albert R. Hollenbeck, Arthur Schatzkin, and Christian C. Abnet. Association between smoking and risk of bladder cancer among men and women. *JAMA : the journal of the American Medical Association*, 306(7):737, aug 2011.

[212] Cosimo De Nunzio, Gerald L. Andriole, Ian M. Thompson, and Stephen J. Freedland. Smoking and Prostate Cancer: A Systematic Review. *European Urology Focus*, 1(1):28–38, aug 2015.

[213] Michael Huncharek, K. Sue Haddock, Rodney Reid, and Bruce Kupelnick. Smoking as a Risk Factor for Prostate Cancer: A Meta-Analysis of 24 Prospective Cohort Studies. *American Journal of Public Health*, 100(4):693, apr 2010.

[214] Ruth Pidsley, Chloe C. Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C. Schalkwyk. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 14(1):1–10, may 2013.

[215] Broad Institute. Analysis pipelines for the GTEx Consortium and TOPMed. `https://github.com/broadinstitute/gtex-pipeline`, 2022. [Online; accessed 24-January-2023].

[216] Jiantao Ma, Roby Joehanes, Chunyu Liu, Amena Keshawarz, Shih-Jen Hwang, Helena Bui, Brandon Tejada, Meera Sooda, Peter J Munson, Cumhur Y Demirkale, Paul Courchesne, Nancy L Heard-Costa, Achilleas N Pitsillides, Mike Feolo, Nataliya Sharopova, Ramachandran S Vasan, Tianxiao Huan, and Daniel Levy. Elucidating the genetic architecture of DNA methylation to identify promising molecular mechanisms of disease. *Scientific reports*, 12(1):19564, nov 2022.

[217] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, nov 2010.

[218] Roby Joehanes, Saixia Ying, Tianxiao Huan, Andrew D. Johnson, Nalini Raghavachari, Richard Wang, Poching Liu, Kimberly A. Woodhouse, Shurjo K. Sen, Kahraman Tanriverdi, Paul Courchesne, Jane E. Freedman, Christopher J. O'Donnell, Daniel Levy, and Peter J. Munson. Gene Expression Signatures of Coronary Heart Disease. *Arteriosclerosis, thrombosis, and vascular biology*, 33(6):1418, jun 2013.

[219] Dorothy A. Rhoades, John Farley, Stephen M. Schwartz, Kimberly M. Malloy, Wenyu Wang, Lyle G. Best, Ying Zhang, Tauqeer Ali, Fawn Yeh, Everett R. Rhoades, Elisa Lee, and Barbara V. Howard. Cancer mortality in a population-based cohort of American Indians - The strong heart study. *Cancer epidemiology*, 74:101978, oct 2021.

[220] Arce Domingo-Relloso, Angela L. Riffo-Campos, Karin Haack, Pilar Rentero-Garrido, Christine Ladd-Acosta, Daniele M. Fallin, Wan Yee Tang, Miguel Herreros-Martinez, Juan R. Gonzalez, Anne K. Bozack, Shelley A. Cole, Ana Navas-Acien, and Maria Tellez-Plaza. Cadmium, Smoking, and Human Blood DNA Methylation Profiles in Adults from the Strong Heart Study. *Environmental health perspectives*, 128(6):067005, 2020.

[221] Roby Joehanes, Xiaoling Zhang, Tianxiao Huan, Chen Yao, Sai xia Ying, Quang Tri Nguyen, Cumhur Yusuf Demirkale, Michael L. Feolo, Nataliya R. Sharopova, Anne Sturcke, Alejandro A. Schäffer, Nancy Heard-Costa, Han Chen, Po ching Liu, Richard Wang, Kimberly A. Woodhouse, Kahraman Tanriverdi, Jane E. Freedman, Nalini Raghavachari, Josée Dupuis, Andrew D. Johnson, Christopher J. O'Donnell, Daniel Levy, and Peter J. Munson. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome biology*, 18(1):16, jan 2017.

[222] Chengwen Luo, Botao Fa, Yuting Yan, Yang WangID, Yiwang ZhouID, Yue ZhangID, and Zhangsheng Yu. High-dimensional mediation analysis in survival models. *PLoS Comput Biol*, 16(4):e1007768, 2020.

[223] Gea Kõks, Mari Liis Uudelepp, Maia Limbach, Pärt Peterson, Ene Reimann, and Sulev Kõks. Smoking-induced expression of the GPR15 gene indicates its potential role in chronic inflammatory pathologies. *American Journal of Pathology*, 185(11):2898–2906, nov 2015.

[224] Gordon W. McLean, Neil O. Carragher, Egle Avizienyte, Jeff Evans, Valerie G. Brunton, and Margaret C. Frame. The role of focal-adhesion kinase in cancer — a new therapeutic opportunity. *Nature Reviews Cancer*, 5(7):505–515, jul 2005.

[225] Myungmi Lee and Inmoo Rhee. Cytokine Signaling in Tumor Progression. *Immune Network*, 17(4):214, aug 2017.

[226] Yi Xiao and Jixin Dong. The Hippo Signaling Pathway in Cancer: A Cell Cycle Perspective. *Cancers*, 13(24):6214, dec 2021.

[227] Tyler J Vanderweele and Onyebuchi A Arah. Unmeasured Confounding for General Outcomes, Treatments, and Confounders: Bias Formulas for Sensitivity Analysis. *Epidemiology*, 22(1):42–52, 2011.

[228] Irene Hernando-Herraez, Brendan Evano, Thomas Stubbs, Pierre-Henri Commere, Marc Jan Bonder, Stephen Clark, Simon Andrews, Shahragim Tajbakhsh, and Wolf Reik. Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nat Commun*, 10(1):4361, 2019.

[229] Shikha Gupta, Oliver M. Dovey, Ana Filipa Domingues, Oliwia W. Cyran, Caitlin M. Cash, George Giotopoulos, Justyna Rak, Jonathan Cooper, Malgorzata Gozdecka, Liza Dijkhuis, Ryan J. Asby, Noor Al-Jabery, Victor Hernandez-Hernandez, Sudhakaran Prabakaran, Brian J. Huntly, George S. Vassiliou, and Cristina Pina. Transcriptional variability accelerates preleukemia by cell diversification and perturbation of protein synthesis. *Science Advances*, 8(31):eabn4886, aug 2022.

[230] Ruochen Jiang, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biology*, 23(1):1–24, jan 2022.

[231] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9(1):1–9, mar 2018.

[232] Kip D Zimmerman, Mark A Espeland, and Carl D Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nat Commun*, 12(1):738, 2021.

[233] Mariia Bilous, Loc Tran, Chiara Cianciaruso, Aurélie Gabriel, Hugo Michel, Santiago J. Carmona, Mikael J. Pittet, and David Gfeller. Metacells untangle large and complex single-cell transcriptome networks. *BMC Bioinformatics*, 23(1):1–24, dec 2022.

[234] Belinda Phipson and Alicia Oshlack. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome biology*, 15(9):465, sep 2014.

[235] Simone Tiberi, Helena L. Crowell, Pantelis Samartsidis, Lukas M. Weber, and Mark D. Robinson. distinct: a novel approach to differential distribution analyses, 2022. bioRxiv.

[236] Keegan D. Korthauer, Li Fang Chu, Michael A. Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):1–15, oct 2016.

[237] George C. Linderman, Jun Zhao, Manolis Roulis, Piotr Bielecki, Richard A. Flavell, Boaz Nadler, and Yuval Kluger. Zero-preserving imputation of single-cell RNA-seq data. *Nature Communications*, 13(1):1–11, jan 2022.

[238] Helena L Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D Robinson. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat Commun*, 11:6077, 2020.

[239] Bhupinder Pal, Yunshun Chen, François Vaillant, Bianca D Capaldo, Rachel Joyce, Xiaoyu Song, Vanessa L Bryant, Jocelyn S Penington, Leon Di Stefano, Nina Tubau Ribera, Stephen Wilcox, Gregory B Mann, KConFab, Anthony T Papenfuss, Geoffrey J Lindeman, Gordon K Smyth, and Jane E Visvader. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *The EMBO Journal*, 40(11):e107333, jun 2021.

[240] Ayshwarya Subramanian, Mikhail Alperovich, Yiming Yang, and Bo Li. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biology*, 23(1):267, dec 2022.

[241] Mengjie Chen and Xiang Zhou. VIPER: Variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biology*, 19(1):1–15, nov 2018.

[242] Tyler J Vanderweele and Yasutaka Chiba. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiol Biostat Public Health*, 11(2):e9027, 2014.

[243] Eric J. Tchetgen Tchetgen. On causal mediation analysis with a survival outcome. *The international journal of biostatistics*, 7(1):Article 33, 2011.

[244] Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology*, 23(1):1–20, dec 2022.

[245] Ariane Mora, Christina Schmidt, Brad Balderson, Christian Frezza, and Mikael Boden. Sircle (signature regulatory clustering) model integration reveals mechanisms of phenotype regulation in renal cancer, 2022. bioRxiv.

[246] Liang Jin, Yingtao Bi, Chenqi Hu, Jun Qu, Shichen Shen, Xue Wang, and Yu Tian. A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, 11(1):1–11, jan 2021.

Appendix A: Supplementary tables and figures for section 3.2.1

Table A1: Mean differences (95 % CI) for the CpGs selected by ISIS - Aenet for BMI and comparison with linear regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	LS	Bayes
cg00047657	10	70639626	*STOX1*	0.26 (0.02, 0.45)	1.22 (0.19, 2.26)	0.92 (0.02, 1.88)
cg00602326	5	31427700	*DROSHA*	-0.71 (-0.93, -0.46)	-1.42 (-2.14, -0.7)	-1.31 (-2.01, -0.59)
cg00831028	20	55043745	*RTF2*	0.39 (0.11, 0.67)	1.2 (0.56, 1.85)	0.94 (0.31, 1.58)
cg01577114	14	61906283	*PRKCH*	0.48 (0.22, 0.7)	0.83 (0.25, 1.42)	0.71 (0.14, 1.27)
cg01880404	19	44079527	*XRCC1*	0.52 (0.19, 0.81)	1.07 (0.51, 1.63)	0.93 (0.38, 1.48)
cg01894508	2	70189111	*ASPRV1*	0.49 (0.22, 0.68)	1.42 (0.37, 2.46)	1.23 (0.23, 2.24)
cg03008286	20	60394429	*CDH4*	0.3 (0.05, 0.54)	1.28 (0.44, 2.11)	0.96 (0.16, 1.79)
cg03078551	17	41656298	*ETV4*	-0.47 (-0.69, -0.24)	-0.76 (-1.4, -0.12)	-0.69 (-1.31, -0.08)
cg03580256	1	78149279	*ZZZ3*	0.5 (0.25, 0.72)	1.67 (0.9, 2.44)	1.4 (0.65, 2.16)
cg03710333	4	1722958	*TMEM129*	0.51 (0.2, 0.82)	0.84 (0.33, 1.34)	0.81 (0.32, 1.31)
cg05253110	7	141130687	*TMEM178B*	0.39 (0.08, 0.65)	0.98 (0.39, 1.58)	0.85 (0.27, 1.44)
cg05575921	5	373378	*AHRR*	0.51 (0.31, 0.7)	2.54 (1.17, 3.91)	1.69 (0.5, 2.87)
cg08389486	9	132377983	*C9orf50*	0.3 (0.05, 0.56)	0.86 (0.21, 1.51)	0.63 (0.03, 1.26)
cg09074260	11	94707049	*KDM4D*	0.26 (0.03, 0.5)	0.72 (0.21, 1.23)	0.61 (0.11, 1.11)
cg09364595	9	139457749	*MIR4674*	-0.5 (-0.72, -0.26)	-0.9 (-1.59, -0.21)	-0.82 (-1.5, -0.16)
cg10092685	16	73090591	*ZFHX3*	-0.32 (-0.55, -0.08)	-1.3 (-2.05, -0.55)	-1.08 (-1.81, -0.35)
cg10251538	3	108800886	*MORC1*	-0.62 (-0.87, -0.36)	-1.16 (-1.86, -0.46)	-1.09 (-1.78, -0.4)
cg10894085	14	91817232	*CCDC88C*	0.39 (0.14, 0.61)	0.96 (0.19, 1.73)	0.85 (0.12, 1.6)
cg10948061	6	110500990	*WASF1*	0.43 (0.15, 0.7)	0.67 (0.1, 1.24)	0.65 (0.1, 1.21)
cg11202345	17	76976057	*LGALS3BP*	0.75 (0.47, 1.02)	1.85 (1.2, 2.51)	1.68 (1.04, 2.32)
cg11591807	2	118888717	*INSIG2*	0.44 (0.19, 0.7)	0.91 (0.03, 1.79)	0.86 (0.05, 1.72)
cg11625476	17	4795410	*MINK1*	0.38 (0.12, 0.66)	0.82 (0.17, 1.48)	0.71 (0.09, 1.36)

cg11743438	6	16238437	*GMPR*	0.36 (0.11, 0.59)	1.01 (0.46, 1.55)	0.84 (0.31, 1.38)
cg13549904	1	154438143	*IL6R*	-0.65 (-0.97, -0.32)	-1.54 (-2.04, -1.04)	-1.37 (-1.86, -0.88)
cg15340629	22	27725596	*MN1*	0.28 (0.05, 0.49)	0.97 (0.34, 1.61)	0.8 (0.19, 1.43)
cg15706574	6	46231809	*RCAN2*	0.32 (0.06, 0.56)	1.27 (0.58, 1.95)	1.02 (0.35, 1.68)
cg15826542	19	539241	*CDC34*	0.48 (0.23, 0.72)	0.84 (0.27, 1.41)	0.78 (0.22, 1.34)
cg16032415	8	95278692	*GEM*	-0.29 (-0.52, -0.02)	-1.72 (-2.48, -0.97)	-1.36 (-2.09, -0.62)
cg16406078	20	825634	*FAM110A*	-0.62 (-0.87, -0.35)	-1.58 (-2.21, -0.94)	-1.41 (-2.04, -0.79)
cg16640008	6	159515404	*TAGAP*	0.42 (0.13, 0.63)	1.06 (0.28, 1.85)	0.9 (0.16, 1.66)
cg16758086	1	6173356	*CHD5*	-0.58 (-0.86, -0.28)	-1.56 (-2.34, -0.77)	-1.3 (-2.07, -0.53)
cg17420142	18	32702783	*MAPRE2*	-0.52 (-0.79, -0.26)	-1.1 (-1.9, -0.31)	-0.91 (-1.69, -0.16)
cg18011760	2	19320928	*MIR4757*	0.41 (0.18, 0.64)	0.97 (0.45, 1.49)	0.8 (0.28, 1.31)
cg18322280	14	57793087	*AP5M1*	-0.66 (-0.95, -0.34)	-1.28 (-1.93, -0.63)	-1.15 (-1.79, -0.52)
cg18391209	1	223747670	*CAPN8*	0.28 (0.03, 0.54)	0.76 (0.25, 1.27)	0.63 (0.13, 1.14)
cg18499545	8	110552416	*EBAG9*	0.35 (0.03, 0.71)	0.89 (0.38, 1.4)	0.74 (0.24, 1.25)
cg18613281	1	39596444	*MACF1*	-0.33 (-0.53, -0.08)	-1.85 (-2.61, -1.09)	-1.47 (-2.21, -0.71)
cg19026621	1	249106516	*SH3BP5L*	0.34 (0.09, 0.56)	0.9 (0.37, 1.44)	0.71 (0.18, 1.24)
cg19685672	17	33402829	*RFFL*	0.32 (0.08, 0.54)	1.02 (0.25, 1.79)	0.76 (0.05, 1.51)
cg20587236	12	109900956	*KCTD10*	0.64 (0.34, 0.92)	1.55 (0.92, 2.19)	1.44 (0.81, 2.07)
cg22648996	10	63946213	*RTKN2*	0.55 (0.33, 0.76)	2.17 (1.02, 3.31)	1.93 (0.81, 3.06)
cg23615467	1	25695799	*RHCE*	0.46 (0.16, 0.73)	0.82 (0.21, 1.43)	0.75 (0.16, 1.35)
cg25240153	16	23890018	*PRKCB*	0.46 (0.21, 0.66)	1.03 (0.11, 1.95)	0.88 (0.04, 1.78)
cg26416168	2	71934434	*DYSF*	0.42 (0.18, 0.67)	0.96 (0.19, 1.73)	0.87 (0.13, 1.61)
cg26467270	17	76718664	*CYTH1*	-1.09 (-1.34, -0.8)	-3.33 (-4.04, -2.62)	-3.01 (-3.7, -2.32)
cg26800893	11	67184596	*CARNS1*	-0.62 (-0.85, -0.38)	-1.21 (-1.85, -0.56)	-1.1 (-1.73, -0.46)
cg27080917	12	11978350	*ETV6*	-0.5 (-0.76, -0.24)	-1.49 (-2.26, -0.72)	-1.19 (-1.95, -0.44)
cg27254295	16	80574757	*DYNLRB2*	0.4 (0.11, 0.67)	1.12 (0.48, 1.76)	0.96 (0.33, 1.59)

Abbreviations: ISIS, Iterative Sure Independence Screening; Aenet, adaptive elastic-net; LS, least squares.

Table A2: Mean differences (95 % CI) for the CpGs selected by ISIS - MSAenet for BMI and comparison with linear regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	LS	Bayes
cg00011855	13	114199892	*TMCO3*	0.84 (0.07, 1.56)	0.95 (0.21, 1.68)	0.68 (0.01, 1.4)
cg01880404	19	44079527	*XRCC1*	1.11 (0.28, 1.94)	1.14 (0.47, 1.8)	0.94 (0.3, 1.59)
cg02997817	14	64929012	*ZBTB25*	1.07 (0.41, 1.77)	1.02 (0.23, 1.81)	0.85 (0.11, 1.62)
cg05575921	5	373378	*AHRR*	2.02 (0.94, 2.98)	2.18 (1.07, 3.28)	2.12 (1.05, 3.12)
cg06534023	21	34100462	*PAXBP1-AS1*	1.27 (0.61, 1.88)	1.31 (0.72, 1.91)	1.05 (0.47, 1.63)
cg07443900	4	1018378	*FGFRL1*	1.37 (0.66, 1.97)	1.47 (0.78, 2.15)	1.11 (0.44, 1.76)
cg14813947	19	47164221	*DACT3*	1.07 (0.39, 1.83)	1.17 (0.47, 1.86)	0.92 (0.24, 1.59)
cg15705813	2	70297499	*PCBP1-AS1*	1.16 (0.12, 2.02)	1.11 (0.12, 2.09)	1.05 (0.15, 1.98)
cg16209444	3	58522771	*ACOX2*	1.15 (0.4, 1.83)	1.11 (0.37, 1.86)	0.99 (0.26, 1.69)
cg16740586	21	43655919	*ABCG1*	1.51 (0.73, 2.29)	1.31 (0.53, 2.08)	1.36 (0.61, 2.11)
cg20437049	6	75918463	*COL12A1*	0.54 (0.27, 0.77)	0.54 (0.29, 0.79)	0.47 (0.22, 0.71)
cg24490227	11	133928292	*JAM3*	1.23 (0.34, 1.94)	1.33 (0.49, 2.18)	0.99 (0.18, 1.82)
cg27243685	21	43642366	*ABCG1*	1.58 (0.82, 2.37)	1.64 (0.84, 2.44)	1.39 (0.62, 2.17)

Abbreviations: ISIS, Iterative Sure Independence Screening, MSAenet, multi-step adaptive elastic-net; LS, least squares.

Table A3: Mean differences (95 % CI) for the CpGs selected by ISIS - enet for BMI and comparison with linear regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	LS	Bayes
cg00141611	5	172264364	*ERGIC1*	0.85 (0.24, 1.47)	1 (0.31, 1.68)	0.71 (0.07, 1.38)
cg00602326	5	31427700	*DROSHA*	-1.14 (-1.86, -0.51)	-1.11 (-1.89, -0.33)	-1.15 (-1.91, -0.39)
cg01577114	14	61906283	*PRKCH*	0.67 (0.11, 1.2)	0.66 (0.03, 1.29)	0.64 (0.06, 1.25)
cg01765545	5	81045418	*SSBP2*	1.29 (0.42, 2.11)	1.49 (0.57, 2.41)	1.17 (0.29, 2.05)
cg01894508	2	70189111	*ASPRV1*	1.24 (0.2, 2.04)	1.26 (0.09, 2.43)	1.19 (0.13, 2.28)
cg03078551	17	41656298	*ETV4*	-0.85 (-1.41, -0.24)	-0.91 (-1.57, -0.25)	-0.79 (-1.42, -0.15)
cg03710333	4	1722958	*TMEM129*	0.79 (0.25, 1.31)	0.8 (0.26, 1.34)	0.78 (0.26, 1.31)
cg03882777	6	33160965	*RXRB*	0.67 (0.06, 1.15)	0.8 (0.29, 1.3)	0.6 (0.11, 1.09)
cg06235693	19	52275119	*FPR2*	0.96 (0.3, 1.47)	1.1 (0.45, 1.75)	0.88 (0.26, 1.51)
cg07504977	10	102131012	*OLMALINC*	1.06 (0.39, 1.79)	1.2 (0.41, 2)	0.97 (0.21, 1.75)
cg08389486	9	132377983	*C9orf50*	1.04 (0.42, 1.56)	1.15 (0.46, 1.84)	0.95 (0.28, 1.63)
cg08633893	13	100068932	*MIR548AN*	1.56 (0.72, 2.28)	1.8 (0.86, 2.74)	1.47 (0.6, 2.36)
cg08851202	9	95999150	*WNK2*	1.32 (0.14, 2.28)	1.55 (0.26, 2.84)	1.18 (0.05, 2.4)
cg09364595	9	139457749	*MIR4674*	-0.95 (-1.58, -0.32)	-1.01 (-1.76, -0.27)	-0.86 (-1.58, -0.14)
cg09554443	1	167487762	*CD247*	-1.01 (-1.81, -0.24)	-1.14 (-2.03, -0.26)	-0.9 (-1.77, -0.08)
cg10092685	16	73090591	*ZFHX3*	-0.97 (-1.64, -0.24)	-1.15 (-1.96, -0.34)	-0.86 (-1.64, -0.1)
cg10251538	3	108800886	*MORC1*	-1.22 (-1.95, -0.58)	-1.28 (-2.02, -0.53)	-1.19 (-1.9, -0.46)
cg11202345	17	76976057	*LGALS3BP*	1.66 (0.94, 2.3)	1.71 (1, 2.42)	1.65 (0.95, 2.34)
cg11591807	2	118888717	*INSIG2*	1.39 (0.56, 2.31)	1.7 (0.76, 2.64)	1.23 (0.32, 2.15)
cg11614060	9	137660527	*COL5A1*	1.09 (0.37, 1.71)	1.27 (0.57, 1.97)	0.91 (0.22, 1.6)
cg12628550	14	91817627	*CCDC88C*	1.39 (0.68, 1.98)	1.5 (0.81, 2.2)	1.33 (0.65, 2)
cg12924402	2	218898511	*RUFY4*	-1.17 (-1.87, -0.37)	-1.45 (-2.36, -0.55)	-1.02 (-1.88, -0.17)
cg13549904	1	154438143	*IL6R*	-1.25 (-1.89, -0.66)	-1.32 (-1.84, -0.8)	-1.22 (-1.73, -0.71)
cg14896076	12	52225262	*FIGNL2*	1.06 (0.23, 1.88)	1.3 (0.39, 2.21)	0.91 (0.07, 1.8)
cg15786705	6	28176104	*TOB2P1*	0.8 (0.05, 1.36)	0.9 (0.24, 1.56)	0.76 (0.15, 1.39)
cg16032415	8	95278692	*GEM*	-1.01 (-1.8, -0.25)	-1.23 (-2.05, -0.4)	-0.88 (-1.67, -0.1)
cg16406078	20	825634	*FAM110A*	-1.36 (-1.94, -0.72)	-1.55 (-2.24, -0.87)	-1.26 (-1.93, -0.6)
cg16740586	21	43655919	*ABCG1*	0.84 (0.06, 1.53)	0.96 (0.16, 1.76)	0.74 (0.02, 1.52)
cg16758086	1	6173356	*CHD5*	-1.24 (-1.93, -0.47)	-1.42 (-2.23, -0.61)	-1.16 (-1.94, -0.36)
cg17468665	12	56221379	*DNAJC14*	1.64 (0.77, 2.43)	1.98 (0.97, 2.98)	1.46 (0.49, 2.42)
cg18322280	14	57793087	*AP5M1*	-1.09 (-1.65, -0.43)	-1.25 (-1.93, -0.56)	-0.99 (-1.66, -0.33)
cg18581607	17	4714084	*PLD2*	-1.13 (-1.78, -0.39)	-1.32 (-2.13, -0.51)	-0.96 (-1.73, -0.2)
cg18613281	1	39596444	*MACF1*	-1.45 (-2.09, -0.74)	-1.71 (-2.52, -0.9)	-1.34 (-2.13, -0.53)
cg19534021	19	16178091	*TPM4*	1.07 (0.38, 1.73)	1.25 (0.44, 2.06)	0.91 (0.13, 1.7)
cg22648996	10	63946213	*RTKN2*	2.22 (1.21, 3.28)	2.62 (1.35, 3.89)	2.29 (1.08, 3.52)
cg25919221	1	9006680	*CA6*	1.28 (0.53, 2.04)	1.52 (0.58, 2.46)	1.12 (0.23, 2.04)
cg26416168	2	71934434	*DYSF*	1.04 (0.34, 1.8)	1.02 (0.2, 1.84)	1.01 (0.21, 1.8)
cg26439401	12	103849421	*C12orf42*	0.63 (0.17, 1)	0.77 (0.26, 1.29)	0.58 (0.09, 1.08)
cg26467270	17	76718664	*CYTH1*	-2.98 (-3.64, -2.22)	-3.26 (-4.02, -2.49)	-2.93 (-3.67, -2.19)
cg26800893	11	67184596	*CARNS1*	-1.48 (-2.1, -0.86)	-1.64 (-2.32, -0.96)	-1.42 (-2.08, -0.76)
cg27080917	12	11978350	*ETV6*	-1.42 (-2.15, -0.72)	-1.67 (-2.49, -0.86)	-1.28 (-2.08, -0.48)

Abbreviations: ISIS, Iterative Sure Independence Screening, enet, elastic-net; LS, least squares.

Table A4: Mean differences (95 % CI) for the CpGs selected by ISIS - LASSO for BMI and comparison with linear regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	LS	Bayes
cg00157319	8	74360923	*STAU2*	0.85 (0.14, 1.57)	0.89 (0.2, 1.59)	0.73 (0.08, 1.4)
cg00602326	5	31427700	*DROSHA*	-2.04 (-2.81, -1.32)	-2.11 (-2.85, -1.37)	-1.84 (-2.57, -1.13)
cg01500427	2	19546666	*MIR4757*	1.12 (0.11, 1.92)	1.14 (0.22, 2.07)	0.86 (0.05, 1.74)
cg01577114	14	61906283	*PRKCH*	1.04 (0.44, 1.6)	1.08 (0.5, 1.66)	0.91 (0.34, 1.48)
cg01880404	19	44079527	*XRCC1*	1.06 (0.39, 1.66)	1.09 (0.53, 1.65)	0.96 (0.41, 1.53)
cg02411354	3	187697534	*LPP-AS2*	-1.83 (-3.05, -0.61)	-2.01 (-3.19, -0.83)	-1.27 (-2.41, -0.18)
cg03078551	17	41656298	*ETV4*	-1.24 (-1.87, -0.61)	-1.28 (-1.9, -0.65)	-1.12 (-1.74, -0.52)
cg03710333	4	1722958	*TMEM129*	0.81 (0.26, 1.35)	0.8 (0.3, 1.3)	0.79 (0.3, 1.28)
cg03748458	8	14711248	*SGCZ*	-1.07 (-1.84, -0.28)	-1.2 (-2, -0.41)	-0.82 (-1.58, -0.08)
cg04086239	16	24067174	*PRKCB*	0.7 (0.02, 1.42)	0.73 (0.08, 1.39)	0.67 (0.05, 1.32)
cg04245590	12	85673412	*ALX1*	1.52 (0.3, 2.35)	1.65 (0.68, 2.61)	1.18 (0.25, 2.11)
cg05575921	5	373378	*AHRR*	1.74 (0.64, 2.77)	1.78 (0.68, 2.87)	1.52 (0.5, 2.51)
cg06235693	19	52275119	*FPR2*	1.13 (0.43, 1.71)	1.21 (0.61, 1.81)	0.91 (0.33, 1.5)
cg06548519	17	34267111	*LYZL6*	0.6 (0.07, 1.12)	0.61 (0.09, 1.13)	0.54 (0.04, 1.05)
cg06548673	15	45473947	*SHF*	0.98 (0.26, 1.73)	1.01 (0.27, 1.74)	0.82 (0.11, 1.54)
cg08018468	1	43768048	*TIE1*	-1.2 (-1.92, -0.45)	-1.34 (-2.05, -0.63)	-0.91 (-1.6, -0.21)
cg08125271	7	22143318	*RAPGEF5*	0.93 (0.16, 1.7)	0.97 (0.21, 1.73)	0.78 (0.07, 1.52)
cg08389486	9	132377983	*C9orf50*	0.87 (0.23, 1.47)	0.89 (0.25, 1.52)	0.75 (0.13, 1.38)
cg08633893	13	100068932	*MIR548AN*	1.21 (0.22, 2.03)	1.27 (0.38, 2.15)	1.1 (0.26, 1.93)
cg09074260	11	94707049	*KDM4D*	0.75 (0.09, 1.29)	0.75 (0.23, 1.28)	0.7 (0.2, 1.21)
cg09364595	9	139457749	*MIR4674*	-1.28 (-1.97, -0.62)	-1.38 (-2.06, -0.69)	-1.13 (-1.8, -0.45)
cg09658645	17	77704767	*ENPP7*	0.93 (0.35, 1.54)	0.96 (0.38, 1.54)	0.81 (0.24, 1.38)
cg10092685	16	73090591	*ZFHX3*	-1.16 (-1.88, -0.4)	-1.18 (-1.91, -0.45)	-1.04 (-1.76, -0.32)
cg11099291	17	1620044	*WDR81*	0.87 (0.33, 1.51)	0.91 (0.36, 1.46)	0.73 (0.19, 1.28)
cg11202345	17	76976057	*LGALS3BP*	1.84 (1.07, 2.45)	1.88 (1.21, 2.54)	1.74 (1.09, 2.39)
cg11473706	4	83733078	*SEC31A*	-1.07 (-1.81, -0.31)	-1.13 (-1.85, -0.42)	-0.83 (-1.53, -0.15)
cg11474081	8	129234520	*MIR1208*	1.93 (0.55, 3.28)	2.2 (0.85, 3.55)	1.36 (0.13, 2.66)
cg11591807	2	118888717	*INSIG2*	1.25 (0.37, 2.18)	1.29 (0.41, 2.17)	1.08 (0.23, 1.94)
cg11614060	9	137660527	*COL5A1*	0.78 (0.1, 1.38)	0.82 (0.17, 1.47)	0.65 (0.04, 1.3)
cg11625476	17	4795410	*MINK1*	1.21 (0.49, 1.87)	1.25 (0.61, 1.9)	1.05 (0.42, 1.69)
cg11743438	6	16238437	*GMPR*	0.99 (0.42, 1.55)	1.02 (0.47, 1.56)	0.9 (0.36, 1.44)
cg12859382	3	52445103	*PHF7*	0.66 (0.07, 1.23)	0.68 (0.09, 1.26)	0.61 (0.06, 1.18)
cg12915892	2	134024093	*NCKAP5*	1.12 (0.13, 2.04)	1.21 (0.31, 2.11)	0.92 (0.09, 1.79)
cg13182145	9	16179702	*C9orf92*	1.59 (0.62, 2.59)	1.72 (0.76, 2.68)	1.28 (0.32, 2.22)
cg13549904	1	154438143	*IL6R*	-1.35 (-1.9, -0.81)	-1.37 (-1.86, -0.89)	-1.28 (-1.76, -0.81)
cg13681954	2	122656302	*TSN*	-0.85 (-1.67, -0.1)	-0.88 (-1.63, -0.13)	-0.73 (-1.47, -0.03)

cg14585186	12	104974102	*CHST11*	1.19 (0.08, 2.12)	1.21 (0.19, 2.23)	1.06 (0.11, 2.04)
cg14969094	3	156848003	*LINC00880*	0.97 (0.26, 1.69)	1.05 (0.3, 1.8)	0.81 (0.1, 1.55)
cg15144123	3	47655398	*SMARCC1*	1.25 (0.37, 2.01)	1.36 (0.56, 2.16)	1.02 (0.25, 1.8)
cg15340629	22	27725596	*MN1*	0.84 (0.19, 1.49)	0.87 (0.23, 1.51)	0.76 (0.15, 1.38)
cg15706574	6	46231809	*RCAN2*	1.13 (0.48, 1.73)	1.14 (0.46, 1.81)	1.05 (0.39, 1.71)
cg16032415	8	95278692	*GEM*	-1.31 (-2.11, -0.5)	-1.38 (-2.14, -0.63)	-1.12 (-1.86, -0.38)
cg16153294	11	2018227	*H19*	0.68 (0.08, 1.2)	0.71 (0.12, 1.3)	0.59 (0.03, 1.17)
cg16406078	20	825634	*FAM110A*	-1.9 (-2.52, -1.21)	-1.98 (-2.61, -1.34)	-1.68 (-2.3, -1.06)
cg16740586	21	43655919	*ABCG1*	0.85 (0.09, 1.57)	0.91 (0.17, 1.65)	0.73 (0.04, 1.46)
cg16758086	1	6173356	*CHD5*	-1.53 (-2.28, -0.76)	-1.59 (-2.35, -0.84)	-1.34 (-2.08, -0.6)
cg16958927	15	51970968	*SCG3*	0.53 (0.25, 0.75)	0.54 (0.22, 0.86)	0.46 (0.15, 0.78)
cg17420142	18	32702783	*MAPRE2*	-0.86 (-1.77, -0.13)	-0.88 (-1.65, -0.11)	-0.82 (-1.58, -0.08)
cg17683449	22	39760036	*SYNGR1*	1.09 (0.32, 1.89)	1.18 (0.41, 1.95)	0.88 (0.13, 1.63)
cg18140642	1	236094805	*NID1*	1.13 (0.21, 2.04)	1.17 (0.25, 2.09)	0.88 (0.05, 1.77)
cg18322280	14	57793087	*AP5M1*	-1.76 (-2.35, -1.05)	-1.87 (-2.53, -1.22)	-1.51 (-2.15, -0.87)
cg18613281	1	39596444	*MACF1*	-2.04 (-2.75, -1.35)	-2.2 (-2.96, -1.44)	-1.67 (-2.42, -0.9)
cg18632602	18	21978020	*OSBPL1A*	0.83 (0.17, 1.43)	0.85 (0.23, 1.48)	0.74 (0.13, 1.34)
cg19992857	2	201936687	*NDUFB3*	1.21 (0.69, 1.76)	1.26 (0.71, 1.81)	1.04 (0.5, 1.59)
cg20587236	12	109900956	*KCTD10*	1.37 (0.65, 2.01)	1.4 (0.74, 2.05)	1.26 (0.61, 1.91)
cg20936142	2	236406027	*AGAP1*	0.82 (0.25, 1.4)	0.88 (0.32, 1.45)	0.64 (0.09, 1.2)
cg21217117	2	65190910	*SLC1A4*	-0.87 (-1.64, -0.3)	-0.95 (-1.59, -0.31)	-0.67 (-1.3, -0.06)
cg22177704	2	241533597	*CAPN10*	0.56 (0.08, 1.02)	0.6 (0.11, 1.08)	0.46 (0.01, 0.94)
cg22371743	1	31192334	*MATN1*	0.7 (0.14, 1.27)	0.75 (0.2, 1.3)	0.57 (0.04, 1.1)
cg22648996	10	63946213	*RTKN2*	2.1 (0.92, 3.35)	2.19 (1, 3.38)	1.83 (0.67, 2.99)
cg23615467	1	25695799	*RHCE*	0.89 (0.18, 1.58)	0.9 (0.26, 1.55)	0.82 (0.2, 1.45)
cg25919221	1	9006680	*CA6*	1.33 (0.43, 2.3)	1.41 (0.47, 2.35)	1.03 (0.15, 1.94)
cg26416168	2	71934434	*DYSF*	0.84 (0.06, 1.63)	0.87 (0.1, 1.64)	0.72 (0.02, 1.48)
cg26439401	12	103849421	*C12orf42*	0.84 (0.33, 1.21)	0.88 (0.41, 1.35)	0.71 (0.25, 1.17)
cg26467270	17	76718664	*CYTH1*	-3.87 (-4.56, -3.11)	-3.98 (-4.68, -3.27)	-3.53 (-4.23, -2.84)
cg26800893	11	67184596	*CARNS1*	-1.48 (-2.21, -0.82)	-1.47 (-2.09, -0.84)	-1.48 (-2.09, -0.86)

Abbreviations: ISIS, Iterative Sure Independence Screening; LASSO, Least Absolute Shrinkage and Selection Operator; LS, least squares.

Table A5: Mean differences (95 % CI) for the CpGs selected by ISIS - SCAD for BMI and comparison with linear regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	LS	Bayes
cg00309970	1	203208795	*CHIT1*	0.13 (0.01, 0.23)	0.13 (0.05, 0.22)	0.11 (0.02, 0.19)
cg00440217	6	71012765	*COL9A1*	0.74 (0.07, 1.25)	0.76 (0.27, 1.26)	0.7 (0.21, 1.2)
cg00602326	5	31427700	*DROSHA*	-2.34 (-3.01, -1.76)	-2.35 (-2.97, -1.73)	-2.18 (-2.8, -1.56)
cg00831028	20	55043745	*RTF2*	0.85 (0.01, 1.52)	0.86 (0.27, 1.45)	0.74 (0.17, 1.33)
cg00880429	5	171520729	*STK10*	1.11 (0.49, 1.74)	1.11 (0.55, 1.67)	0.95 (0.4, 1.51)
cg01529701	18	46572234	*DYM*	1.5 (0.76, 2.19)	1.42 (0.77, 2.08)	1.23 (0.57, 1.88)
cg01874871	20	2644800	*IDH3B*	1.11 (0.02, 1.94)	1.19 (0.44, 1.94)	1.02 (0.28, 1.77)
cg01953134	6	90348409	*LYRM2*	0.39 (0.01, 0.66)	0.39 (0.11, 0.67)	0.34 (0.06, 0.61)
cg03078551	17	41656298	*ETV4*	-1.73 (-2.29, -1.08)	-1.73 (-2.25, -1.2)	-1.61 (-2.14, -1.09)
cg03710333	4	1722958	*TMEM129*	1.02 (0.48, 1.59)	1.01 (0.54, 1.47)	0.96 (0.5, 1.42)
cg05575921	5	373378	*AHRR*	2.27 (1.6, 2.89)	2.24 (1.64, 2.85)	2.13 (1.53, 2.73)
cg06235693	19	52275119	*FPR2*	0.95 (0.06, 1.53)	0.94 (0.42, 1.46)	0.8 (0.29, 1.31)
cg06834534	10	15001276	*DCLRE1C*	0.19 (0.02, 0.31)	0.19 (0.08, 0.3)	0.16 (0.06, 0.27)
cg07060261	13	90557520	*LINC00559*	1.37 (0.1, 2.17)	1.39 (0.64, 2.15)	1.13 (0.39, 1.87)
cg07599607	4	114682255	*CAMK2D*	0.86 (0.03, 1.42)	0.86 (0.35, 1.38)	0.67 (0.16, 1.18)
cg07706844	10	102509510	*PAX2*	0.87 (0.41, 1.38)	0.88 (0.47, 1.29)	0.85 (0.44, 1.26)
cg08125271	7	22143318	*RAPGEF5*	1.07 (0.04, 1.8)	1.05 (0.38, 1.72)	0.94 (0.27, 1.6)
cg08486432	6	33598003	*ITPR3*	1.28 (0.54, 1.97)	1.26 (0.63, 1.9)	1.14 (0.5, 1.77)
cg08633893	13	100068932	*MIR548AN*	1.41 (0.62, 2.37)	1.52 (0.75, 2.28)	1.42 (0.66, 2.18)
cg09043226	6	32146099	*RNF5*	0.47 (0.03, 0.87)	0.46 (0.15, 0.78)	0.39 (0.08, 0.71)
cg09074260	11	94707049	*KDM4D*	1 (0.41, 1.54)	0.98 (0.52, 1.44)	0.84 (0.38, 1.3)
cg09506600	1	248100228	*OR2L13*	0.84 (0.18, 1.41)	0.84 (0.34, 1.34)	0.73 (0.23, 1.23)
cg09554443	1	167487762	*CD247*	-1.58 (-2.3, -0.84)	-1.54 (-2.24, -0.85)	-1.4 (-2.1, -0.71)
cg10251538	3	108800886	*MORC1*	-1.78 (-2.44, -1.15)	-1.78 (-2.39, -1.17)	-1.63 (-2.25, -1.02)
cg11157034	1	168344184	*MIR557*	0.22 (0.07, 0.34)	0.22 (0.1, 0.34)	0.19 (0.07, 0.3)
cg11202345	17	76976057	*LGALS3BP*	1.24 (0.65, 1.88)	1.25 (0.69, 1.82)	1.27 (0.71, 1.83)
cg13089947	12	26277925	*BHLHE41*	1.06 (0.38, 1.7)	1.04 (0.46, 1.63)	0.95 (0.37, 1.53)
cg13381660	1	44432698	*IPO13*	1.06 (0.09, 1.51)	1.07 (0.56, 1.57)	0.86 (0.36, 1.36)
cg13549904	1	154438143	*IL6R*	-1.25 (-1.89, -0.66)	-1.26 (-1.71, -0.81)	-1.2 (-1.64, -0.75)
cg14108978	11	73021145	*ARHGEF17*	0.96 (0.46, 1.44)	0.96 (0.48, 1.44)	0.9 (0.41, 1.38)
cg14782266	2	42565367	*EML4*	1.17 (0.05, 1.88)	1.11 (0.44, 1.78)	1.02 (0.35, 1.68)
cg16368504	21	33415990	*LINC00159*	1.04 (0.53, 1.62)	1.07 (0.55, 1.58)	0.95 (0.44, 1.47)
cg16774354	4	6576608	*MAN2B2*	0.97 (0.08, 1.75)	0.99 (0.4, 1.59)	0.88 (0.29, 1.47)
cg17468665	12	56221379	*DNAJC14*	2.11 (1.15, 3.01)	2.11 (1.22, 3.01)	1.6 (0.71, 2.5)
cg17495627	13	114321698	*GRK1*	0.82 (0.12, 1.3)	0.81 (0.34, 1.28)	0.75 (0.28, 1.22)
cg18322280	14	57793087	*AP5M1*	-1.34 (-1.96, -0.71)	-1.32 (-1.9, -0.75)	-1.21 (-1.79, -0.64)

cg20165604	8	134686291	*ST3GAL1*	0.94 (0.08, 1.53)	0.92 (0.33, 1.5)	0.8 (0.22, 1.38)
cg20481941	16	80604148	*DYNLRB2*	1.1 (0.51, 1.73)	1.1 (0.58, 1.61)	0.97 (0.46, 1.49)
cg20993361	7	116503444	*CAPZA2*	1.1 (0.21, 1.82)	1.09 (0.47, 1.71)	0.93 (0.31, 1.54)
cg21096502	7	56174374	*PSPH*	0.66 (0.08, 1.17)	0.68 (0.23, 1.13)	0.64 (0.19, 1.08)
cg21790695	10	35070092	*PARD3*	0.95 (0.27, 1.44)	0.97 (0.48, 1.45)	0.83 (0.36, 1.32)
cg25240153	16	23890018	*PRKCB*	1.61 (0.72, 2.43)	1.57 (0.79, 2.35)	1.41 (0.64, 2.18)
cg25695193	2	45010550	*CAMKMT*	0.53 (0.03, 0.93)	0.51 (0.11, 0.9)	0.44 (0.05, 0.83)
cg26337592	1	153642686	*ILF2*	1.88 (1.05, 2.68)	1.86 (1.09, 2.63)	1.53 (0.78, 2.29)
cg26439401	12	103849421	*C12orf42*	1.31 (0.75, 1.65)	1.3 (0.88, 1.71)	1.09 (0.67, 1.5)
cg26467270	17	76718664	*CYTH1*	-4.04 (-4.62, -3.36)	-4.05 (-4.66, -3.45)	-3.76 (-4.36, -3.15)
cg26542597	19	38682823	*SIPA1L3*	1.32 (0.66, 1.96)	1.34 (0.74, 1.95)	1.17 (0.57, 1.78)
cg26800893	11	67184596	*CARNS1*	-1.97 (-2.54, -1.39)	-1.98 (-2.52, -1.44)	-1.82 (-2.36, -1.27)

Abbreviations: ISIS, Iterative Sure Independence Screening; SCAD, Smoothly Clipped Absolute Deviation; LS, least squares.

Table A6: Mean differences (95 % CI) for the CpGs selected by ISIS - MCP for BMI and comparison with linear regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	LS	Bayes
cg00005164	19	35511522	*GRAMD1A*	1.15 (0.52, 1.8)	1.14 (0.55, 1.74)	1.04 (0.45, 1.64)
cg00602326	5	31427700	*DROSHA*	-2.1 (-2.69, -1.5)	-2.1 (-2.67, -1.53)	-1.98 (-2.57, -1.4)
cg01530962	3	9027157	*SRGAP3*	1.17 (0.49, 1.71)	1.16 (0.57, 1.74)	0.9 (0.31, 1.49)
cg01855290	21	45132261	*PDXK*	1.48 (0.76, 2.12)	1.47 (0.86, 2.09)	1.31 (0.69, 1.93)
cg02715531	17	77709027	*ENPP7*	1.02 (0.52, 1.5)	1.02 (0.54, 1.5)	0.91 (0.43, 1.4)
cg02753444	6	158492604	*SYNJ2*	0.24 (0.13, 0.33)	0.24 (0.14, 0.34)	0.21 (0.11, 0.3)
cg03078551	17	41656298	*ETV4*	-2.1 (-2.65, -1.46)	-2.08 (-2.6, -1.56)	-1.88 (-2.4, -1.37)
cg03272499	13	66919912	*PCDH9*	0.52 (0.03, 0.83)	0.52 (0.18, 0.87)	0.4 (0.06, 0.75)
cg03710333	4	1722958	*TMEM129*	1.01 (0.48, 1.51)	1.01 (0.56, 1.46)	0.94 (0.5, 1.39)
cg03944143	17	2595123	*CLUH*	0.96 (0.2, 1.6)	0.95 (0.37, 1.53)	0.82 (0.24, 1.4)
cg05575921	5	373378	*AHRR*	2.69 (2.14, 3.27)	2.68 (2.12, 3.24)	2.59 (2.03, 3.16)
cg07060261	13	90557520	*LINC00559*	2.13 (1.24, 2.88)	2.13 (1.4, 2.86)	1.8 (1.07, 2.53)
cg07443900	4	1018378	*FGFRL1*	0.79 (0.14, 1.27)	0.79 (0.31, 1.27)	0.67 (0.19, 1.15)
cg07706844	10	102509510	*PAX2*	0.93 (0.44, 1.42)	0.94 (0.54, 1.33)	0.9 (0.5, 1.29)
cg08242024	7	157551111	*PTPRN2*	0.36 (0.01, 0.63)	0.35 (0.08, 0.63)	0.35 (0.07, 0.62)
cg08297094	19	33166201	*ANKRD27*	0.75 (0.26, 1.19)	0.75 (0.33, 1.17)	0.68 (0.26, 1.1)
cg08633893	13	100068932	*MIR548AN*	1.81 (1.01, 2.6)	1.8 (1.13, 2.48)	1.68 (1, 2.35)
cg08703857	7	2653651	*IQCE*	1.03 (0.24, 1.66)	1 (0.41, 1.6)	0.9 (0.32, 1.49)
cg09718708	4	48272297	*TEC*	0.92 (0.17, 1.53)	0.93 (0.36, 1.5)	0.79 (0.23, 1.36)
cg10251538	3	108800886	*MORC1*	-1.21 (-1.82, -0.6)	-1.2 (-1.78, -0.62)	-1.15 (-1.73, -0.57)
cg10948061	6	110500990	*WASF1*	0.9 (0.37, 1.46)	0.9 (0.41, 1.4)	0.84 (0.34, 1.34)
cg11058916	16	34257749	*UBE2MP1*	0.9 (0.34, 1.46)	0.89 (0.36, 1.43)	0.81 (0.27, 1.34)
cg11625476	17	4795410	*MINK1*	1.57 (0.9, 2.18)	1.57 (0.99, 2.15)	1.38 (0.79, 1.97)
cg11739303	3	39952693	*MYRIP*	0.3 (0.08, 0.5)	0.31 (0.12, 0.49)	0.26 (0.08, 0.44)
cg12998942	4	103781683	*UBE2D3*	0.92 (0.14, 1.61)	0.93 (0.33, 1.53)	0.81 (0.21, 1.41)
cg13414270	2	45465395	*LINC01121*	0.9 (0.35, 1.53)	0.91 (0.35, 1.46)	0.83 (0.28, 1.38)
cg13549904	1	154438143	*IL6R*	-1.59 (-2.23, -0.94)	-1.59 (-2.02, -1.16)	-1.49 (-1.93, -1.07)
cg14037728	9	116645936	*ZNF618*	0.84 (0.03, 1.46)	0.84 (0.26, 1.41)	0.79 (0.21, 1.37)
cg14566095	16	55876964	*CES5A*	1.91 (1.03, 2.69)	1.9 (1.17, 2.63)	1.69 (0.96, 2.42)
cg14782266	2	42565367	*EML4*	1.27 (0.59, 1.94)	1.26 (0.61, 1.91)	1.13 (0.48, 1.78)
cg15243454	2	233415061	*TIGD1*	1.01 (0.46, 1.57)	1 (0.47, 1.52)	0.89 (0.36, 1.42)
cg15251779	7	150929295	*CHPF2*	0.89 (0.31, 1.45)	0.89 (0.4, 1.39)	0.77 (0.27, 1.27)
cg15706574	6	46231809	*RCAN2*	1.01 (0.47, 1.6)	1.01 (0.47, 1.56)	0.94 (0.39, 1.48)
cg16774354	4	6576608	*MAN2B2*	0.97 (0.18, 1.67)	0.96 (0.38, 1.54)	0.84 (0.26, 1.42)
cg16958927	15	51970968	*SCG3*	0.53 (0.24, 0.81)	0.53 (0.25, 0.81)	0.51 (0.22, 0.79)
cg17995403	2	95831296	*ZNF2*	1.05 (0.51, 1.57)	1.04 (0.53, 1.55)	0.88 (0.37, 1.4)
cg18322280	14	57793087	*AP5M1*	-1.71 (-2.33, -1.05)	-1.7 (-2.27, -1.13)	-1.52 (-2.1, -0.96)
cg19141201	14	23388712	*RBM23*	0.86 (0.32, 1.34)	0.85 (0.37, 1.32)	0.83 (0.35, 1.3)
cg20223677	8	7332846	*DEFB104B*	0.73 (0.18, 1.22)	0.72 (0.29, 1.16)	0.61 (0.18, 1.05)

cg20315590	1	186003041	*HMCN1*	0.87 (0.12, 1.36)	0.86 (0.34, 1.38)	0.7 (0.18, 1.22)
cg20481941	16	80604148	*DYNLRB2*	1.34 (0.85, 1.94)	1.34 (0.86, 1.83)	1.23 (0.74, 1.71)
cg20562176	19	8008963	*TIMM44*	0.94 (0.38, 1.46)	0.94 (0.43, 1.45)	0.82 (0.31, 1.34)
cg20587236	12	109900956	*KCTD10*	1.39 (0.81, 1.99)	1.39 (0.83, 1.95)	1.29 (0.73, 1.85)
cg21687775	1	146989469	*LINC00624*	0.97 (0.5, 1.47)	0.98 (0.47, 1.48)	0.88 (0.37, 1.39)
cg22699725	1	207242586	*PFKFB2*	0.85 (0.07, 1.51)	0.83 (0.28, 1.39)	0.81 (0.26, 1.36)
cg24106020	1	181452827	*CACNA1E*	0.75 (0.01, 1.24)	0.74 (0.3, 1.19)	0.64 (0.19, 1.09)
cg24523250	1	241230132	*RGS7*	0.85 (0.36, 1.3)	0.85 (0.42, 1.28)	0.73 (0.3, 1.16)
cg24591090	3	125094085	*ZNF148*	0.75 (0.07, 1.32)	0.74 (0.24, 1.24)	0.67 (0.17, 1.17)
cg26337592	1	153642686	*ILF2*	1.29 (0.49, 2.04)	1.3 (0.57, 2.04)	1.13 (0.39, 1.87)
cg26416168	2	71934434	*DYSF*	1.23 (0.49, 1.91)	1.21 (0.53, 1.88)	1.08 (0.4, 1.76)
cg26439401	12	103849421	*C12orf42*	1.06 (0.63, 1.42)	1.07 (0.67, 1.46)	0.9 (0.51, 1.31)
cg26467270	17	76718664	*CYTH1*	-4.61 (-5.13, -3.87)	-4.61 (-5.18, -4.04)	-4.27 (-4.85, -3.7)
cg26800893	11	67184596	*CARNS1*	-1.86 (-2.42, -1.29)	-1.86 (-2.38, -1.35)	-1.75 (-2.27, -1.24)
cg27604402	6	31765590	*LSM2*	1.56 (0.75, 2.33)	1.55 (0.83, 2.28)	1.3 (0.58, 2.02)

Abbreviations: ISIS, Iterative Sure Independence Screening; MCP, Minimax Concave Penalty; LS, least squares.

Table A7: Hazard ratios (95 % CI) for the CpGs selected by ISIS - Aenet comparing percentile 90th vs 10th for lung cancer and comparison with Cox regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	Cox	Bayes
cg03362418	22	50965563	*TYMP*	0.47 (0.25, 0.88)	0.2 (0.06, 0.66)	0.65 (0.36, 1.14)
cg03650729	1	47692625	*TAL1*	2.73 (1.45, 5.12)	8.12 (2.75, 23.95)	1.67 (1, 2.91)
cg04000528	4	26702926	*TBC1D19*	1.15 (1, 1.35)	5.14 (1.06, 24.93)	1.04 (0.76, 1.41)
cg04227931	4	180386617	*LINC01098*	1.15 (1, 1.31)	3.33 (1.1, 10.08)	1.05 (0.85, 1.32)
cg06285727	11	72524028	*ATG16L2*	0.51 (0.31, 0.86)	0.23 (0.07, 0.83)	0.73 (0.42, 1.23)
cg10113527	9	134109740	*NUP214*	2.03 (1.06, 3.33)	4.91 (1.6, 15.13)	1.52 (0.9, 2.6)
cg10684686	6	28557047	*ZBED9*	1.88 (1.26, 2.65)	6.52 (2.27, 18.75)	1.33 (0.9, 2.05)
cg13777023	22	20964020	*MED15*	1.21 (1.1, 1.33)	1.77 (1.07, 2.95)	1.09 (0.97, 1.25)
cg14273031	14	22320448	*OR4E2*	1.44 (1.1, 2.05)	4.95 (1.7, 14.44)	1.16 (0.84, 1.62)
cg17746033	5	153828051	*SAP30L*	1.52 (1.25, 1.94)	4.74 (1.64, 13.73)	1.18 (0.96, 1.53)
cg21990700	12	7260776	*C1RL*	0.42 (0.22, 0.91)	0.27 (0.1, 0.77)	0.65 (0.4, 1.05)
cg21999471	11	128555317	*FLI1*	2.59 (1.46, 4.62)	9.52 (3.29, 27.59)	1.61 (0.98, 2.73)
cg27209729	11	64428925	*NRXN2*	0.45 (0.25, 0.77)	0.28 (0.11, 0.7)	0.67 (0.4, 1.08)
ch.9.1286602F	9	93982668	*AUH*	1.69 (1.15, 2.39)	5.49 (1.61, 18.65)	1.29 (0.87, 1.95)

Abbreviations: ISIS, Iterative Sure Independence Screening; Aenet, adaptive elastic-net.

Table A8: Hazard ratios (95 % CI) for the CpGs selected by ISIS - MSAenet comparing percentile 90th vs 10th for lung cancer and comparison with Cox regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	Cox	Bayes
cg00780810	15	70780653	*UACA*	0.33 (0.09, 0.84)	0.27 (0.08, 0.89)	0.64 (0.36, 1.09)
cg00841849	2	8683604	*ID2*	0.47 (0.1, 1)	0.21 (0.05, 0.8)	0.72 (0.39, 1.28)
cg03259188	2	118860242	*INSIG2*	2.24 (1.15, 7.96)	3.96 (1.37, 11.43)	1.31 (0.91, 1.9)
cg03362418	22	50965563	*TYMP*	0.61 (0.14, 1)	0.26 (0.08, 0.89)	0.72 (0.4, 1.29)
cg03650729	1	47692625	*TAL1*	3.31 (1.36, 10.89)	4.6 (1.66, 12.74)	1.54 (0.93, 2.57)
cg05021589	6	6588931	*LY86-AS1*	3.2 (1.32, 9.52)	4.75 (1.52, 14.82)	1.6 (0.95, 2.71)
cg07012499	2	11618241	*E2F6*	2.47 (1.24, 6.02)	4.5 (1.89, 10.7)	1.35 (0.97, 1.91)
cg09650907	17	71224983	*FAM104A*	2.56 (1, 8.05)	3.46 (1.03, 11.63)	1.48 (0.82, 2.76)
cg09984392	8	126011784	*SQLE*	2.84 (1.27, 9.07)	3.24 (1.3, 8.08)	1.61 (0.97, 2.69)
cg10113527	9	134109740	*NUP214*	3.33 (1.43, 11.67)	5.82 (1.44, 23.48)	1.7 (0.99, 3.06)
cg11911122	8	71316769	*NCOA2*	2.28 (1, 6.29)	3.63 (1.25, 10.61)	1.47 (0.87, 2.52)
cg22454769	2	106015767	*FHL2*	3.05 (1.73, 8.96)	4.06 (1.68, 9.79)	1.77 (1.15, 2.77)
cg25544931	19	12097624	*ZNF763*	4.41 (1.93, 15.12)	4.48 (1.61, 12.47)	1.82 (1.16, 2.83)
cg27209729	11	64428925	*NRXN2*	0.51 (0.19, 0.95)	0.36 (0.14, 0.89)	0.66 (0.41, 1.05)
ch.2.1365132F	2	59899176	*LINC01122*	1.48 (1, 3.28)	2.84 (1.51, 5.33)	1.07 (0.9, 1.28)

Abbreviations: ISIS, Iterative Sure Independence Screening; MSAenet, multi-step adaptive elastic-net.

Table A9: Hazard ratios (95 % CI) for the CpGs selected by ISIS - enet comparing percentile 90th vs 10th for lung cancer and comparison with Cox regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	Cox	Bayes
cg03259188	2	118860242	*INSIG2*	1.7 (1.16, 2.79)	3.22 (1.29, 8.02)	1.2 (0.88, 1.7)
cg03650729	1	47692625	*TAL1*	3.38 (1.58, 7.3)	6.37 (2.2, 18.46)	1.61 (0.98, 2.72)
cg10113527	9	134109740	*NUP214*	2.41 (1.14, 4.5)	2.94 (1.03, 8.43)	1.54 (0.92, 2.63)
cg11185549	12	116996871	*MAP1LC3B2*	1.86 (1, 3.48)	4.68 (1.23, 17.78)	1.23 (0.74, 2.15)
cg11469818	17	78093132	*GAA*	1.46 (1.18, 1.84)	2.45 (1.08, 5.52)	1.14 (0.96, 1.4)
cg13777023	22	20964020	*MED15*	1.22 (1.1, 1.38)	1.74 (1.02, 2.96)	1.08 (0.97, 1.22)
cg22454769	2	106015767	*FHL2*	2.41 (1.41, 4.52)	5.57 (2.17, 14.29)	1.45 (0.95, 2.24)
cg27209729	11	64428925	*NRXN2*	0.52 (0.26, 0.96)	0.33 (0.13, 0.82)	0.74 (0.46, 1.16)
ch.1.374405F	1	10177152	*UBE4B*	2.15 (1.17, 3.07)	6.57 (3.38, 12.77)	1.28 (0.99, 1.69)

Abbreviations: ISIS, Iterative Sure Independence Screening; enet, elastic-net.

Table A10: Hazard ratios (95 % CI) for the CpGs selected by ISIS - LASSO comparing percentile 90th vs 10th for lung cancer and comparison with Cox regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	Cox	Bayes
cg00480032	1	111459642	*CD53*	2.87 (1, 5.94)	3.55 (1.77, 7.11)	1.23 (0.89, 1.74)
cg02371147	4	37839219	*PGM2*	3.58 (1, 9.18)	4.33 (1.83, 10.23)	1.55 (0.96, 2.54)
cg03650729	1	47692625	*TAL1*	2.55 (1, 6.59)	2.57 (1.11, 5.93)	1.65 (1, 2.73)
cg06318011	1	36567283	*COL8A2*	1.53 (1, 2.13)	1.63 (1.01, 2.63)	1.16 (0.98, 1.42)
cg07428919	11	31391095	*DCDC1*	7.04 (1, 16.83)	9.7 (4.39, 21.43)	1.93 (1.2, 3.01)
cg10514538	8	136602088	*KHDRBS3*	4.93 (1, 17.95)	7.95 (2.47, 25.56)	1.17 (0.84, 1.69)
cg12410530	22	32001086	*SFI1*	2.09 (1, 3.66)	2.82 (1.41, 5.66)	1.15 (0.98, 1.42)
cg13559022	12	54117783	*CALCOCO1*	1.8 (1, 2.95)	1.92 (1.01, 3.65)	1.25 (1, 1.65)
cg14096595	2	187420141	*ITGAV*	2.93 (1, 9.4)	3.34 (1.31, 8.49)	1.43 (0.97, 2.18)
cg17178502	8	17929088	*ASAH1*	5.5 (1, 15.28)	7.29 (2.51, 21.13)	1.49 (0.97, 2.43)
cg19965693	2	163175743	*IFIH1*	0.27 (0.12, 1)	0.24 (0.11, 0.54)	0.51 (0.31, 0.81)
cg22998476	10	74058092	*DDIT4*	0.41 (0.17, 1)	0.32 (0.15, 0.7)	0.7 (0.43, 1.14)

Abbreviations: ISIS, Iterative Sure Independence Screening; LASSO, Least Absolute Shrinkage and Selection Operator.

Table A11: Hazard ratios (95 % CI) for the CpGs selected by ISIS - SCAD comparing percentile 90th vs 10th for lung cancer and comparison with Cox regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	Cox	Bayes
cg00067702	5	166721765	*TENM2*	1.2 (1, 5.97)	3.33 (1.24, 8.99)	1.36 (0.84, 2.2)
cg00837619	14	21077925	*RNASE12*	2.93 (1, 9.23)	10.81 (4.04, 28.9)	1.63 (0.99, 2.76)
cg01309343	2	28668518	*FOSL2*	4.71 (1, 11.39)	4.35 (1.8, 10.55)	1.31 (0.85, 1.99)
cg02915225	10	93812043	*CPEB3*	6.58 (1.04, 25.52)	11.93 (4.59, 31)	1.71 (1.09, 2.77)
cg05926943	16	50100517	*HEATR3*	1.08 (1, 5.54)	3.33 (1.47, 7.51)	1.18 (0.74, 1.89)
cg06252810	18	77378261	*CTDP1*	1.73 (1, 4.79)	7.12 (1.71, 29.74)	1.66 (1.07, 2.75)
cg06318011	1	36567283	*COL8A2*	1.1 (1, 1.27)	2.19 (1.07, 4.48)	1.13 (0.95, 1.39)
cg06342317	2	121105259	*INHBB*	1.1 (1, 4.11)	4.49 (1.84, 10.92)	1.24 (0.82, 1.83)
cg06647068	12	104853274	*CHST11*	0.23 (0.06, 1)	0.15 (0.06, 0.39)	0.49 (0.29, 0.81)
cg07888917	2	12108682	*MIR4262*	3.54 (1, 8.85)	7.23 (2.8, 18.69)	1.62 (1.01, 2.66)
cg08162948	6	32374184	*BTNL2*	1.11 (1, 1.41)	2.22 (1.23, 4)	1.14 (0.93, 1.44)
cg12666727	1	42128487	*HIVEP3*	4.01 (1, 12.12)	8.94 (3.88, 20.6)	1.99 (1.25, 3.25)
cg13559022	12	54117783	*CALCOCO1*	1.22 (1, 1.6)	3.59 (1.67, 7.71)	1.28 (1.01, 1.67)
cg15997319	12	123778445	*SBNO1*	1.3 (1, 1.89)	2.62 (1.06, 6.5)	1.34 (0.98, 1.88)
cg16546976	3	171618566	*TMEM212*	7.75 (1, 21.83)	27.98 (9.9, 79.12)	1.74 (1.1, 2.83)
cg17172877	11	12863965	*TEAD1*	2.92 (1, 6.12)	9.66 (4.31, 21.62)	1.55 (1.06, 2.26)
cg17373649	8	669578	*ERICH1*	1.08 (1, 1.16)	3.9 (1.49, 10.22)	1.07 (0.98, 1.17)
cg17697043	20	52224625	*ZNF217*	1.31 (1, 1.62)	7.71 (2.81, 21.12)	1.26 (1.04, 1.62)
cg18277467	4	2180030	*POLN*	1.04 (1, 1.09)	1.98 (1.03, 3.81)	1.09 (0.98, 1.24)
cg19832312	10	74855378	*P4HA1*	2.78 (1, 5.74)	8.55 (3.9, 18.72)	1.67 (1.1, 2.53)
cg22660578	17	35294029	*LHX1*	3.46 (1, 10.65)	9.74 (4.02, 23.61)	1.82 (1.15, 2.99)
cg24377437	6	142047924	*NMBR*	1.11 (1, 2.47)	4.09 (1.71, 9.8)	1.19 (0.88, 1.64)
cg24650120	7	150096722	*ZNF775*	1.52 (1, 5.18)	5.42 (1.57, 18.7)	1.61 (1.01, 2.72)
cg25695116	22	37436981	*KCTD17*	0.11 (0.03, 1)	0.07 (0.02, 0.2)	0.41 (0.23, 0.69)

Abbreviations: ISIS, Iterative Sure Independence Screening; SCAD, Smoothly Clipped Absolute Deviation.

Table A12: Hazard ratios (95 % CI) for the CpGs selected by ISIS - MCP comparing percentile 90th vs 10th for lung cancer and comparison with Cox regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	Cox	Bayes
cg00926226	8	33462191	*DUSP26*	1.63 (1, 12.31)	5.66 (2.54, 12.65)	1.73 (1.16, 2.67)
cg02431184	6	88624913	*SPACA1*	3.5 (1, 78.04)	17.44 (5.96, 51.05)	2.01 (1.29, 3.41)
cg02468627	1	57043571	*PLPP3*	1.42 (1, 10.81)	4.98 (2.13, 11.65)	1.53 (1.12, 2.18)
cg04847932	17	40271031	*KAT2A*	1.19 (1, 27.51)	10 (4.09, 24.43)	2.1 (1.25, 3.55)
cg06647068	12	104853274	*CHST11*	0.51 (0.07, 1)	0.19 (0.08, 0.43)	0.49 (0.29, 0.8)
cg09018918	2	238607244	*LRRFIP1*	3.09 (1, 10.15)	3.7 (1.77, 7.74)	2.11 (1.31, 3.44)
cg09992204	15	99500303	*IGF1R*	1.28 (1, 9.64)	3.1 (1.26, 7.61)	1.56 (1.05, 2.45)
cg10662093	19	38747234	*PPP1R14A*	4.03 (1, 12.55)	6.52 (3.73, 11.41)	2.14 (1.45, 3.16)
cg14968926	5	64267561	*CWC27*	1.62 (1, 36.5)	8.47 (3.12, 22.98)	1.65 (1.06, 2.65)
cg15181928	8	2375845	*MYOM2*	1.92 (1, 19.44)	4.35 (1.83, 10.35)	1.7 (1.11, 2.69)
cg17372101	7	147500722	*CNTNAP2*	2.79 (1, 17.77)	7.25 (3.56, 14.76)	2.06 (1.28, 3.35)
cg21990700	12	7260776	*C1RL*	0.42 (0.05, 1)	0.15 (0.07, 0.35)	0.49 (0.29, 0.8)
cg25544931	19	12097624	*ZNF763*	3.25 (1, 31.61)	12.68 (5.47, 29.37)	2.27 (1.46, 3.59)
cg26248066	11	70303464	*SHANK2*	1 (1, 12.07)	5.24 (2.87, 9.6)	1.61 (1.08, 2.35)
cg26808749	10	121378908	*TIAL1*	1.1 (1, 2.41)	3.31 (1.45, 7.57)	1.27 (1.03, 1.63)
cg26928531	6	32407714	*HLA-DRA*	3.08 (1, 15.2)	9.92 (5.53, 17.79)	1.87 (1.34, 2.59)

Abbreviations: ISIS, Iterative Sure Independence Screening; MCP, Minimax Concave Penalty.

Table A13: Odds ratios (95 % CI) for the CpGs selected by ISIS - Aenet comparing percentile 90th vs 10th for diabetes and comparison with logistic regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	GLM	Bayes
cg00574958	11	68607622	*CPT1A*	0.9 (0.87, 0.92)	0.32 (0.2, 0.5)	0.36 (0.23, 0.55)
cg01411468	4	36283588	*DTHD1*	1.04 (1.01, 1.06)	1.72 (1.03, 2.86)	1.56 (0.99, 2.54)
cg04839616	16	86787604	*FOXL1*	1.05 (1.02, 1.08)	2.26 (1.51, 3.38)	2.0 (1.36, 2.95)
cg05291965	2	47382583	*STPG4*	1.07 (1.02, 1.09)	2.01 (1.32, 3.05)	1.88 (1.25, 2.82)
cg05746809	3	141131788	*ZBTB38*	0.93 (0.9, 0.95)	0.57 (0.38, 0.85)	0.59 (0.39, 0.87)
cg06002198	10	6187994	*PFKFB3*	0.93 (0.91, 0.96)	0.49 (0.27, 0.9)	0.57 (0.32, 0.98)
cg06865772	10	17725026	*STAM*	0.94 (0.92, 0.97)	0.6 (0.39, 0.92)	0.65 (0.43, 0.97)
cg08309687	21	35320596	*LINC00649*	0.91 (0.89, 0.95)	0.33 (0.19, 0.56)	0.39 (0.23, 0.64)
cg08686493	3	9670152	*MTMR14*	0.96 (0.94, 0.99)	0.47 (0.34, 0.67)	0.53 (0.38, 0.74)
cg11253148	11	104540147	*CASP12*	0.94 (0.93, 0.98)	0.54 (0.37, 0.81)	0.6 (0.4, 0.88)
cg12538681	11	34196039	*ABTB2*	1.05 (1.03, 1.09)	1.96 (1.35, 2.84)	1.81 (1.27, 2.59)
cg15340727	9	882695	*DMRT1*	0.95 (0.94, 0.99)	0.42 (0.27, 0.65)	0.46 (0.3, 0.71)
cg15910469	6	30804271	*DDR1*	1.04 (1.02, 1.08)	2.02 (1.41, 2.91)	1.86 (1.32, 2.65)
cg16340030	11	59554873	*STX3*	0.95 (0.93, 0.98)	0.32 (0.2, 0.51)	0.39 (0.25, 0.61)
cg16611584	17	19809078	*AKAP10*	1.08 (1.04, 1.11)	2.16 (1.39, 3.35)	1.95 (1.28, 2.99)
cg16740586	21	43655919	*ABCG1*	1.12 (1.09, 1.15)	2.09 (1.27, 3.43)	1.95 (1.22, 3.14)
cg17075888	7	95225339	*PDK4*	0.9 (0.87, 0.93)	0.44 (0.27, 0.7)	0.46 (0.29, 0.73)
cg19266329	1	145456128	*NBPF20*	0.92 (0.89, 0.95)	0.45 (0.28, 0.73)	0.52 (0.33, 0.83)
cg19466702	9	81749776	*TLE4*	0.93 (0.92, 0.97)	0.36 (0.22, 0.58)	0.41 (0.26, 0.66)
cg19693031	1	145441552	*TXNIP*	0.79 (0.76, 0.8)	0.03 (0.02, 0.06)	0.04 (0.02, 0.07)
cg21079041	7	151108012	*WDR86-AS1*	1.05 (1.01, 1.07)	2.71 (1.69, 4.35)	2.34 (1.49, 3.72)
cg22675726	18	3179889	*MYOM1*	1.08 (1.04, 1.1)	1.65 (1.07, 2.54)	1.55 (1.03, 2.35)
cg22757957	1	91440433	*ZNF644*	1.05 (1.02, 1.07)	2.76 (1.66, 4.59)	2.36 (1.46, 3.86)
cg25551219	22	50909865	*SBF1*	1.08 (1.05, 1.11)	2.63 (1.65, 4.17)	2.35 (1.51, 3.69)
cg26403843	5	158634085	*RNF145*	1.11 (1.07, 1.14)	2.48 (1.6, 3.85)	2.32 (1.52, 3.57)
cg27243685	21	43642366	*ABCG1*	1.06 (1.03, 1.09)	2.04 (1.29, 3.23)	1.74 (1.12, 2.72)

Abbreviations: ISIS, Iterative Sure Independence Screening; Aenet, adaptive elastic-net; GLM, Generalized linear models.

Table A14: Odds ratios (95 % CI) for the CpGs selected by ISIS - MSAenet comparing percentile 90th vs 10th for diabetes and comparison with logistic regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	GLM	Bayes
cg00574958	11	68607622	*CPT1A*	0.43 (0.29, 0.61)	0.43 (0.28, 0.66)	0.45 (0.29, 0.67)
cg00923870	20	37698529	*DHX35*	0.59 (0.34, 0.85)	0.29 (0.18, 0.47)	0.34 (0.21, 0.55)
cg01411468	4	36283588	*DTHD1*	1.71 (1.15, 2.63)	1.88 (1.11, 3.19)	1.79 (1.08, 2.97)
cg03803818	12	110637414	*IFT81*	1.52 (1.06, 2.2)	1.89 (1.17, 3.06)	1.79 (1.13, 2.88)
cg05028010	1	145437567	*NBPF20*	0.68 (0.43, 0.92)	0.67 (0.47, 0.96)	0.71 (0.5, 0.99)
cg05746809	3	141131788	*ZBTB38*	0.46 (0.32, 0.61)	0.43 (0.3, 0.64)	0.45 (0.31, 0.66)
cg05919202	4	5914761	*MIR378D1*	0.7 (0.41, 0.97)	0.3 (0.18, 0.52)	0.36 (0.21, 0.61)
cg06865772	10	17725026	*STAM*	0.62 (0.44, 0.82)	0.55 (0.37, 0.8)	0.56 (0.39, 0.82)
cg06998286	1	20473001	*PLA2G2F*	0.62 (0.37, 0.91)	0.37 (0.23, 0.59)	0.44 (0.28, 0.69)
cg08263236	7	129846037	*TMEM209*	0.61 (0.4, 0.84)	0.47 (0.31, 0.73)	0.54 (0.35, 0.82)
cg08309687	21	35320596	*LINC00649*	0.51 (0.3, 0.71)	0.26 (0.16, 0.44)	0.3 (0.18, 0.49)
cg10251538	3	108800886	*MORC1*	0.64 (0.45, 0.92)	0.63 (0.42, 0.93)	0.63 (0.43, 0.92)
cg10405605	10	6188149	*PFKFB3*	0.73 (0.52, 0.97)	0.57 (0.34, 0.95)	0.61 (0.37, 0.97)
cg10898277	20	24821220	*CST7*	1.76 (1.16, 2.87)	2.88 (1.72, 4.83)	2.47 (1.49, 4.11)
cg11126497	3	52558566	*NT5DC2*	0.54 (0.35, 0.74)	0.42 (0.27, 0.63)	0.46 (0.31, 0.7)
cg13898430	1	25292274	*RUNX3*	1.77 (1.17, 3.15)	2.8 (1.47, 5.31)	2.4 (1.29, 4.48)
cg15092039	6	148657019	*SASH1*	0.58 (0.3, 0.87)	0.28 (0.15, 0.5)	0.34 (0.19, 0.6)
cg16504526	9	73025362	*KLF9*	1.93 (1.4, 2.94)	2.44 (1.62, 3.68)	2.27 (1.53, 3.37)
cg16611584	17	19809078	*AKAP10*	2.0 (1.48, 3.08)	3.75 (2.46, 5.73)	3.33 (2.22, 5.08)
cg16740586	21	43655919	*ABCG1*	2.24 (1.58, 3.42)	2.72 (1.69, 4.38)	2.5 (1.57, 3.98)
cg19466702	9	81749776	*TLE4*	0.73 (0.49, 0.98)	0.55 (0.35, 0.88)	0.58 (0.37, 0.9)
cg19693031	1	145441552	*TXNIP*	0.08 (0.05, 0.12)	0.04 (0.02, 0.07)	0.05 (0.03, 0.08)
cg24488001	1	238644629	*LINC01139*	1.51 (1.13, 2.28)	2.52 (1.7, 3.73)	2.23 (1.53, 3.27)
cg25551219	22	50909865	*SBF1*	2.43 (1.71, 3.81)	3.81 (2.43, 5.98)	3.29 (2.13, 5.13)
cg26403843	5	158634085	*RNF145*	1.83 (1.33, 2.67)	3.16 (2.04, 4.9)	2.72 (1.79, 4.15)
cg27243685	21	43642366	*ABCG1*	1.72 (1.19, 2.54)	2.0 (1.28, 3.12)	1.83 (1.19, 2.83)

Abbreviations: ISIS, Iterative Sure Independence Screening; MSAenet, multi-step adaptive elastic-net; GLM, Generalized linear models.

Table A15: Odds ratios (95 % CI) for the CpGs selected by ISIS - enet comparing percentile 90th vs 10th for diabetes and comparison with logistic regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	GLM	Bayes
cg00574958	11	68607622	*CPT1A*	0.62 (0.51, 0.75)	0.45 (0.29, 0.7)	0.48 (0.32, 0.74)
cg01065697	1	3320738	*PRDM16*	1.35 (1.12, 1.61)	2.88 (1.75, 4.76)	2.33 (1.45, 3.81)
cg01268711	3	137799724	*DZIP1L*	0.7 (0.58, 0.85)	0.12 (0.05, 0.27)	0.21 (0.09, 0.47)
cg04244036	6	38207681	*BTBD9*	0.74 (0.6, 0.91)	0.61 (0.41, 0.92)	0.65 (0.44, 0.95)
cg05291965	2	47382583	*STPG4*	1.32 (1.07, 1.62)	1.65 (1.1, 2.48)	1.58 (1.07, 2.35)
cg05746809	3	141131788	*ZBTB38*	0.64 (0.53, 0.77)	0.41 (0.27, 0.62)	0.46 (0.31, 0.68)
cg08309687	21	35320596	*LINC00649*	0.57 (0.47, 0.7)	0.34 (0.2, 0.58)	0.4 (0.24, 0.67)
cg14725534	22	31516431	*INPP5J*	0.63 (0.53, 0.78)	0.11 (0.05, 0.24)	0.17 (0.08, 0.37)
cg16504526	9	73025362	*KLF9*	1.47 (1.21, 1.83)	1.77 (1.17, 2.69)	1.67 (1.12, 2.5)
cg16740586	21	43655919	*ABCG1*	1.62 (1.33, 1.99)	2.08 (1.28, 3.38)	1.96 (1.24, 3.1)
cg17075888	7	95225339	*PDK4*	0.61 (0.49, 0.76)	0.46 (0.29, 0.73)	0.49 (0.31, 0.76)
cg19264738	11	72925488	*P2RY2*	1.5 (1.25, 1.83)	2.86 (1.83, 4.44)	2.46 (1.61, 3.78)
cg19266329	1	145456128	*NBPF20*	0.61 (0.5, 0.74)	0.42 (0.26, 0.67)	0.48 (0.31, 0.75)
cg19466702	9	81749776	*TLE4*	0.69 (0.56, 0.84)	0.49 (0.3, 0.78)	0.53 (0.34, 0.83)
cg19693031	1	145441552	*TXNIP*	0.26 (0.22, 0.31)	0.06 (0.03, 0.09)	0.07 (0.04, 0.11)
cg25551219	22	50909865	*SBF1*	1.57 (1.28, 1.91)	2.46 (1.56, 3.87)	2.18 (1.42, 3.39)
cg26403843	5	158634085	*RNF145*	1.59 (1.28, 1.92)	2.06 (1.34, 3.19)	1.97 (1.29, 3.0)

Abbreviations: ISIS, Iterative Sure Independence Screening; enet, elastic-net; GLM, Generalized linear models.

Table A16: Odds ratios (95 % CI) for the CpGs selected by ISIS - LASSO comparing percentile 90th vs 10th for diabetes and comparison with logistic regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	GLM	Bayes
cg00574958	11	68607622	*CPT1A*	0.48 (0.33, 0.66)	0.35 (0.22, 0.57)	0.38 (0.24, 0.6)
cg04141459	3	183852329	*EIF2B5*	1.68 (1.22, 2.35)	2.43 (1.6, 3.68)	2.16 (1.45, 3.24)
cg04244036	6	38207681	*BTBD9*	0.6 (0.42, 0.87)	0.45 (0.29, 0.71)	0.5 (0.32, 0.76)
cg04422019	7	50611949	*DDC*	1.95 (1.4, 2.78)	2.96 (1.86, 4.69)	2.62 (1.68, 4.12)
cg04893683	5	178779526	*ADAMTS2*	0.61 (0.41, 0.94)	0.43 (0.27, 0.71)	0.48 (0.29, 0.76)
cg05241075	1	55776045	*MIR4422*	0.63 (0.39, 0.99)	0.34 (0.19, 0.62)	0.42 (0.24, 0.75)
cg05291965	2	47382583	*STPG4*	1.83 (1.29, 2.67)	2.73 (1.74, 4.28)	2.42 (1.58, 3.78)
cg05746809	3	141131788	*ZBTB38*	0.5 (0.34, 0.71)	0.33 (0.21, 0.51)	0.37 (0.24, 0.57)
cg05868469	15	93128777	*LINC00930*	1.96 (1.05, 3.72)	4.41 (1.98, 9.82)	3.3 (1.54, 7.23)
cg05878073	14	74766220	*ABCD4*	2.24 (1.27, 3.9)	5.46 (2.68, 11.13)	4.17 (2.11, 8.27)
cg08309687	21	35320596	*LINC00649*	0.45 (0.28, 0.68)	0.32 (0.18, 0.56)	0.35 (0.2, 0.6)
cg08896067	7	5867617	*ZNF815P*	0.55 (0.39, 0.77)	0.38 (0.25, 0.58)	0.43 (0.29, 0.65)
cg08972190	7	2138995	*MAD1L1*	1.95 (1.28, 2.98)	2.7 (1.55, 4.7)	2.52 (1.5, 4.32)
cg10322118	9	33173043	*B4GALT1*	1.63 (1.08, 2.49)	2.34 (1.4, 3.91)	2.08 (1.27, 3.43)
cg10933573	18	72212692	*CNDP1*	1.85 (1.3, 2.73)	3.32 (2.04, 5.39)	2.76 (1.74, 4.45)
cg11024682	17	17730094	*SREBF1*	1.79 (1.07, 3.02)	2.58 (1.35, 4.93)	2.26 (1.23, 4.23)
cg11406521	2	86226603	Unknown	1.61 (1.13, 2.27)	1.93 (1.21, 3.09)	1.85 (1.19, 2.89)
cg14734059	17	12698291	*ARHGAP44*	0.58 (0.41, 0.83)	0.39 (0.25, 0.62)	0.44 (0.28, 0.69)
cg15643381	3	58510065	*ACOX2*	0.5 (0.27, 0.91)	0.15 (0.07, 0.35)	0.23 (0.1, 0.51)
cg16611584	17	19809078	*AKAP10*	1.55 (1.07, 2.27)	1.84 (1.14, 2.95)	1.74 (1.11, 2.75)
cg16615151	3	111409324	*PLCXD2*	1.9 (1.25, 2.9)	2.19 (1.3, 3.71)	2.11 (1.28, 3.52)
cg16740586	21	43655919	*ABCG1*	2.17 (1.45, 3.52)	2.77 (1.65, 4.67)	2.62 (1.6, 4.35)
cg17075888	7	95225339	*PDK4*	0.5 (0.33, 0.75)	0.37 (0.22, 0.61)	0.4 (0.25, 0.65)
cg19266329	1	145456128	*NBPF20*	0.62 (0.43, 0.91)	0.43 (0.26, 0.69)	0.48 (0.3, 0.77)
cg19466702	9	81749776	*TLE4*	0.65 (0.42, 0.96)	0.45 (0.27, 0.76)	0.51 (0.31, 0.84)
cg19693031	1	145441552	*TXNIP*	0.07 (0.04, 0.1)	0.03 (0.02, 0.06)	0.04 (0.02, 0.07)
cg22675726	18	3179889	*MYOM1*	1.68 (1.19, 2.42)	2.51 (1.62, 3.9)	2.21 (1.45, 3.41)
cg22757957	1	91440433	*ZNF644*	1.65 (1.05, 2.63)	2.74 (1.54, 4.86)	2.31 (1.34, 4.03)
cg23065813	2	43617024	*THADA*	1.75 (1.18, 2.74)	2.52 (1.51, 4.18)	2.23 (1.38, 3.64)
cg26955383	10	105218660	*CALHM1*	1.45 (1.04, 2.04)	1.7 (1.12, 2.58)	1.63 (1.1, 2.44)
cg27243685	21	43642366	*ABCG1*	1.57 (1.08, 2.21)	2.25 (1.4, 3.61)	1.97 (1.25, 3.12)
cg27531842	2	131672685	*ARHGEF4*	1.77 (1.21, 2.52)	2.66 (1.66, 4.25)	2.36 (1.52, 3.73)

Abbreviations: ISIS, Iterative Sure Independence Screening; LASSO, Least Absolute Shrinkage and Selection Operator; GLM, Generalized linear models.

Table A17: Odds ratios (95 % CI) for the CpGs selected by ISIS - SCAD comparing percentile 90th vs 10th for diabetes and comparison with logistic regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	GLM	Bayes
cg00574958	11	68607622	*CPT1A*	0.4 (0.22, 0.88)	0.39 (0.24, 0.62)	0.42 (0.26, 0.66)
cg02592976	7	140746692	*TMEM178B*	0.21 (0.09, 0.85)	0.17 (0.09, 0.34)	0.23 (0.11, 0.44)
cg05291965	2	47382583	*STPG4*	2.26 (1.05, 3.86)	2.15 (1.39, 3.34)	2.01 (1.32, 3.1)
cg05746809	3	141131788	*ZBTB38*	0.29 (0.18, 0.45)	0.3 (0.2, 0.46)	0.33 (0.22, 0.49)
cg05919202	4	5914761	*MIR378D1*	0.25 (0.12, 0.42)	0.23 (0.13, 0.41)	0.28 (0.15, 0.49)
cg08309687	21	35320596	*LINC00649*	0.28 (0.14, 0.73)	0.27 (0.16, 0.47)	0.3 (0.18, 0.52)
cg08594547	4	178426921	*AGA*	2.71 (1.69, 4.72)	2.9 (1.93, 4.37)	2.57 (1.74, 3.86)
cg11697045	16	13416540	*SHISA9*	3.29 (1.94, 5.94)	3.29 (2.11, 5.12)	3.0 (1.96, 4.61)
cg12603039	1	179678692	*FAM163A*	0.22 (0.11, 1)	0.2 (0.1, 0.4)	0.28 (0.14, 0.53)
cg15340727	9	882695	*DMRT1*	0.36 (0.21, 0.97)	0.4 (0.24, 0.65)	0.43 (0.27, 0.69)
cg17075888	7	95225339	*PDK4*	0.33 (0.18, 0.7)	0.35 (0.21, 0.57)	0.37 (0.23, 0.59)
cg18572606	14	74691002	*VSX2*	0.18 (0.06, 0.97)	0.14 (0.07, 0.3)	0.19 (0.09, 0.39)
cg19520763	3	38534710	*ACVR2B*	6.01 (2.88, 13.23)	5.23 (2.83, 9.67)	4.45 (2.48, 8.11)
cg19693031	1	145441552	*TXNIP*	0.03 (0.01, 0.05)	0.03 (0.02, 0.05)	0.04 (0.02, 0.06)
cg22757957	1	91440433	*ZNF644*	3.31 (1.15, 6.01)	2.95 (1.71, 5.07)	2.62 (1.56, 4.49)
cg25551219	22	50909865	*SBF1*	3.03 (1.73, 5.22)	3.02 (1.88, 4.85)	2.8 (1.78, 4.44)
cg26403843	5	158634085	*RNF145*	3.19 (1.92, 5.77)	3.23 (2.02, 5.17)	2.94 (1.88, 4.68)

Abbreviations: ISIS, Iterative Sure Independence Screening; SCAD, Smoothly Clipped Absolute Deviation; GLM, Generalized linear models.

Table A18: Odds ratios (95 % CI) for the CpGs selected by ISIS - MCP comparing percentile 90th vs 10th for diabetes and comparison with logistic regression and Bayesian elastic-net.

CpG	chr	pos	Gene	ISIS	GLM	Bayes
cg00506811	2	235860443	*SH3BP4*	0.46 (0.28, 0.87)	0.44 (0.31, 0.62)	0.49 (0.34, 0.68)
cg00574958	11	68607622	*CPT1A*	0.22 (0.11, 0.39)	0.22 (0.13, 0.37)	0.25 (0.15, 0.42)
cg00976328	2	208689522	*PLEKHM3*	2.83 (1.25, 5.24)	2.78 (1.67, 4.64)	2.42 (1.49, 3.97)
cg05291965	2	47382583	*STPG4*	3.39 (1.8, 6.47)	3.23 (1.96, 5.31)	2.84 (1.77, 4.62)
cg05746809	3	141131788	*ZBTB38*	0.29 (0.14, 0.59)	0.28 (0.17, 0.46)	0.31 (0.19, 0.5)
cg08594547	4	178426921	*AGA*	2.62 (1.31, 4.45)	2.59 (1.67, 4.01)	2.35 (1.55, 3.62)
cg08896067	7	5867617	*ZNF815P*	0.27 (0.15, 0.43)	0.26 (0.16, 0.41)	0.3 (0.19, 0.46)
cg10933573	18	72212692	*CNDP1*	2.46 (1.02, 4.64)	2.44 (1.49, 4.01)	2.19 (1.37, 3.54)
cg12538681	11	34196039	*ABTB2*	2.77 (1.36, 5.03)	2.74 (1.75, 4.28)	2.46 (1.6, 3.77)
cg13259095	8	1455354	*DLGAP2*	3.51 (1.88, 6.8)	3.51 (2.07, 5.95)	3.02 (1.83, 5.05)
cg13718666	12	101800969	*ARL1*	2.44 (1.05, 4.34)	2.49 (1.57, 3.95)	2.23 (1.44, 3.47)
cg14734059	17	12698291	*ARHGAP44*	0.27 (0.14, 0.42)	0.26 (0.16, 0.42)	0.3 (0.18, 0.48)
cg15340727	9	882695	*DMRT1*	0.29 (0.15, 0.56)	0.3 (0.18, 0.5)	0.33 (0.2, 0.54)
cg15630743	10	134863617	*ADGRA1*	0.18 (0.08, 0.37)	0.18 (0.09, 0.34)	0.22 (0.12, 0.42)
cg16611584	17	19809078	*AKAP10*	2.62 (1.1, 4.98)	2.55 (1.57, 4.15)	2.36 (1.48, 3.79)
cg17075888	7	95225339	*PDK4*	0.29 (0.14, 0.53)	0.3 (0.18, 0.5)	0.31 (0.19, 0.51)
cg19264738	11	72925488	*P2RY2*	3.69 (2.37, 7.72)	3.78 (2.26, 6.32)	3.43 (2.11, 5.67)
cg19693031	1	145441552	*TXNIP*	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)
cg22030766	21	34608899	*IFNAR2*	0.41 (0.24, 0.9)	0.4 (0.25, 0.62)	0.46 (0.29, 0.7)
cg22033732	10	76975755	*VDAC2*	2.46 (1.1, 4.35)	2.47 (1.57, 3.87)	2.23 (1.45, 3.45)
cg22939839	2	46586862	*EPAS1*	2.53 (1.02, 4.52)	2.6 (1.62, 4.15)	2.32 (1.49, 3.68)
cg25386579	6	139571728	*TXLNB*	3.68 (1.99, 7.59)	3.64 (2.09, 6.36)	3.17 (1.87, 5.44)
cg25551219	22	50909865	*SBF1*	3.62 (1.76, 7.15)	3.71 (2.19, 6.27)	3.32 (2.02, 5.54)
cg26403843	5	158634085	*RNF145*	2.84 (1.32, 5.79)	2.76 (1.66, 4.56)	2.56 (1.57, 4.18)
cg26940541	8	134584367	*ST3GAL1*	3.48 (1.81, 7.24)	3.55 (2.08, 6.07)	3 (1.79, 5.1)

Abbreviations: ISIS, Iterative Sure Independence Screening; MCP, Minimax Concave Penalty; GLM, Generalized linear models.

Figure A1: Overlap of significantly enriched pathways for genes annotated to the identified BMI-DMPs, separately for each of the specific methods, and, also, for the union set of genes annotated to BMI-DMPs across all methods.

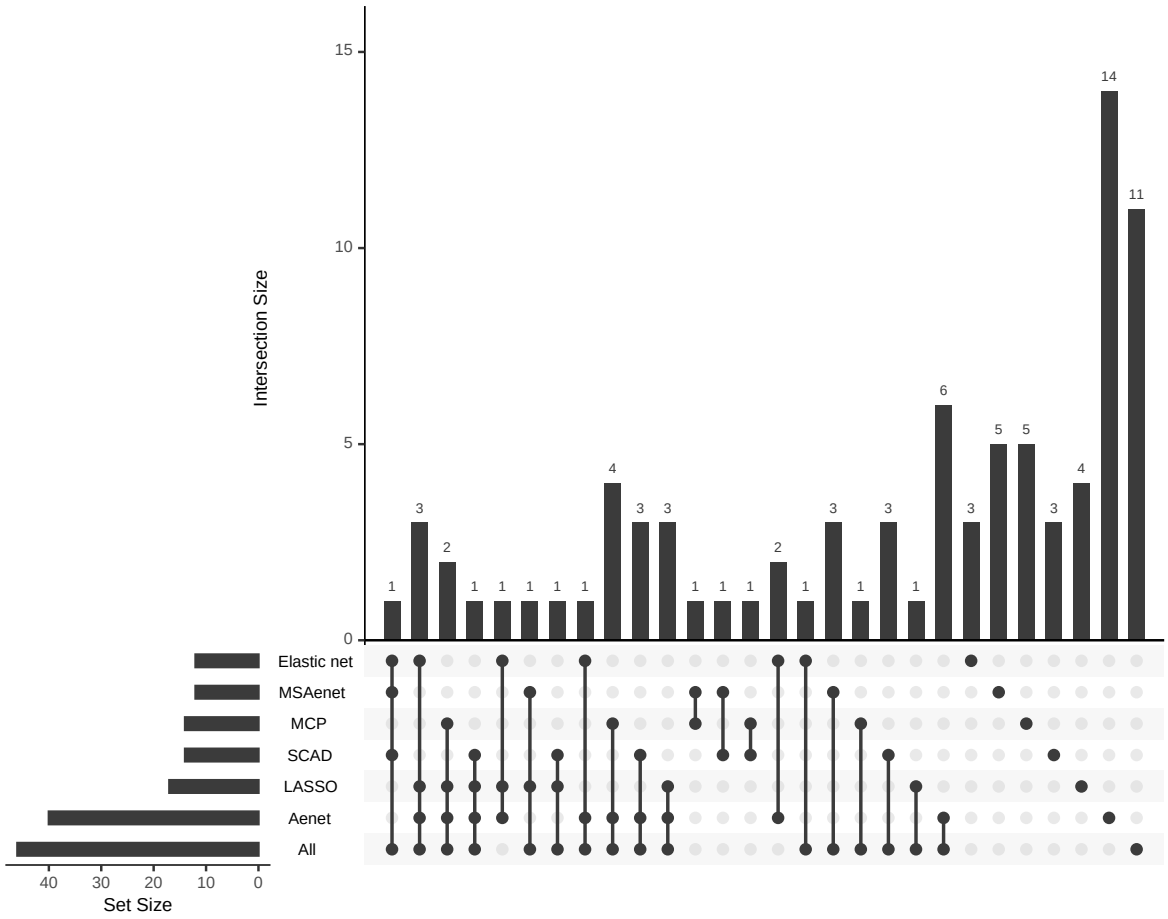
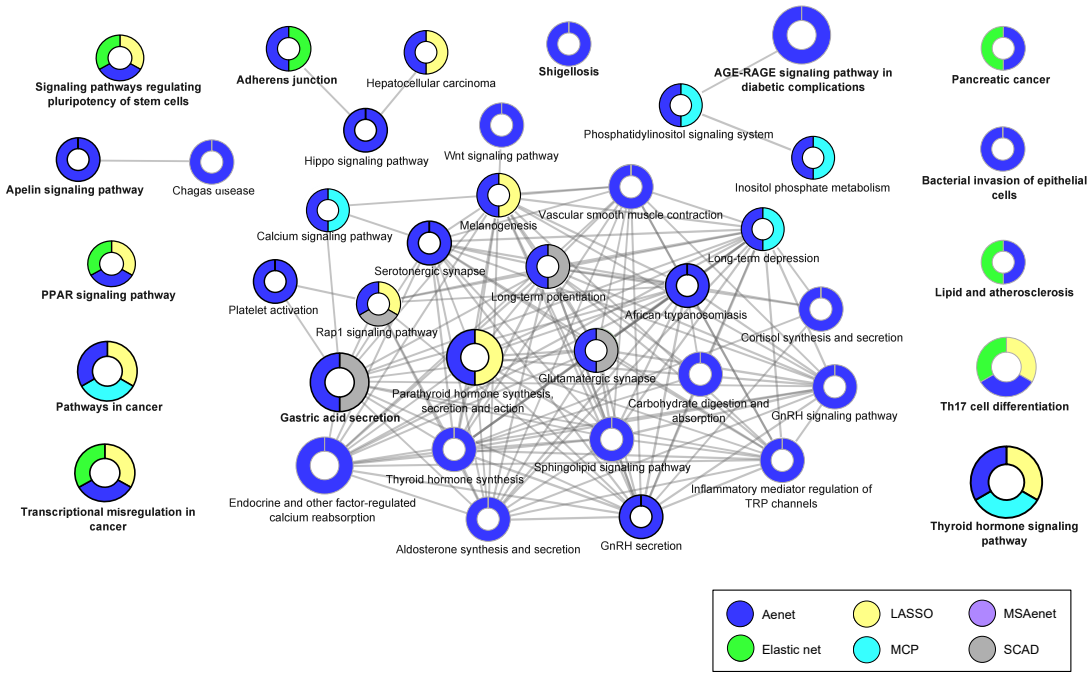
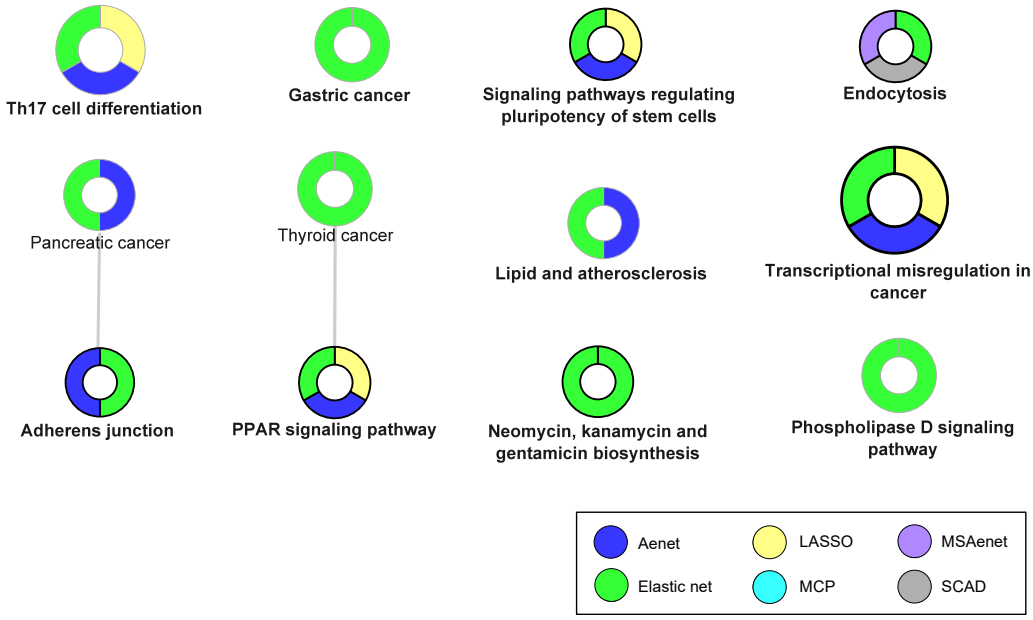


Figure A2: Network of the enriched pathways for BMI out of the genes annotated to identified differentially methylated positions for ISIS-Aenet.



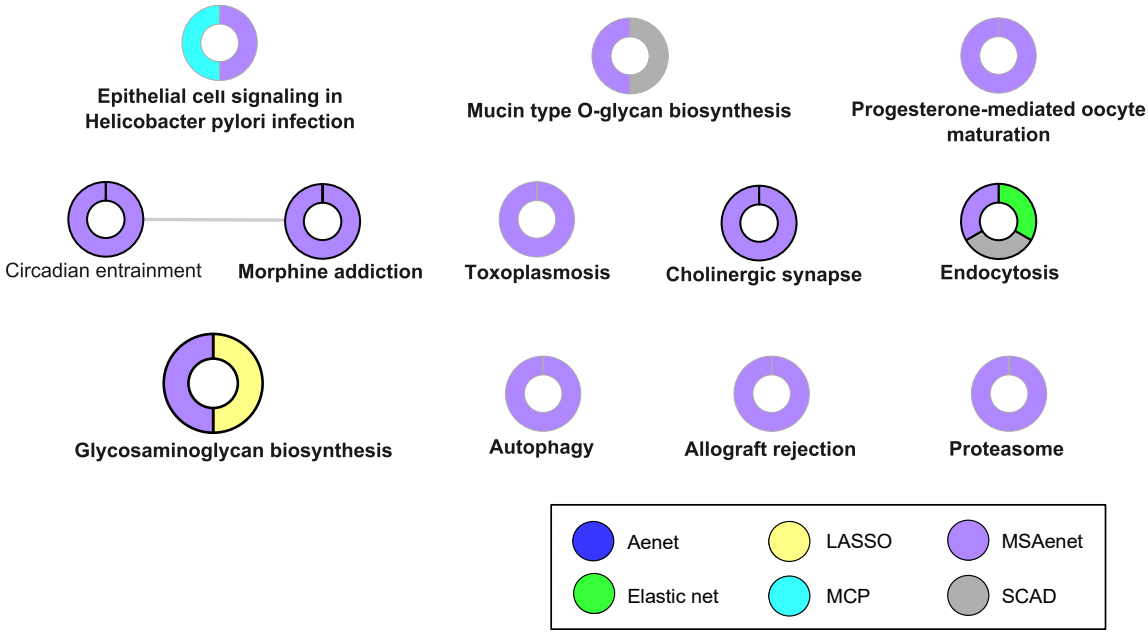
KEGG pathways are represented as nodes and the node size represents the term enrichment significance. The size of the slices represents the proportion of the genes that contribute to the metabolic pathway for each method.

Figure A3: Network of the enriched pathways for BMI out of the genes annotated to identified differentially methylated positions for ISIS-enet.



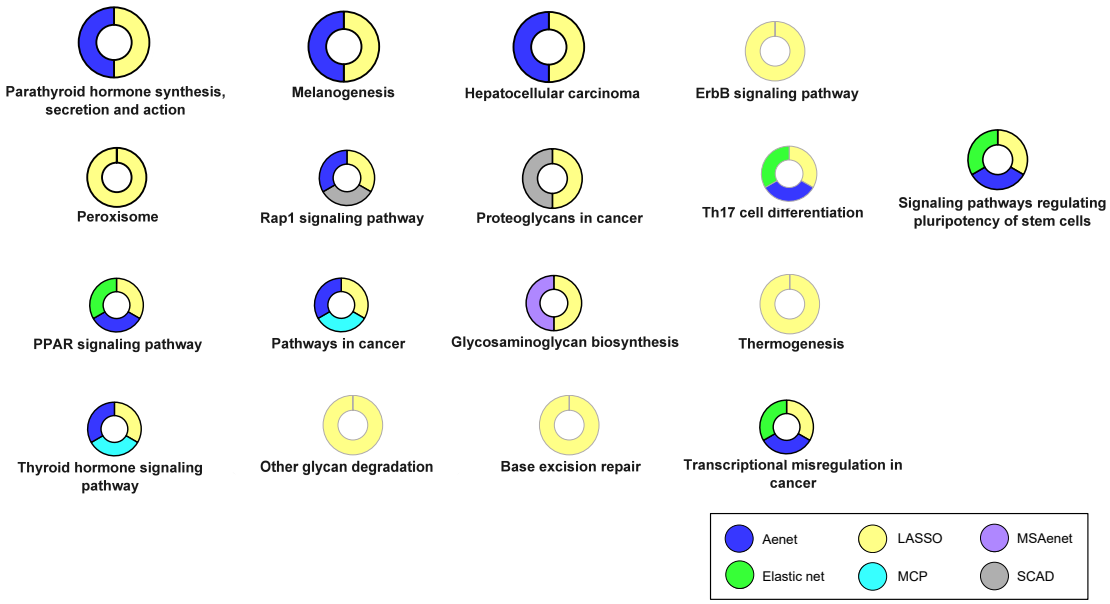
KEGG pathways are represented as nodes and the node size represents the term enrichment significance. The size of the slices represents the proportion of the genes that contribute to the metabolic pathway for each method.

Figure A4: Network of the enriched pathways for BMI out of the genes annotated to identified differentially methylated positions for ISIS-MSAenet.



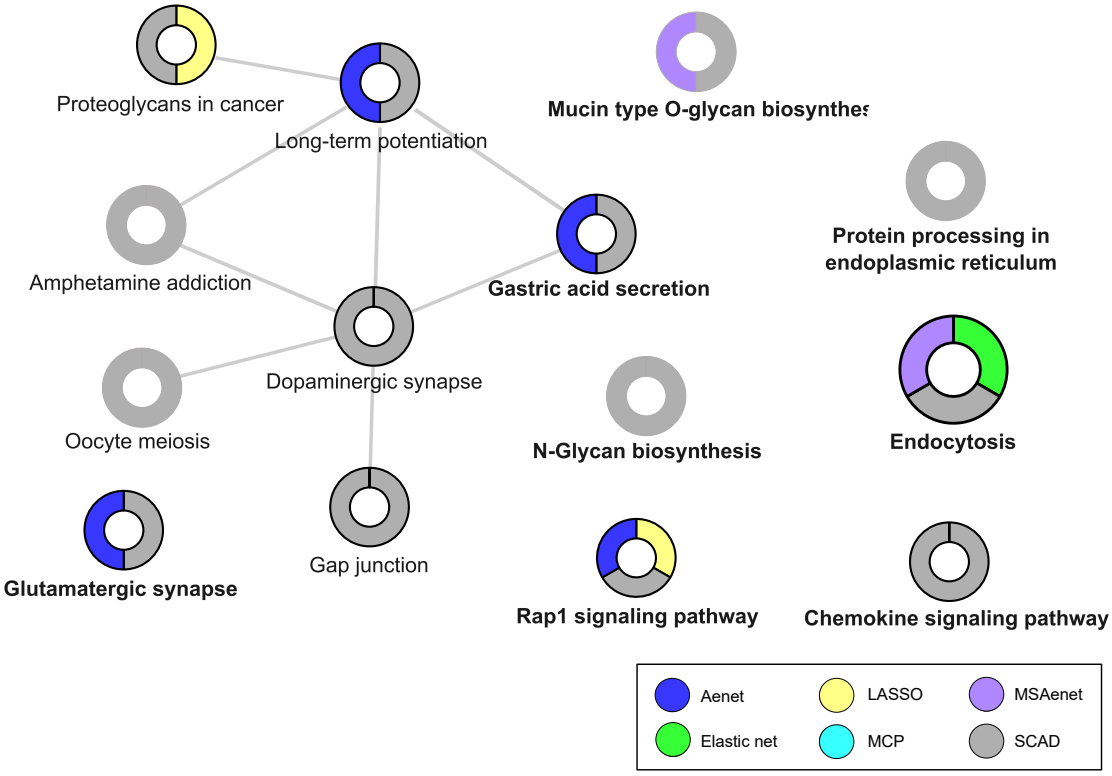
KEGG pathways are represented as nodes and the node size represents the term enrichment significance. The size of the slices represents the proportion of the genes that contribute to the metabolic pathway for each method.

Figure A5: Network of the enriched pathways for BMI out of the genes annotated to identified differentially methylated positions for ISIS-LASSO.



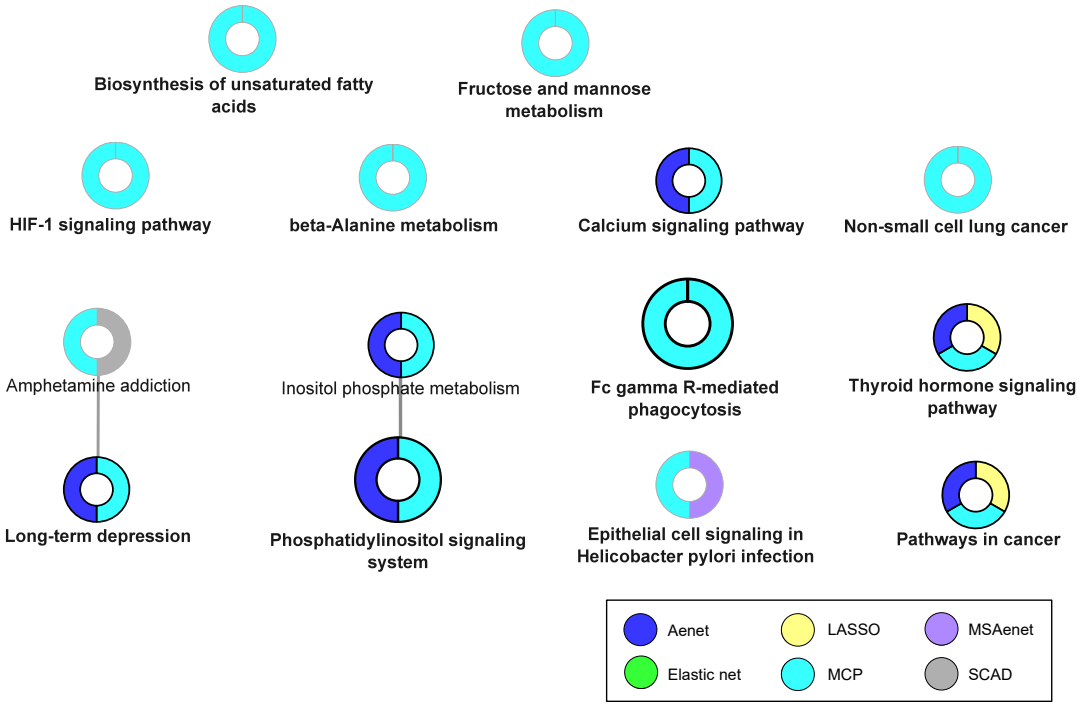
KEGG pathways are represented as nodes and the node size represents the term enrichment significance. The size of the slices represents the proportion of the genes that contribute to the metabolic pathway for each method.

Figure A6: Network of the enriched pathways for BMI out of the genes annotated to identified differentially methylated positions for ISIS-SCAD.



KEGG pathways are represented as nodes and the node size represents the term enrichment significance. The size of the slices represents the proportion of the genes that contribute to the metabolic pathway for each method.

Figure A7: Network of the enriched pathways for BMI out of the genes annotated to identified differentially methylated positions for ISIS-MCP.



KEGG pathways are represented as nodes and the node size represents the term enrichment significance. The size of the slices represents the proportion of the genes that contribute to the metabolic pathway for each method.

Appendix B: Supplementary Tables for section 3.2.2

Table B1: Hazard ratios (95 % CI) of CpGs selected by ISIS-Aenet as associated with CVD incidence comparing percentile 90th vs 10th.

CpG	Chr	Position	Gene	HR (95 % CIs)
cg00168459	10	52134715	*SGMS1*	2.6 (1.48, 4.87)
cg00451635	16	10675030	*EMP2*	1.08 (0.98, 1.33)
cg00841849	2	8683604	*ID2*	0.64 (0.45, 0.9)
cg01120308	11	85780971	*PICALM*	1.48 (1.0, 2.13)
cg01127300	22	38614796	*TMEM184B*	0.91 (0.73, 1.0)
cg01270753	9	101944336	*TGFBR1*	0.75 (0.57, 0.97)
cg01695954	2	10262019	*RRM2*	1.1 (0.78, 1.64)
cg02862467	1	19407897	*UBR4*	1.29 (1.0, 2.24)
cg03061719	12	56687956	*CS*	0.71 (0.37, 1.0)
cg03258257	16	89786032	*VPS9D1*	3.07 (1.79, 5.66)
cg03362418	22	50965563	*TYMP*	0.81 (0.56, 1.0)
cg03725309	1	109757585	*SARS*	0.94 (0.63, 1.18)
cg03877179	2	220468716	*STK11IP*	1.04 (0.79, 1.78)
cg05168229	13	45390049	*LINC00330*	0.86 (0.59, 1.0)
cg05438378	15	67383736	*SMAD3*	0.88 (0.66, 1.0)
cg05487345	2	237657127	*ACKR3*	1.44 (1.0, 2.3)
cg05709770	3	59703569	*FHIT*	1.34 (1.0, 1.83)
cg06084585	13	50797125	*DLEU1*	0.52 (0.29, 0.81)
cg06090426	3	129712103	*TRH*	1.03 (0.81, 1.29)
cg06639320	2	106015739	*FHL2*	1.05 (0.88, 1.48)
cg06647068	12	104853274	*CHST11*	0.97 (0.75, 1.07)
cg06668829	8	144948781	*EPPK1*	1.49 (1.21, 1.9)
cg06779020	14	70183064	*SUSD6*	0.59 (0.36, 0.98)

cg06970472	4	40910981	*APBB2*	1.07 (0.94, 1.45)
cg08232510	20	52527051	*BCAS1*	1.52 (1.0, 2.79)
cg08374298	2	161431244	*RBMS1*	1.27 (1.0, 2.52)
cg08919791	1	174135405	*RABGAP1L*	1.36 (1.0, 1.94)
cg09110394	6	25301880	*CARMIL1*	1.18 (1.0, 1.55)
cg09883673	9	124160814	*STOM*	0.61 (0.35, 1.0)
cg10465805	8	134911885	*ST3GAL1*	1.18 (1.0, 1.78)
cg10972897	1	165859714	*UCK2*	0.9 (0.51, 1.0)
cg11499323	11	45114907	*PRDM11*	1.18 (1.0, 1.58)
cg11961845	7	129008179	*AHCYL2*	1.54 (1.04, 2.29)
cg12747277	19	33754366	*CEBPA*	1.05 (0.97, 1.72)
cg13092901	22	50965373	*TYMP*	0.87 (0.65, 1.0)
cg13251119	1	110302560	*EPS8L3*	0.57 (0.33, 1.0)
cg13984040	12	125258948	*SCARB1*	1.01 (0.95, 1.6)
cg14066163	17	17311866	-	0.75 (0.44, 1.0)
cg14625801	1	32216877	*ADGRB2*	1.03 (0.94, 1.4)
cg15551201	5	67948405	*PIK3R1*	1.2 (1.0, 2.2)
cg16355382	6	7009231	*RREB1*	1.22 (1.0, 1.64)
cg16410715	6	139090474	*GVQW2*	0.61 (0.34, 1.0)
cg17075045	18	10305098	*APCDD1*	1.4 (0.89, 2.18)
cg17677968	12	121100923	*CABP1*	1.34 (1.0, 1.92)
cg18130370	22	37270153	*NCF4*	0.89 (0.58, 1.0)
cg18442362	7	44677772	*OGDH*	0.71 (0.5, 0.99)
cg18618815	17	48275324	*COL1A1*	0.84 (0.64, 1.0)
cg18638581	2	75059602	*HK2*	0.7 (0.4, 1.0)
cg20070323	15	72520350	*PKM*	0.97 (0.68, 1.05)
cg20241876	6	32180045	*NOTCH4*	0.71 (0.42, 1.0)
cg22399484	2	234153621	*ATG16L1*	0.65 (0.36, 1.0)
cg22454769	2	106015767	*FHL2*	1.22 (1.0, 1.52)
cg22640868	17	26661374	*IFT20*	1.67 (1.13, 2.55)
cg23018612	1	155960280	*ARHGEF2*	0.51 (0.32, 0.89)
cg23027596	9	138854008	*UBAC1*	0.77 (0.59, 1.0)
cg23772226	22	29225779	*XBP1*	0.58 (0.32, 1.0)
cg23877117	21	27298484	*APP*	0.93 (0.53, 1.28)
cg24223075	11	119137279	*CBL*	0.98 (0.7, 1.05)
cg24507742	19	47288244	*SLC1A5*	0.9 (0.6, 1.0)
cg24924295	17	999104	*ABR*	1.3 (1.0, 2.14)
cg25371036	11	94500749	*AMOTL1*	0.8 (0.61, 1.0)
cg25375916	3	155570275	*SLC33A1*	1.12 (1.0, 1.75)
cg25444339	7	75194698	*HIP1*	0.55 (0.28, 1.0)
cg25452273	15	75335524	*PPCDC*	1.17 (1.0, 1.6)
cg25553730	7	101874141	*CUX1*	0.94 (0.5, 1.39)
cg26269286	14	103848982	*MARK3*	1.14 (0.98, 1.75)
cg26292691	10	111654561	*XPNPEP1*	0.88 (0.56, 1.0)
cg26855629	5	131353815	*ACSL6*	1.53 (1.0, 2.55)
cg27076680	2	219336877	*USP37*	1.0 (0.79, 1.68)
cg27260684	16	85063742	*KIAA0513*	1.48 (1.0, 2.63)

Table B2: Hazard ratios (95 % CIs) of CpGs selected by ISIS-Aenet as associated with CVD mortality comparing percentile 90th vs 10th.

CpG	Chr	Position	Gene	HR (95 % CIs)
cg00218914	3	129146731	*EFCAB12*	3.18 (1, 7.06)
cg00451635	16	10675030	*EMP2*	0.68 (0.46, 1)
cg00724332	12	129342261	*GLT1D1*	0.29 (0.11, 1)
cg00841849	2	8683604	*ID2*	0.63 (0.32, 1.01)
cg00918877	14	23584303	*CEBPE*	0.3 (0.11, 0.77)
cg00950718	1	159859904	*CFAP45*	1.03 (0.51, 2.32)
cg01034993	15	68992751	*CORO2B*	0.85 (0.47, 1.73)
cg01309511	10	125799171	*CHST15*	1.5 (0.87, 2.83)
cg01695954	2	10262019	*RRM2*	1.69 (0.79, 3.17)
cg01858712	16	12156525	*SNX29*	0.3 (0.12, 1)
cg01990133	19	51898727	*C19orf84*	1.43 (0.96, 3.28)
cg02182504	17	44292898	*KANSL1*	0.35 (0.16, 1)
cg02848648	14	100807801	*WARS1*	1.57 (0.89, 2.85)
cg03026982	11	19953699	*NAV2*	1.2 (0.8, 1.73)
cg03362418	22	50965563	*TYMP*	0.51 (0.29, 0.94)
cg03794433	15	90548061	*ZNF710*	1.61 (1, 4.49)
cg04009045	2	74804817	*M1AP*	1.71 (0.92, 3.65)
cg04415535	3	46034485	*FYCO1*	3.96 (1.26, 8.13)
cg04987734	14	103415873	*CDC42BPB*	1.55 (1, 2.22)
cg05569131	6	36665620	*RAB44*	4.72 (1, 14.64)
cg06090426	3	129712103	*TRH*	1.06 (0.74, 1.64)
cg06285727	11	72524028	*ATG16L2*	0.77 (0.38, 1.18)
cg06970472	4	40910981	*APBB2*	0.69 (0.43, 1.05)
cg06983026	7	2344903	*SNX8*	0.64 (0.41, 1)
cg07009821	16	17451418	*XYLT1*	3.37 (1, 8.86)
cg07235066	15	76013321	*ODF3L1*	0.42 (0.18, 1)
cg08317046	9	101010744	*TBC1D2*	2.45 (1, 8.53)
cg09047229	16	11330882	*RMI2*	2.92 (1, 7.81)
cg09340693	9	112890072	*PALM2AKAP2*	0.23 (0.09, 0.91)
cg09608814	16	17448390	*XYLT1*	0.41 (0.14, 1)
cg09849237	12	52404250	*GRASP*	1.39 (0.88, 2.61)
cg10809358	2	10530799	*HPCAL1*	2.15 (1, 4.06)
cg11009736	2	119699682	*MARCO*	2.25 (1, 4.44)
cg11012616	14	35835542	*NFKBIA*	0.26 (0.13, 0.81)
cg11197422	5	149319802	*PDE6A*	0.51 (0.22, 1.11)
cg11496404	6	88428902	*AKIRIN2*	0.3 (0.13, 0.92)
cg11697092	19	3365736	*NFIC*	4.72 (1.2, 14.68)
cg11964145	2	190523706	*ASNSD1*	1.97 (1.05, 3.25)
cg12484135	3	10270544	*IRAK2*	0.65 (0.38, 0.98)
cg12592772	12	123166661	*HCAR2*	3.61 (1.1, 13.4)
cg12668854	17	944076	*ABR*	0.46 (0.17, 1)
cg12737520	7	149116704	*ZNF777*	1.39 (1.16, 1.84)
cg13251119	1	110302560	*EPS8L3*	0.18 (0.06, 0.63)

cg13509638	1	12240593	*TNFRSF1B*	1.3 (0.86, 3.8)
cg13681496	9	95817516	*SUSD3*	2.68 (1, 8.03)
cg13708436	20	43292835	*LINC01260*	0.5 (0.22, 1)
cg14034677	17	38637569	*TNS4*	0.84 (0.43, 1.3)
cg14066163	17	17311866	-	0.67 (0.31, 1.17)
cg15865606	17	75853127	*LINC01973*	1.99 (1, 3.9)
cg16571794	4	7648627	*SORCS2*	1.53 (0.87, 2.62)
cg16783053	10	112261195	*DUSP5*	1.17 (0.63, 2.81)
cg17166812	1	161169574	*NDUFS2*	2.88 (1.41, 5.03)
cg17470213	7	134836467	*CYREN*	0.34 (0.13, 1)
cg17762936	5	135159685	*SLC25A48*	0.67 (0.46, 1)
cg18130370	22	37270153	*NCF4*	0.44 (0.19, 0.99)
cg18339493	17	78681826	*RPTOR*	0.23 (0.09, 0.74)
cg19275653	2	175532338	*WIPF1*	0.37 (0.11, 1)
cg20131875	7	72759068	*FKBP6*	1.47 (0.7, 2.89)
cg20363271	8	55181528	*MRPL15*	1.34 (0.92, 2.15)
cg21122051	10	97515319	*ENTPD1*	0.64 (0.25, 1)
cg21204620	15	77281418	*PSTPIP1*	3.28 (1.31, 6.75)
cg21990700	12	7260776	*C1RL*	0.81 (0.53, 1.16)
cg22530144	2	46300869	*PRKCE*	0.56 (0.34, 0.88)
cg24104249	1	247600998	*NLRP3*	0.34 (0.17, 1)
cg24254842	1	42193353	*HIVEP3*	1.54 (0.87, 2.92)
cg24487151	3	73025772	*GXYLT2*	1.8 (1, 3.29)
cg25371036	11	94500749	*AMOTL1*	0.42 (0.27, 0.73)
cg25452273	15	75335524	*PPCDC*	1.8 (1, 3.42)
cg25627098	17	7328734	*SPEM2*	1.86 (0.98, 3.88)
cg25998745	8	142028625	*PTK2*	1.48 (0.99, 2.28)
cg26341457	11	72523885	*ATG16L2*	0.68 (0.35, 1.33)
cg27631602	14	21484862	*NDRG2*	2.14 (1.02, 3.53)

Table B3: CVD deaths per 100,000 person-years for the doubling of urinary arsenic levels not attributable (direct effect) and attributable (indirect effect) to changes in DNA methylation for each CpG (one marker at a time approach).

CpG	Gene	Function	Deaths attributable to a doubling of urinary As (95% CI) (direct effect)	Deaths attributable to a doubling of urinary As through DNAm (95% CI) (indirect effect)	% deaths attributable to a doubling of urinary As explained by DNAm (95 % CI)
cg05779585	*LOC286083*	Unknown function	91.9 (-9.7, 193.3)	52.9 (6.1, 120.4)	36.5 (4.3, 109.0)
cg19693031	*TXNIP*	Binding partner for redox signaling protein thioredoxin	70.7 (-35.4, 176.4)	43.5 (18.1, 75.4)	38.1 (9.9, 198.5)
cg06716655	*ADAR*	RNA editing enzyme involved in innate immunity	88.9 (-16.1, 193.6)	25.1 (7.2, 47.5)	22 (3.2, 114.5)
cg17608381	*HLA-A*	Central role in the immune system	91.9 (-14.2, 197.8)	24.1 (6.5, 45.8)	20.8 (2.8, 112.3)
cg22294740	*LINGO3*	Unknown function	89.9 (-14.6, 194.2)	22.7 (1.4, 47.7)	20.2 (-3.4, 108.3)
cg03362418	*TYMP**	Angiogenesis in vivo. Possible therapeutic target for CVD	93.4 (-11.1, 197.6)	21.3 (4.6, 43.0)	18.5 (1.6, 94.7)
cg11946459	*HLA-A*	Central role in the immune system	98.2 (-6.2, 202.3)	18.4 (3.6, 37.1)	15.8 (1.2, 81.3)
cg21990700	*C1RL**	Complement protein in the endoplasmic reticulum	92.3 (-11.9, 196.3)	18.3 (5.7, 34.9)	16.6 (2.6, 91.3)
cg06970472	*APBB2**	Beta cell function and insulin secretion	99.2 (-4.6, 202.8)	16.4 (5.0, 31.4)	14.2 (2.8, 71.3)
cg03026982	*NAV2**	Blood pressure regulation	101.4 (-3.2, 205.8)	15.5 (1.9, 34.8)	13.2 (0.3, 66.1)
cg05527044	*EGR4*	Transcription regulation	101.3 (-2.2, 204.7)	13.5 (0.6, 30.7)	11.7 (-1.6, 60.6)
cg00451635	*EMP2**	Endothelial cell migration and angiogenesis	106.8 (2.9, 210.4)	9.9 (0.2, 24.2)	8.5 (-1.0, 43.1)
cg27523527	*BARHL2*	Potential regulator of neural basic helix-loop-helix genes	104.8 (0.9, 208.4)	7.7 (0.1, 19.5)	6.9 (-1.3, 37.8)
cg19301366	*HLA-DQB1*	Type 1 diabetes susceptibility	106.8 (3.2, 210.2)	3.5 (0.04, 8.8)	3.2 (-0.7, 18.6)
cg06970472	*APBB2**	Beta cell function, insulin secretion	205.7 (13.7, 397.3)	27.8 (7.7, 54.8)	11.9 (2.6, 52.3)
cg06716655	*ADAR2*	RNA editing enzyme involved in innate immunity	203.3 (7.0, 399.2)	25.7 (3.9, 56.5)	11.2 (0.9, 55.7)
cg18618815	*COL1A1**	Extracellular matrix. As-induced remodeling mice model	198.5 (3.1, 393.4)	23.7 (4.8, 49.8)	10.7 (1.2, 54.9)
cg01178924	*LMO7*	Development of muscle and heart tissues. Pancreatic cancer	208.7 (13.6, 403.4)	23.7 (0.4, 54.7)	10.2 (-0.8, 48.8)
cg01542019	*TECR*	Sphingolipid synthesis and oxidoreductase activity	202.1 (7.7, 396.1)	21.4 (2.3, 48.4)	9.6 (0.2, 48.8)
cg02047803	*RELL2*	Apoptosis	206.3 (13.3, 398.8)	18.7 (0.7, 45.6)	8.3 (-0.3, 43.5)
cg16335098	*SMOC2*	Angiogenesis in tumor growth and myocardial ischemia	219.2 (25.7, 412.2)	13.1 (2.7, 26.9)	5.7 (0.8, 25.4)

Abbreviations: DNAm, DNA methylation; CI, confidence interval.

The sum of the direct and indirect effect represents the total effect for a doubling of urinary arsenic in CVD deaths.

Models adjusted for age, sex, smoking status, BMI, LDL cholesterol, study center (Arizona, Oklahoma or North and South Dakota), cell counts (CD8T, CD4T, NK, B cells and monocytes) and genetic PCs.

* CpG sites selected by ISIS – Aenet as predictive of CVD mortality.

To account for the withdrawal of one of the Tribal Nations, models were weighted with approximately 1/3 of weight for each center (33.0 % Arizona, 33.6 % Oklahoma, 33.4 % North Dakota / South Dakota) using inverse probability weighting.

Table B4: Gene Ontology enrichment for differentially methylated positions that were significant in the mediation analysis for CVD incidence.

GO number	Ontology	Term	N	DE	p-value	FDR
GO:0009605	BP	Response to external stimulus	28	6	0,0043	1
GO:0003725	MF	Double-stranded RNA binding	2	2	0,0047	1
GO:0016246	BP	RNA interference	2	2	0,0047	1
GO:0031054	BP	Pre-miRNA processing	2	2	0,0047	1
GO:0035280	BP	miRNA loading onto RISC involved in gene silencing by miRNA	2	2	0,0047	1
GO:0060147	BP	Regulation of posttranscriptional gene silencing	2	2	0,0047	1
GO:0060966	BP	Regulation of gene silencing by RNA	2	2	0,0047	1
GO:0060968	BP	Regulation of gene silencing	2	2	0,0047	1
GO:0070922	BP	Small RNA loading onto RISC	2	2	0,0047	1
GO:0034340	BP	Response to type I interferon	2	2	0,0092	1
GO:0060337	BP	Type I interferon signaling pathway	2	2	0,0092	1
GO:0071357	BP	Cellular response to type I interferon	2	2	0,0092	1
GO:0022613	BP	Ribonucleoprotein complex biogenesis	3	2	0,0099	1
GO:0022618	BP	Ribonucleoprotein complex assembly	3	2	0,0099	1
GO:0071826	BP	Ribonucleoprotein complex subunit organization	3	2	0,0099	1

Abbreviations: GO, gene ontology; BP, biological process; MF, molecular funcion; N, total number of terms; DE, number of enriched terms; FDR, false discovery rate.

Table B5: Gene Ontology enrichment of differentially methylated positions that were significant in the mediation analysis for CVD mortality.

GO number	Ontology	Term	N	DE	p-value	FDR
GO:0002252	BP	Immune effector process	13	5	0,00015	0,37
GO:0002449	BP	Lymphocyte mediated immunity	7	4	0,00022	0,37
GO:0002460	BP	Adaptive immune response	7	4	0,00022	0,37
GO:0002250	BP	Adaptive immune response	8	4	0,00036	0,46
GO:0002443	BP	Leukocyte mediated immunity	9	4	0,00058	0,51
GO:0045087	BP	Innate immune response	10	4	0,00061	0,51
GO:0033218	MF	Amide binding	6	3	0,0011	0,71
GO:0042277	MF	Peptide binding	6	3	0,0011	0,71
GO:0050776	BP	Regulation of immune response	14	4	0,0023	0,88
GO:0006955	BP	Immune response	24	5	0,0025	0,88
GO:0034341	BP	Response to interferon-gamma	2	2	0,0032	0,88
GO:0042611	CC	MHC protein complex	2	2	0,0032	0,88
GO:0060333	BP	Interferon-gamma-mediated signaling pathway	2	2	0,0032	0,88
GO:0071346	BP	Cellular response to interferon-gamma	2	2	0,0032	0,88
GO:0071556	CC	Integral component of lumenal side of endoplasmic reticulum membrane	2	2	0,0032	0,88
GO:0098553	CC	Lumenal side of endoplasmic reticulum membrane	2	2	0,0032	0,88
GO:0034340	BP	Response to type I interferon	2	2	0,0034	0,88
GO:0060337	BP	Type I interferon signaling pathway	2	2	0,0034	0,88
GO:0071357	BP	Cellular response to type I interferon	2	2	0,0034	0,88
GO:0002455	BP	Humoral immune response mediated by circulating immunoglobulin	2	2	0,0039	0,88
GO:0006959	BP	Humoral immune response	2	2	0,0039	0,88
GO:0016064	BP	Immunoglobulin mediated immune response	2	2	0,0039	0,88
GO:0019724	BP	B cell mediated immunity	2	2	0,0039	0,88
GO:0019221	BP	Cytokine-mediated signaling pathway	8	3	0,0054	1
GO:0003073	BP	Regulation of systemic arterial blood pressure	3	2	0,0055	1
GO:0001906	BP	Cell killing	3	2	0,0069	1
GO:0001909	BP	Leukocyte mediated cytotoxicity	3	2	0,0069	1
GO:0001913	BP	T cell mediated cytotoxicity	3	2	0,0069	1
GO:0002478	BP	Antigen processing and presentation of exogenous peptide antigen	3	2	0,0069	1
GO:0019882	BP	Antigen processing and presentation	3	2	0,0069	1
GO:0019884	BP	Antigen processing and presentation of exogenous antigen	3	2	0,0069	1
GO:0048002	BP	Antigen processing and presentation of peptide antigen	3	2	0,0069	1
GO:0042605	MF	Peptide antigen binding	3	2	0,0069	1
GO:0050778	BP	Positive regulation of immune response	10	3	0,0076	1
GO:0006952	BP	Defense response	20	4	0,0082	1
GO:0002682	BP	Regulation of immune system process				

Abbreviations: GO, gene ontology; BP, biological process; MF, molecular funcion; N, total number of terms; DE, number of enriched terms; FDR, False dicovery rate.

Table B6: KEGG enrichment for differentially methylated positions that were significant in the mediation analysis for CVD mortality.

Pathway	Description	N	DE	P.DE	FDR
path:hsa04514	Cell adhesion molecules	2	2	0,0032	0,10
path:hsa04612	Antigen processing and presentation	2	2	0,0032	0,10
path:hsa04940	Type I diabetes mellitus	2	2	0,0032	0,10
path:hsa05320	Autoimmune thyroid disease	2	2	0,0032	0,10
path:hsa05330	Allograft rejection	2	2	0,0032	0,10
path:hsa05332	Graft-versus-host disease	2	2	0,0032	0,10
path:hsa05416	Viral myocarditis	3	2	0,0062	0,17
path:hsa05166	Human T-cell leukemia virus 1 infection	4	2	0,0093	0,20
path:hsa05168	Herpes simplex virus 1 infection	3	2	0,011	0,20
path:hsa05169	Epstein-Barr virus infection	3	2	0,011	0,20
path:hsa04145	Phagosome	4	2	0,011	0,20
path:hsa05164	Influenza A	5	2	0,023	0,38

Abbreviations: N, total number of pathways; DE, number of enriched pathways; FDR, False dicovery rate.

Table B7: Other traits associated with CpGs showing significant mediated effects for CVD in our study according to EWAS Catalog [1].

CpG	Gene	Author	PMID	Trait	N
cg19693031	*TXNIP*	Chambers JC	26095709	Type 2 diabetes	3805
cg03497652	*ANKS3*	Sikdar S	31536415	Smoking	15907
cg22294740	*LINGO3*	Liu C	27843151	Alcohol intake	2423
cg17608381	*HLA-A*	Dugue P-A	31789449	Alcohol intake	5606
cg21990700	*C1RL*	Joehanes R	27651444	Smoking	13474
cg14827056	*AGO2*	Sikdar S	31536415	Smoking	15907
cg13092901	*TYMP*	Marioni R	29311653	Cognitive ability	4794
cg11946459	*HLA-A*	Liu C	27843151	Alcohol intake	2423
cg18618815	*COL1A1*	Sharp GC	29016858	Maternal BMI and offspring DNA methylation	7523
cg01178924	*LMO7*	Kazmi N	31230546	Pregnancy-related hypertension	5242
cg01542019	*TECR*	Singmann P	26500701	Sex (autosomal differences)	1799
cg02047803	*RELL2*	Albao D	31691802	Type 2 diabetes	365
cg06970472	*APBB2*	Sikdar S	31536415	Smoking	15907
cg02145701	*BANP*	Bohlin J	27717397	Gestational age	1068
cg05527044	*EGR4*	Liu J	31197173	Fasting insulin	4808
cg00451635	*EMP2*	Liu C	27843151	Alcohol intake	2423
cg27523527	*BARHL2*	Bonder MJ	25282492	Fetal vs adult liver	195

For those CpGs for which associations with several traits were found in the EWAS catalog, either the most relevant trait for this work or the study with the larger sample size are shown.

Table B8: Significant genes in mediation analysis in the Strong Heart Study that were differentially methylated in liver samples from the mouse model of in utero arsenic exposure compared to controls.

Mouse gene	Outcome in mediation analysis in the Strong Heart Study	Number of DMRs (male / female) annotated to the gene in the mouse model	Number of DMPs (male / female) annotated to the gene in the mouse model	Genomic position of the DMPs
Tgfbr1	CVD incidence	5 / 4	1 / 0	47429393
Arrdc2	CVD incidence	5 / 2	1 / 0	73359785
Ago2	CVD incidence	8 / 2	2 / 0	72999018, 72977447
Nisch	CVD incidence	2 / 0	1 / 0	32008471
Lmo7	CVD incidence	23 / 7	6 / 0	102168435, 102232355, 102232332, 102232208, 102296394, 102136457
Adar	CVD mortality	4 / 5	1 / 0	89534367
Apbb2	CVD mortality	3 / 16	4 / 0	66999334, 66978308, 66724458, 66733745
Nav2	CVD mortality	31 / 15	8 / 1	56849475, 56830246, 56621107, 56724015, 56583581, 56804011, 56665515, 56605002, 56747173
Egr4	CVD mortality	2 / 2	1 / 0	85463274
Lingo3	CVD incidence and mortality	0 / 1	2 / 0	80308751, 80306748
Ubac1	CVD incidence	1 / 0	0 / 0	-
Eppk1	CVD incidence	1 / 2	0 / 0	-
Tecr	CVD incidence	3 / 1	0 / 0	-
Smoc2	CVD incidence	4 / 11	0 / 0	-
Klf9	CVD mortality	4 / 4	0 / 0	-
C1rl	CVD mortality	4 / 0	0 / 0	-
Emp2	CVD mortality	4 / 9	0 / 0	-
Barhl2	CVD mortality	8 / 10	0 / 0	-
Txnip	CVD incidence and mortality	4 / 4	0 / 0	-
Tymp	CVD incidence and mortality	1 / 0	0 / 0	-

Appendix C: Supplementary tables for section 4.4.2

Table C1: Hazards ratios (95 % CI) of CpGs selected by ISIS-enet as associated with lung cancer comparing percentile 90th vs 10th.

CpG	Chr	Position	Gene	HR (95 % CIs)
cg00367135	11	61722666	*BEST1*	1.24 (0.87, 1.9)
cg00524773	16	121668	*RHBDF1*	1.25 (1, 1.59)
cg01513913	14	106329158	*FAM30A*	0.68 (0.44, 1)
cg01571467	5	432252	*AHRR*	1.09 (0.87, 1.41)
cg01692968	9	108005349	*SLC44A1*	1.34 (0.99, 2.05)
cg01765406	2	129231478	*HS6ST1*	1.35 (1, 1.93)
cg01899089	5	369969	*AHRR*	0.87 (0.59, 1.22)
cg01940273	2	233284934	*ALPG*	0.98 (0.63, 1.3)
cg02560069	15	39425615	*C15orf54*	1.15 (0.89, 1.65)
cg02738868	14	74221164	*ELMSAN1*	0.86 (0.6, 1.06)
cg03062284	2	122994061	*TSN*	0.76 (0.5, 1.03)
cg03329539	2	233283329	*ALPG*	0.74 (0.48, 1.13)
cg03636183	19	17000585	*F2RL3*	0.91 (0.63, 1.2)
cg03707168	19	49379127	*PPP1R15A*	0.72 (0.46, 1.02)
cg04009588	9	35619585	*CD72*	1.14 (0.85, 1.72)
cg04428531	3	109525931	*LINC01205*	0.98 (0.85, 1.05)
cg04885881	1	11123118	*SRM*	0.84 (0.54, 1.11)
cg05049335	11	66103889	*RIN1*	1.63 (1.08, 2.4)
cg05221370	7	110738836	*IMMP2L*	0.83 (0.61, 1.04)
cg05284742	14	93552128	*ITPK1*	1.35 (1, 1.92)
cg05575921	5	373378	*AHRR*	0.63 (0.45, 0.9)
cg06521527	6	11217462	*NEDD9*	0.74 (0.53, 1.01)
cg07251887	17	73641809	*RECQL5*	0.58 (0.38, 0.9)

cg07267541	9	12784592	*LURAP1L*	0.99 (0.67, 1.33)
cg08371497	22	27029030	*CRYBA4*	1.07 (0.75, 1.6)
cg09834951	19	1265877	*CIRBP-AS1*	1.1 (0.79, 1.67)
cg09842685	12	4492769	*FGF23*	0.67 (0.44, 0.97)
cg10041129	11	117685550	*DSCAML1*	0.92 (0.63, 1.25)
cg11556164	7	110738315	*IMMP2L*	1.17 (1, 1.42)
cg11660018	11	86510915	*PRSS23*	1.47 (1.03, 2.2)
cg11902777	5	368843	*AHRR*	0.7 (0.5, 0.97)
cg12144776	6	25166749	*CMAHP*	0.73 (0.45, 1.03)
cg13772414	2	222383060	*EPHA4*	1.54 (1.02, 2.19)
cg13937905	12	53612551	*RARG*	0.84 (0.58, 1.08)
cg14391737	11	86513429	*PRSS23*	0.65 (0.43, 0.99)
cg14580211	5	150161299	*SMIM3*	0.93 (0.59, 1.33)
cg14624207	11	68142198	*LRP5*	0.66 (0.42, 1)
cg16201146	20	19191526	*SLC24A3*	0.67 (0.45, 1)
cg16207944	14	106331592	*FAM30A*	0.86 (0.58, 1.06)
cg16727193	3	126627910	*CHCHD6*	1.05 (0.83, 1.54)
cg16998502	18	71347435	*LINC02582*	1.32 (1, 1.97)
cg17738628	15	67155520	*SMAD6*	1.33 (0.99, 2.12)
cg17739917	17	38477572	*RARA*	1.06 (0.81, 1.59)
cg18158149	1	162138215	*NOS1AP*	0.93 (0.63, 1.33)
cg19136686	16	17464401	*XYLT1*	1.37 (0.98, 2.11)
cg19578936	17	2163849	*SMG6*	0.88 (0.63, 1.08)
cg19885130	11	68146832	*LRP5*	0.89 (0.55, 1.23)
cg20295214	1	206226794	*AVPR1B*	1.14 (0.89, 1.67)
cg20731257	2	87883497	*RMND5A*	1.35 (1, 1.98)
cg21217140	6	131981534	*ENPP3*	1.69 (1.13, 2.45)
cg21566642	2	233284661	*ALPG*	0.9 (0.59, 1.21)
cg21733098	12	127931219	*LINC02393*	0.89 (0.61, 1.27)
cg23025288	2	129278724	*HS6ST1*	0.9 (0.57, 1.26)
cg23771366	11	86510998	*PRSS23*	0.9 (0.63, 1.14)
cg23916896	5	368804	*AHRR*	0.91 (0.6, 1.21)
cg24021808	13	40805588	*LINC00548*	1.06 (0.78, 1.5)
cg24556382	4	174173455	*GALNT7*	0.9 (0.63, 1.29)
cg24859433	6	30720203	*IER3*	0.7 (0.48, 0.97)
cg24947681	15	39760933	*THBS1*	0.75 (0.51, 1.07)
cg25799109	3	57102900	*ARHGEF3*	0.78 (0.51, 1.09)
cg26916621	17	46657346	*HOXB3*	1.35 (0.94, 2.09)
cg27241845	2	233250370	*ECEL1P2*	0.77 (0.53, 1.06)

Table C2: Hazards ratios (95 % CI) of CpGs selected by ISIS-enet as associated with smoking-related cancer comparing percentile 90th vs 10th.

CpG	Chr	Position	Gene	HR (95 % CIs)
cg00073090	19	1265879	*CIRBP-AS1*	0.64 (0.45, 0.95)
cg00524773	16	121668	*RHBDF1*	1.18 (0.98, 1.5)
cg01002722	7	5608879	*FSCN1*	1.75 (1.26, 2.4)
cg01513913	14	106329158	*FAM30A*	0.58 (0.39, 0.86)
cg01692968	9	108005349	*SLC44A1*	1.15 (0.82, 1.71)
cg01899089	5	369969	*AHRR*	0.76 (0.55, 1.03)
cg01901332	11	75031054	*ARRB1*	1.41 (1, 2.04)
cg01940273	2	233284934	*ALPG*	1.3 (1, 1.95)
cg02738868	14	74221164	*ELMSAN1*	0.76 (0.5, 1.04)
cg03062284	2	122994061	*TSN*	0.76 (0.55, 1.04)
cg03368099	7	27184521	*HOXA-AS3*	1.05 (0.8, 1.45)
cg03636183	19	17000585	*F2RL3*	0.88 (0.61, 1.21)
cg03707168	19	49379127	*PPP1R15A*	0.85 (0.57, 1.16)
cg03977382	11	99564911	*CNTN5*	1.23 (0.94, 1.67)
cg04009588	9	35619585	*CD72*	1.2 (0.89, 1.69)
cg05049335	11	66103889	*RIN1*	1.37 (1, 2.03)
cg05547483	7	80741676	*SEMA3C*	1.39 (0.96, 2.25)
cg05575921	5	373378	*AHRR*	0.96 (0.64, 1.28)
cg05934812	5	334322	*AHRR*	0.85 (0.54, 1.2)
cg07251887	17	73641809	*RECQL5*	0.69 (0.45, 1)
cg07267541	9	12784592	*LURAP1L*	0.82 (0.59, 1.06)
cg07943658	5	352001	*AHRR*	1.17 (0.84, 1.66)
cg08371497	22	27029030	*CRYBA4*	0.81 (0.55, 1.08)
cg09338374	22	39888390	*MGAT3*	1.13 (0.81, 1.65)
cg09834951	19	1265877	*CIRBP-AS1*	1.49 (1, 2.16)
cg09842685	12	4492769	*FGF23*	0.8 (0.58, 1.06)
cg10258214	14	106330534	*FAM30A*	1.09 (0.83, 1.6)
cg11464806	7	156837580	*MNX1-AS1*	1.06 (0.82, 1.48)
cg11556164	7	110738315	*IMMP2L*	1.13 (0.93, 1.39)
cg11931220	12	49276387	*RND1*	1.27 (0.95, 1.77)
cg12144776	6	25166749	*CMAHP*	0.87 (0.58, 1.24)
cg12409728	13	31150700	*HMGB1*	1.51 (1.08, 2.02)
cg12571376	9	89019429	*TUT7*	0.98 (0.81, 1.18)
cg12615852	14	106330121	*FAM30A*	1.07 (0.81, 1.49)
cg13772414	2	222383060	*EPHA4*	1.56 (1.11, 2.25)
cg14391737	11	86513429	*PRSS23*	0.68 (0.45, 1)
cg14580211	5	150161299	*SMIM3*	1.06 (0.78, 1.54)
cg15310518	14	106330520	*FAM30A*	0.64 (0.42, 0.95)
cg15559352	18	74785799	*MBP*	0.88 (0.59, 1.22)
cg16201146	20	19191526	*SLC24A3*	0.57 (0.39, 0.83)
cg16519923	16	30485810	*ITGAL*	0.92 (0.58, 1.24)
cg16727193	3	126627910	*CHCHD6*	0.81 (0.52, 1.14)
cg17569124	7	27183643	*HOXA-AS3*	1.22 (0.92, 1.69)

cg17738628	15	67155520	*SMAD6*	1.16 (0.83, 1.71)
cg18110140	15	75350380	*PPCDC*	0.82 (0.57, 1.1)
cg18158149	1	162138215	*NOS1AP*	0.86 (0.6, 1.17)
cg18446336	7	2847575	*GNA12*	1.05 (0.78, 1.43)
cg19578936	17	2163849	*SMG6*	1.04 (0.76, 1.54)
cg19859270	3	98251294	*GPR15*	0.77 (0.61, 1)
cg19885130	11	68146832	*LRP5*	0.72 (0.46, 1.02)
cg20174472	20	61283288	*SLCO4A1*	0.7 (0.48, 1)
cg20295214	1	206226794	*AVPR1B*	1.17 (0.92, 1.58)
cg21217140	6	131981534	*ENPP3*	1.15 (0.85, 1.64)
cg21322436	7	145812842	*CNTNAP2*	0.82 (0.56, 1.17)
cg21566642	2	233284661	*ALPG*	0.94 (0.61, 1.28)
cg21704177	3	98257530	*GPR15*	0.85 (0.64, 1.1)
cg21911711	19	16998668	*F2RL3*	0.9 (0.64, 1.22)
cg22222502	5	150161551	*SMIM3*	1.08 (0.83, 1.5)
cg22851561	14	74214183	*ELMSAN1*	1.16 (0.87, 1.64)
cg23025288	2	129278724	*HS6ST1*	0.98 (0.71, 1.3)
cg24021808	13	40805588	*LINC00548*	0.97 (0.68, 1.25)
cg24859433	6	30720203	*IER3*	0.77 (0.58, 1)
cg25189904	1	68299493	*GNG12-AS1*	0.96 (0.67, 1.26)
cg25648203	5	395444	*AHRR*	1.18 (0.87, 1.68)
cg25799109	3	57102900	*ARHGEF3*	0.81 (0.57, 1.1)
cg25845814	14	74224613	*ELMSAN1*	0.78 (0.56, 1.03)
cg26337070	2	85999873	*ATOH8*	1.19 (0.89, 1.75)
cg26764244	1	68299511	*GNG12-AS1*	1.11 (0.83, 1.62)
cg27271698	14	106330538	*FAM30A*	1.27 (0.98, 1.78)

Table C3: Differences in lung cancer cases per 100,000 person-years for a 10 pack-years change attributable to differences in DNA methylation for each CpG ('mediated effects') in the Strong Heart Study and replication in the Framingham Heart Study.

		Strong Heart Study			Framingham Heart Study		
		Mediated (i.e., indirect) effect of cigarette pack-years through DNAm[b]		Direct effect of cigarette pack-years[a]	Mediated (i.e., indirect) effect of cigarette pack-years through DNAm[b]		Direct effect of cigarette pack-years[a]
CpG	Gene	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases for a 10 pack-years increase (95 % CI) per 100,000 person-years	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases for a 10 pack-years increase (95 % CI) per 100,000 person-years
cg14391737*	*PRSS23*	17.3 (7.5, 27.6)	14.2 (5.5, 31.9)	104.9 (45.1, 164.5)	-	-	-
cg05575921	*AHRR*	14.4 (6.7, 22.6)	11.9 (4.6, 28.9)	106.8 (45.5, 167.9)	10.1 (-16.2, 36.5)	4.7 (-7.3, 23.6)	206.4 (85.4, 327.2)
cg03636183	*F2RL3*	10.2 (3.3, 17.5)	8.4 (2.4, 20.6)	112.0 (51.9, 171.9)	21.5 (-2.9, 46.2)	10.9 (-1.7, 28.5)	175.8 (71.5, 279.7)
cg21566642	*ALPG*	9.4 (2.0, 17.0)	7.6 (1.5, 19.6)	113.1 (52.9, 173.1)	13.8 (-5.4, 33.1)	7.0 (-2.6, 25.9)	182.3 (69.1, 295.3)
cg24859433	*IER3*	6.9 (2.8, 11.8)	5.6 (2.1, 13.1)	115.7 (56.9, 174.4)	11.8 (-1.8, 25.9)	6.0 (-1.0, 17.5)	186.6 (79.0, 293.9)
cg03329539	*ALPG*	6.3 (0.9, 12.3)	5.2 (0.7, 14)	116.3 (56.4, 176.0)	9.6 (0.2, 19.5)	4.9 (0.1, 14.0)	186.9 (79.5, 294.2)
cg09842685*	*FGF23*	5.4 (2.3, 9.2)	4.4 (1.7, 10)	117.2 (58.8, 175.5)	-	-	-
cg11902777	*AHRR*	4.4 (1.7, 7.7)	3.6 (1.4, 7.9)	119.0 (60.9, 176.9)	6.5 (-1.3, 14.8)	3.3 (-0.7, 10.2)	191.0 (83.7, 298.1)
cg03707168	*PPP1R15A*	4.2 (0.9, 8.4)	3.4 (0.7, 8)	119.2 (61.4, 176.9)	11.1 (0.4, 22.1)	5.6 (0.2, 15.3)	186.8 (79.8, 293.7)
cg14624207	*LRP5*	3.8 (1.1, 7.3)	3.1 (0.8, 7.5)	119.2 (60.7, 177.5)	2.2 (-5.5, 10.1)	1.1 (-3.3, 6.0)	195.9 (89.2, 302.4)
cg27241845	*ECEL1P2*	3.4 (0.4, 7.2)	2.8 (0.3, 7.3)	120.0 (61.5, 178.4)	5.3 (-3.1, 14.1)	2.7 (-1.6, 9.7)	193.1 (85.1, 300.7)
cg16207944*	*FAM30A*	3.3 (0.2, 6.9)	2.7 (0.1, 7.3)	120.1 (61.4, 178.6)	-	-	-
cg01513913	*FAM30A*	3.3 (0.3, 6.9)	2.7 (0.2, 7.5)	119.8 (60.8, 178.7)	4.2 (-5.1, 13.8)	2.1 (-3, 8.2)	194.2 (87.9, 300.3)
cg01899089	*AHRR*	3.1 (0.8, 6.3)	2.6 (0.6, 6.8)	119.1 (60.3, 177.8)	9.2 (-0.1, 19.1)	4.7 (0.0, 13.0)	188.6 (81.7, 295.1)
cg07251887	*RECQL5*	3.2 (0.5, 6.6)	2.6 (0.4, 6.9)	120.3 (61.5, 178.9)	10.3 (1.9, 19.4)	5.2 (1.0, 12.9)	186.8 (81.1, 292.3)
cg23916896	*AHRR*	2.7 (0.6, 5.7)	2.2 (0.4, 5.8)	121.7 (63.1, 180.1)	14.8 (3.9, 26.2)	7.5 (1.9, 19.2)	181.3 (74.5, 287.9)
cg24947681*	*THBS1*	2.7 (0.3, 5.7)	2.2 (0.2, 5.7)	121.3 (62.8, 179.6)	-	-	-
cg06521527*	*NEDD9*	2.6 (0.3, 5.6)	2.1 (0.3, 5.4)	120.6 (62.5, 178.7)	-	-	-
cg04885881	*SRM*	2.6 (0.1, 5.8)	2.1 (0.1, 5.9)	120.4 (61.9, 178.7)	5.3 (-3.1, 14.1)	2.7 (-1.6, 9.7)	193.1 (85.2, 300.7)
cg18158149*	*NOS1AP*	1.9 (0.2, 4.3)	1.5 (0.1, 4)	122.1 (63.9, 180.1)	-	-	-
cg04885881	*SRM*	2.6 (0.1, 5.8)	2.1 (0.1, 5.9)	120.4 (61.9, 178.7)	5.3 (-3.1, 14.1)	2.7 (-1.6, 9.7)	193.1 (85.2, 300.7)
cg18158149*	*NOS1AP*	1.9 (0.2, 4.3)	1.5 (0.1, 4)	122.1 (63.9, 180.1)	-	-	-

Abbreviations: CI, confidence interval; DNAm, DNA methylation. * CpGs not present in the 450K array, therefore not evaluated in the Framingham Heart Study.
Models were adjusted for age, sex, current smoking, BMI and cell counts (CD8T, CD4T, NK, B cells and monocytes). Additionally adjusted for study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs in the Strong Heart Study.
[a] Absolute changes in cancer incidence (per 100,000 person-years) for a 10 pack-years change were obtained from additive hazards models.
[b] Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for a 10 pack-years increase from the mediator model by the absolute change in cancer incidence cases for a 10 pack-years increase (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were derived by simulation from the estimated model coefficients and covariance matrices.

Table C4: Differences in smoking-related cancer cases per 100,000 person-years comparing current to never smokers attributable to changes in DNA methylation for each CpG ('mediated effects') in the Strong Heart Study and replication in the Framingham Heart Study.

		Strong Heart Study			Framingham Heart Study		
		Mediated (i.e., indirect) effect of current vs never smoking through DNAm[b]		Direct effect of current vs never smoking[a]	Mediated (i.e., indirect) effect of current vs never smoking through DNAm[b]		Direct effect of current vs never smoking[a]
CpG	Gene	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases for current vs never smoking (95 % CI) per 100,000 person-years	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases for current vs never smoking (95 % CI) per 100,000 person-years
cg05575921	*AHRR*	291.7 (180.5, 404.3)	68.7 (40.4, 116.9)	133.1 (-49.3, 314.9)	178.4 (3.3, 353.8)	139.3 (-495.3, 928.2)	-50.3 (-239.2, 138.5)
cg21566642	*ALPG*	239.1 (141.6, 338.5)	55.3 (30.9, 96.2)	193.2 (11.3, 374.6)	126.9 (-3.2, 257.4)	100.8 (-453.0, 779.6)	-1.0 (-185.9, 183.7)
cg14391737*	*PRSS23*	217.5 (138.9, 299.2)	48 (29.5, 78.2)	235.8 (68.1, 402.9)	-	-	-
cg03636183	*F2RL3*	200.4 (116.9, 286.1)	46.8 (26.2, 80.8)	227.6 (55.8, 399.0)	235.1 (84.4, 386.5)	180.8 (-785.0, 1319.2)	-105.0 (-293.4, 83.2)
cg01940273	*ALPG*	150.3 (68.7, 233.4)	34.8 (15.2, 65.7)	281.4 (100.6, 461.7)	119.9 (4.9, 235.4)	96.5 (-446.7, 745.4)	4.3 (-170.5, 179.0)
cg19859270	*GPR15*	148.4 (69.9, 230.3)	33.4 (16.2, 57.1)	296.3 (134.6, 457.9)	71.6 (-5.0, 148.7)	58.0 (-235.3, 413.8)	51.8 (-98.9, 202.6)
cg25845814*	*ELMSAN1*	108.8 (56.8, 163.6)	25.1 (12.9, 43.8)	324.0 (163.2, 484.5)	-	-	-
cg18110140*	*PPCDC*	94.6 (41.6, 150.3)	21.8 (9.3, 40.6)	340.1 (172.9, 506.8)	-	-	-
cg25648203	*AHRR*	89.9 (29.9, 151.8)	20.9 (7, 40.2)	340.7 (175.6, 505.4)	86.4 (8.6, 164.7)	69.4 (-249.4, 450.0)	38.1 (-102.6, 178.7)
cg21911711*	*F2RL3*	86.5 (33.6, 141.6)	19.8 (7.6, 37.3)	351.6 (185.8, 516.9)	-	-	-
cg24859433	*IER3*	80.1 (35.2, 127.8)	18.3 (7.9, 33.9)	357.9 (193.7, 521.7)	97.0 (10.2, 184.7)	77.7 (-367.2, 603.5)	27.9 (-136.4, 192.2)
cg01899089	*AHRR*	67.1 (30.4, 107.0)	15.4 (6.8, 28.7)	369.5 (205.5, 533.1)	-5.1 (-44.6, 34.3)	-4.3 (-141.7, 112.6)	122.1 (-26.5, 270.7)
cg09842685*	*FGF23*	65.4 (24.6, 108.7)	15.0 (5.5, 28.8)	371.9 (207.9, 535.5)	-	-	-
cg25189904	*GNG12-AS1*	57.3 (8.1, 108.2)	13.2 (1.9, 27.9)	375. (211.7, 539.6)	60.9 (-7.7, 130.3)	49.9 (-273.3, 433.4)	61.2 (-104.4, 226.9)
cg01513913	*FAM30A*	56.1 (24.1, 91.6)	13 (5.4, 24.5)	376.7 (213.9, 539.3)	27.8 (-4.1, 60.5)	23.1 (-123.9, 195.9)	92.5 (-61.6, 246.5)
cg02738868*	*ELMSAN1*	55.8 (20.4, 94.4)	12.8 (4.5, 25.3)	380.6 (215.3, 545.4)	-	-	-
cg03707168	*PPP1R15A*	56.3 (16.7, 99.2)	12.8 (3.8, 25.4)	382.1 (219.5, 544.3)	39.6 (-10.3, 90.1)	32.4 (-140.7, 245.1)	82.6 (-67.5, 232.7)
cg15310518*	*FAM30A*	52.6 (19.8, 88.7)	12.2 (4.5, 24)	377.9 (214.4, 541.1)	-	-	-
cg07943658*	*AHRR*	51.5 (15.1, 91.0)	11.8 (3.5, 23.4)	385.6 (222.5, 548.4)	-	-	-
cg05934812*	*AHRR*	44.1 (17.1, 75.6)	9.9 (3.9, 18.4)	402.1 (241.7, 562.3)	-	-	-
cg00073090	*CIRBP-AS1*	42.7 (12.0, 76.2)	9.8 (2.7, 20.1)	390.9 (227.9, 553.4)	-0.7 (-43.7, 42.2)	-0.6 (-130.7, 114.3)	118.4 (-31.3, 268.1)
cg12615852*	*FAM30A*	40.6 (10.1, 73.8)	9.4 (2.4, 19)	393.8 (232.2, 555.0)	-	-	-
cg16201146	*SLC24A3*	41.4 (13.7, 72.9)	9.4 (3.2, 18.3)	400.2 (239.0, 561.4)	26.9 (-3.0, 58.4)	22.2 (-106.6, 176.1)	94.5 (-57.3, 246.2)
cg19885130*	*LRP5*	40.9 (9.3, 75.2)	9.4 (2.2, 18.9)	393.2 (233.9, 552.0)	-	-	-
cg21322436	*CNTNAP2*	40.4 (5.9, 76.9)	9.3 (1.3, 20.5)	394.6 (229.0, 559.7)	30.9 (-16.7, 78.4)	21.7 (-52.1, 95.8)	18.1 (-211.7, 301.1)
cg20174472*	*SLCO4A1*	37.9 (7.4 71.2)	8.7 (1.7, 18.4)	398.8 (235.7, 561.4)	-	-	-
cg27271698*	*FAM30A*	34.9 (2.7, 69.2)	8.1 (0.6, 18)	397.6 (235.0, 559.9)	-	-	-
cg10258214*	*FAM30A*	33.9 (3.1, 67.0)	7.9 (0.7, 17.3)	398.8 (236.5, 560.8)	-	-	-
cg07267541*	*LURAP1L*	33.3 (7.4, 62.3)	7.6 (1.7, 16.2)	402.3 (239.1, 565.1)	-	-	-

cg07251887	*RECQL5*	32.5 (5.7, 62.5)	7.5 (1.3, 16)	401.6 (239.8, 562.9)	33.7 (-13.3, 80.7)	7.6 (-40.2, 55.5)	6.4 (-120.6, 141.4)
cg18158149*	*NOS1AP*	32.1 (2.8, 64.3)	7.3 (0.7, 15.6)	409.7 (249.7, 569.3)	-	-	-
cg09338374*	*MGAT3*	31.9 (3.5, 62.9)	7.2 (0.8, 15.7)	408.4 (246.8, 569.8)	-	-	-
cg23025288*	*HS6ST1*	26.2 (0.4, 54.6)	6 (0.1, 13.9)	406.9 (245.4, 567.9)	-	-	-
cg16519923	*ITGAL*	22.9 (1.1, 48.0)	5.2 (0.3, 11.9)	418.2 (256.4, 579.7)	27.1 (-16.8, 71.1)	41.8 (5.1, 79.7)	33.8 (-152.8, 253.0)
cg12409728*	*HMGB1*	15.6 (0.6, 34.5)	3.6 (0.1, 8.8)	417.7 (255.5, 579.5)	-	-	-
cg01002722	*FSCN1*	15.1 (2.8, 31.9)	3.4 (0.6, 7.6)	429.1 (267.3, 590.9)	32.1 (-14.1, 78.3)	0.9 (-4.3, 7.6)	0.8 (-11.6, 15.4)
cg21704177*	*GPR15*	14.9 (0.7, 33.1)	3.4 (0.2, 8.2)	422.2 (261.3, 582.7)	-	-	-

Abbreviations: CI, confidence interval; DNAm, DNA methylation.

* CpGs not present in the 450K array, therefore not evaluated in the Framingham Heart Study.

Models were adjusted for age, sex, former smoking, BMI and cell counts (CD8T, CD4T, NK, B cells and monocytes). Additionally adjusted for study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs in the Strong Heart Study.

[a] Absolute changes in cancer incidence (per 100,000 person-years) for current versus never smokers were obtained from additive hazards models.

[b] Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for the current versus never smoking comparison from the mediator model by the absolute change in cancer incidence cases for the current versus never smoking comparison (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were derived by simulation from the estimated model coefficients and covariance matrices.

Table C5: Differences in smoking-related cancer cases per 100,000 person-years for a 10 pack-years change attributable to differences in DNA methylation for each CpG ('mediated effects') in the Strong Heart Study and replication in the Framingham Heart Study.

		Strong Heart Study			Framingham Heart Study		
		Mediated (i.e., indirect) effect of cigarette pack-years through DNAm[b]		Direct effect of cigarette pack-years[a]	Mediated (i.e., indirect) effect of cigarette pack-years through DNAm[b]		Direct effect of cigarette pack-years[a]
CpG	Gene	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases for a 10 pack-years increase (95 % CI) per 100,000 person-years	Difference in cancer cases attributable to DNAm (95 % CI) per 100,000 person-years	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases for a 10 pack-years increase (95 % CI) per 100,000 person-years
cg14391737*	*PRSS23*	28.4 (14.2, 43.2)	18.7 (8.3, 39.5)	123.4 (52.5, 194.1)	-	-	-
cg05575921	*AHRR*	19.5 (8.0, 31.5)	12.9 (4.6, 29.6)	131.3 (59.9, 202.6)	42.2 (8.6, 76.2)	51.7 (-4.6, 337.8)	39.5 (-49.4, 128.2)
cg21566642	*ALPG*	18.3 (7.4, 29.8)	12 (4.4, 26.7)	133.6 (63.4, 203.5)	28.5 (0.5, 56.8)	34.9 (-12.9, 218.4)	53.1 (-30.3, 136.3)
cg03636183	*F2RL3*	16.9 (6.9, 27.7)	11.2 (4.2, 24.6)	134.7 (65.2, 204.1)	60.6 (26.9, 94.8)	71.3 (20.8, 365.3)	24.4 (-55.5, 104.0)
cg19859270	*GPR15*	14.2 (4.8, 24.4)	9.2 (3.1, 19.3)	140.6 (72.9, 208.2)	7.5 (-9.9, 24.9)	8.7 (-23.7, 57.1)	78.1 (-0.1, 155.9)
cg01940273	*ALPG*	9.6 (1.1, 18.6)	6.3 (0.7, 15.9)	143.0 (73.4, 212.4)	23.7 (3.8, 44.1)	29.0 (-3.5, 172.1)	58.2 (-22.1, 138.2)
cg25845814*	*ELMSAN1*	9.1 (3.5, 15.6)	5.9 (2.1, 13.1)	143.7 (75.4, 211.8)	-	-	-
cg18110140*	*PPCDC*	8.4 (1.8, 15.9)	5.5 (1.1, 13.5)	144.1 (74.9, 213.1)	-	-	-
cg21911711*	*F2RL3*	7.8 (0.8, 15.3)	5.1 (0.5, 12.4)	145.8 (77.3, 214.2)	-	-	-
cg24859433	*IER3*	6.9 (1.6, 1)	4.5 (1, 11)	146.2 (77.1, 215.1)	19.1 (1.8, 37.1)	21.9 (-1.8, 126.0)	67.9 (-13.7, 149.2)
cg07943658*	*AHRR*	5.9 (0.9, 12.1)	3.9 (0.6, 9.1)	149.1 (81.3, 216.8)	-	-	-
cg01513913	*FAM30A*	5.8 (1.6, 11.2)	3.8 (0.9, 9.3)	147.2 (78.5, 215.8)	4.4 (-3.1, 12.3)	5.1 (-6.8, 33.0)	82.2 (2.8, 161.4)
cg05934812*	*AHRR*	5.8 (1.6, 11.1)	3.7 (1, 8.6)	148.3 (80.6, 215.9)	-1.1 (-10.9, 8.6)	-1.3 (-25.0, 18.5)	87.3 (74.1, 166.9)
cg25648203	*AHRR*	5.5 (0.2, 11.4)	3.6 (0.1, 9.3)	148.0 (79.3, 216.4)	13.4 (1.8, 25.8)	16 (-1.7, 106.4)	70.4 (-11.8, 152.4)
cg09842685*	*FGF23*	5.4 (0.9, 10.6)	3.5 (0.6, 8.7)	147.7 (79.0, 216.2)	-	-	-
cg15310518*	*FAM30A*	5.3 (0.9, 10.6)	3.4 (0.6, 8.8)	147.9 (79.0, 216.7)	-	-	-
cg03707168	*PPP1R15A*	4.8 (0.7, 10.0)	3.1 (0.4, 7.7)	149.0 (81.2, 216.7)	8.6 (-4.5, 21.9)	10.0 (-10.0, 65.6)	77.4 (-3.1, 157.8)
cg00073090	*CIRBP-AS1*	4.4 (0.4, 9.3)	2.9 (0.3, 7.3)	149.6 (81.5, 217.7)	-1.6 (-13.4, 10.1)	-1.9 (-38.7, 17.9)	87.8 (10.3, 165.1)
cg01899089	*AHRR*	4.2 (1.0, 8.6)	2.8 (0.6, 6.9)	147.9 (79.5, 216.3)	-	-	-
cg02738868*	*ELMSAN1*	3.8 (0.5, 8.2)	2.5 (0.3, 6.8)	148.3 (79.4, 216.9)	-	-	-
cg00073090	*CIRBP-AS1*	42.7 (12.0, 76.2)	9.8 (2.7, 20.1)	390.9 (227.9, 553.4)	-0.7 (-43.7, 42.2)	-0.6 (-130.7, 114.3)	118.4 (-31.3, 268.1)
cg12615852*	*FAM30A*	40.6 (10.1, 73.8)	9.4 (2.4, 19)	393.8 (232.2, 555.0)	-	-	-
cg16201146	*SLC24A3*	41.4 (13.7, 72.9)	9.4 (3.2, 18.3)	400.2 (239.0, 561.4)	26.9 (-3.0, 58.4)	22.2 (-106.6, 176.1)	94.5 (-57.3, 246.2)
cg19885130*	*LRP5*	40.9 (9.3, 75.2)	9.4 (2.2, 18.9)	393.2 (233.9, 552.0)	-	-	-
cg21322436	*CNTNAP2*	40.4 (5.9, 76.9)	9.3 (1.3, 20.5)	394.6 (229.0, 559.7)	30.9 (-16.7, 78.4)	21.7 (-52.1, 95.8)	18.1 (-211.7, 301.1)
cg20174472*	*SLCO4A1*	37.9 (7.4 71.2)	8.7 (1.7, 18.4)	398.8 (235.7, 561.4)	-	-	-
cg27271698*	*FAM30A*	34.9 (2.7, 69.2)	8.1 (0.6, 18)	397.6 (235.0, 559.9)	-	-	-
cg10258214*	*FAM30A*	33.9 (3.1, 67.0)	7.9 (0.7, 17.3)	398.8 (236.5, 560.8)	-	-	-
cg07267541*	*LURAP1L*	33.3 (7.4, 62.3)	7.6 (1.7, 16.2)	402.3 (239.1, 565.1)	-	-	-

cg07251887	*RECQL5*	32.5 (5.7, 62.5)	7.5 (1.3, 16)	401.6 (239.8, 562.9)	33.7 (-13.3, 80.7)	7.6 (-40.2, 55.5)	6.4 (-120.6, 141.4)	
cg18158149*	*NOS1AP*	32.1 (2.8, 64.3)	7.3 (0.7, 15.6)	409.7 (249.7, 569.3)	-	-	-	
cg09338374*	*MGAT3*	31.9 (3.5, 62.9)	7.2 (0.8, 15.7)	408.4 (246.8, 569.8)	-	-	-	
cg23025288*	*HS6ST1*	26.2 (0.4, 54.6)	6 (0.1, 13.9)	406.9 (245.4, 567.9)	-	-	-	
cg16519923	*ITGAL*	22.9 (1.1, 48.0)	5.2 (0.3, 11.9)	418.2 (256.4, 579.7)	27.1 (-16.8, 71.1)	41.8 (5.1, 79.7)	33.8 (-152.8, 253.0)	
cg12409728*	*HMGB1*	15.6 (0.6, 34.5)	3.6 (0.1, 8.8)	417.7 (255.5, 579.5)	-	-	-	
cg01002722	*FSCN1*	15.1 (2.8, 31.9)	3.4 (0.6, 7.6)	429.1 (267.3, 590.9)	32.1 (-14.1, 78.3)	0.9 (-4.3, 7.6)	0.8 (-11.6, 15.4)	
cg21704177*	*GPR15*	14.9 (0.7, 33.1)	3.4 (0.2, 8.2)	422.2 (261.3, 582.7)	-	-	-	

Abbreviations: CI, confidence interval; DNAm, DNA methylation.

* CpGs not present in the 450K array; therefore not evaluated in the Framingham Heart Study.
Models were adjusted for age, sex, current smoking, BMI and cell counts (CD8T, CD4T, NK, B cells and monocytes). Additionally adjusted for study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs in the Strong Heart Study.

[a] Absolute changes in cancer incidence (per 100,000 person-years) for a 10 pack-years change were obtained from additive hazards models.

[b] Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for a 10 pack-years increase from the mediator model by the absolute change in cancer incidence cases for a 10 pack-years increase (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % confidence intervals (CIs) in the table were derived by simulation from the estimated model coefficients and covariance matrices.

Table C6: Medians (IQR) of blood DNA methylation proportions of CpGs with statistically significant mediated effect both in the Strong Heart Study and the Framingham Heart Study.

CpG	Lung cancer		Smoking-related cancer		Non cancer		Overall	
	SHS (N=97)	FHS (N=56)	SHS (N=222)	FHS (N=251)	SHS (N=2013)	FHS (N=3919)	SHS (N=2235)	FHS (N=4170)
cg01899089	47.9 (44.4, 50.8)	45.9 (41.9, 47.8)	49.2 (45.7, 52.0)	47.2 (43.4, 50.4)	50.9 (47.9, 53.7)	49.3 (46.0, 52.18)	50.7 (47.7, 53.6)	49.2 (45.9, 52.1)
cg01940273	53.5 (49.5, 58.3)	53.5 (49.4, 57.6)	56.7 (51.0, 62.4)	56.1 (51.2, 59.5)	60.4 (55.0, 64.7)	59.1 (54.5, 62.6)	60.0 (54.5, 64.5)	58.8 (54.2, 62.5)
cg03329539	30.4 (26.6, 34.9)	30.4 (28.3, 33.1)	32.7 (27.9, 36.9)	32.9 (29.6, 35.6)	35.2 (30.9, 38.8)	35.1 (31.8, 38.8)	34.9 (30.6, 38.8)	34.9 (31.7, 38.7)
cg03636183	61.6 (56.9, 66.7)	58.6 (53.3, 62.4)	65.6 (59.6, 71.5)	61.9 (58.2, 66.7)	69.9 (64.7, 73.5)	65.6 (61.4, 68.8)	69.7 (64.2, 73.4)	65.4 (61.2, 68.7)
cg03707168	19.8 (14.6, 25.5)	21.1 (17.8, 24.8)	22.9 (16.9, 27.4)	23.1 (19.4, 26.9)	23.6 (18.4, 29.1)	26.1 (22.4, 30.2)	23.5 (18.2, 28.9)	25.9 (22.2, 30.0)
cg04885881	34.7 (29.4, 39.4)	37.1 (33.8, 41.2)	36.3 (31.9, 41.9)	40.9 (36.8, 44.7)	38.4 (32.9, 43.6)	43.8 (39.5, 47.8)	38.1 (32.7, 43.3)	43.6 (39.3, 47.6)
cg05575921	67.6 (59.9, 79.7)	72.01(63.9, 78.4)	76.7 (65.7, 89.6)	78.2 (70.9, 81.8)	87.0 (75.9, 91.6)	80.9 (75.9, 84.9)	86.6 (74.8, 91.5)	80.7 (75.5, 84.8)
cg07251887	51.8 (45.5, 58.3)	40.4 (37.1, 43.1)	54.5 (47.8, 60.9)	41.3 (38.4, 44.5)	56.3 (51.5, 61.4)	43.9 (40.9, 47.0)	56.2 (51.2, 61.3)	43.8 (40.7, 46.9)
cg11902777	2.2 (1.8, 2.6)	6.6 (5.7, 7.6)	2.3 (1.8, 3.1)	7.5 (6.2, 9.4)	2.8 (2.1, 3.6)	7.5 (6.4, 9.2)	2.8 (2.1, 3.6)	7.5 (6.3, 9.2)
cg21566642	42.6 (38.9, 49.1)	38.5 (34.3, 43.3)	46.2 (40.8, 55.1)	44.2 (37.5, 50.6)	52.5 (45.7, 57.7)	47.1 (41.3, 51.4)	52.0 (44.9, 57.4)	46.9 (40.9, 51.4)
cg23771366	42.9 (38.6, 49.0)	37.6 (31.4, 40.6)	46.0 (40.3, 50.0)	38.2 (33.9, 41.5)	46.7 (42.7, 50.7)	40.8 (37.4, 43.8)	46.7 (42.5, 50.7)	40.6 (37.2, 43.7)
cg23916896	15.9 (13.1, 18.8)	21.2 (18.1, 23.8)	16.8 (13.5, 19.9)	23.8 (20.1, 27.8)	18.2 (14.9, 22.2)	25.2 (21.8, 28.8)	18.1 (14.7, 22.1)	25.1 (21.6, 28.7)
cg24859433	91.4 (88.3, 92.6)	76.4 (73.9, 79.0)	92.5 (90.5, 93.7)	77.7 (75.4, 79.5)	93.1 (91.5, 94.3)	79.9 (77.4, 81.9)	93.0 (91.4, 94.3)	79.7 (77.3, 81.8)
cg25648203	85.6 (81.6, 87.9)	73.7 (70.7, 75.8)	87.2 (83.8, 89.6)	74.4 (72.2, 76.4)	88.3 (85.8, 90.1)	75.8 (73.4, 77.9)	88.2 (85.6, 90.1)	75.7 (73.3, 77.9)
cg27241845	67.3 (63.6, 72.2)	59.2 (55.0, 62.3)	69.2 (64.8, 75.6)	59.4 (55.8, 62.7)	71.9 (67.1, 75.6)	62.7 (59.2, 65.7)	71.7 (66.9, 75.6)	62.5 (58.9, 65.6)

Table C7: Expression quantitative trait methylation (eQTM) for the CpG sites that were significant for both the Strong Heart Study and the Framingham Heart Study in the mediation analysis, and the CpG sites that were significant for the SHS in the multimediator model.

DMP	DNAm gene symbol	Cancer endpoint (smoking-related variable)	N cis-eQTMs	N trans-eQTMs	Direction of association	CpG location
cg16519923	*ITGAL*	Smoking-related (current)	1	697	Inverse	Body
cg05575921[a]	*AHRR*	Lung (current), smoking-related (current, pack-years)	3	655	Inverse	Body
cg03636183	*F2RL3*	Lung (current, pack-years), smoking-related (current, pack-years)	0	347	Inverse	Body
cg03707168	*PPP1R15A*	Lung (current, pack-years), smoking-related (current, pack-years)	1	276	Inverse	Body
cg25648203	*AHRR*	Smoking-related (current, pack-years)	3	248	Inverse	Body
cg01899089	*AHRR*	Lung (current, pack-years), smoking-related (current, pack-years)	1	63	Inverse	Body
cg07251887	*RECQL5*	Lung (current, pack-years), smoking-related (current)	1	43	Inverse	TSS1500
cg23771366	*PRSS23*	Lung (current)	0	37	Inverse	TSS1500
cg11902777[a]	*AHRR*	Lung (current, pack-years)	1	24	Inverse	Body
cg23916896	*AHRR*	Lung (current, pack-years)	1	17	Inverse	Body
cg01940273	*ALPG*	Lung (current), smoking-related (current, pack-years)	0	2	Positive	Intergenic
cg19859270[a]	*GPR15*	Smoking-related (current, pack-years)	0	1	Inverse	1st Exon
cg03329539	*ALPG*	Lung (current, pack-years)	0	1	Inverse	Intergenic
cg04885881	*SRM*	Lung (current, pack-years)	0	1	Inverse	Intergenic
cg14624207	*LRP5*	Lung (current, pack-years)	0	1	Inverse	Body
cg21566642	*ALPG*	Lung (current), smoking-related (current, pack-years)	0	1	Inverse	Intergenic
cg24859433[a]	*IER3*	Lung (current, pack-years), smoking-related (current, pack-years)	0	1	Inverse	Intergenic
cg27241845	*ECEL1P2*	Lung (current, pack-years)	0	1	Inverse	Intergenic

Abbreviations: eQTM, expression quantitative trait methylation; IQR, interquartile range; DNAm, DNA methylation; N, number.

[a] Significant CpGs in the multimediator model.

Table C8: Differences in lung cancer cases per 100,000 person-years comparing current to never smokers attributable to differences in DNA methylation for each CpG ('mediated effects') using the difference of coefficients method in the Strong Heart Study.

Level of adjustment	Absolute difference in cancer cases (95 % CI)[a]	Difference in change of cancer cases attributable to DNAm (95 % CI)[b]	% of difference in cancer cases attributable to DNAm (95 % CI)
Adjusted for risk factors[c]	339.4 (227.2, 452.1)	0.0 (Ref)	0.0 (Ref)
Further adjusted for cg05575921	78.0 (-28.1, 190.9)	261.4 (171.1, 349.5)	77.0 (39.7, 100.6)
Further adjusted for cg24859433	251.1 (149.2, 353.4)	88.3 (42.4, 129.4)	26.0 (10.8, 37.6)
Further adjusted for cg11902777	301.2 (195.6, 412.6)	38.2 (17.8, 53.5)	11.3 (3.4, 16)
Further adjusted for cg0557592, cg24859433 and cg11902777	73.1 (-36.5, 186.6)	266.4 (169.6, 356.8)	78.5 (40.4, 102.3)
Adjusted for risk factors[c] and cg05575921	78.0 (-28.1, 190.9)	0.0 (Ref)	0.00 (Ref)
Further adjusted for cg24859433 and cg11902777	73.1 (-36.5, 186.6)	4.9 (-12.7, 18.2)	6.3 (-79, 71.3)
Adjusted for risk factors[c] and cg24859433	251.1 (149.2, 353.4)	0.0 (Ref)	0.00 (Ref)
Further adjusted for cg05575921 and cg11902777	73.1 (-36.5, 186.6)	178.0 (104.4, 246.2)	70.9 (14.9, 98.6)
Adjusted for risk factors[c] and cg11902777	301.2 (195.6, 412.6)	0.0 (Ref)	0.00 (Ref)
Further adjusted for cg05575921 and cg24859433	73.1 (-36.5, 186.6)	228.1 (140.8, 313.2)	75.7 (31.4, 101)

Abbreviations: CI, confidence interval; DNAm, DNA methylation; Ref, reference.

[a] Absolute change in cancer cases for current versus never smokers was calculated using additive hazards models.

[b] Effects mediated by DNA methylation were estimated using the 'difference of coefficient method' as the absolute change in cancer incidence of current versus never smokers in the model unadjusted for DNA methylation minus that absolute change in the model further adjusted for DNA methylation, expressed both in absolute terms (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were calculated using bootstrap.

[c] Age, sex, former smoking, BMI, study center (Arizona, Oklahoma or North and South Dakota) cell counts (CD8T, CD4T, NK, B cells and monocytes) and five genetic PCs.

Table C9: Differences in smoking-related cancer cases per 100,000 person-years comparing current to never smokers attributable to differences in DNA methylation for each CpG ('mediated effects') using the difference of coefficients method in the Strong Heart Study.

Level of adjustment	Absolute difference in cancer cases (95 % CI)[a]	Difference in change of cancer cases attributable to DNAm (95 % CI)[b]	% of difference in cancer cases attributable to DNAm (95 % CI)
Adjusted for risk factors[c]	436.6 (281.5, 599.2)	Ref	Ref
Further adjusted for cg19859270	296.3 (136.2, 450.6)	140.3 (65.3, 207.8)	32.7 (15.2, 48.4)
Further adjusted for cg16201146	400.2 (241.6, 555.9)	36.4 (8, 61.1)	8.5 (1.9, 14.3)
Further adjusted for cg01513913	376.7 (224.3, 532.4)	59.9 (21.3, 94.8)	14 (5, 22.2)
Further adjusted for cg01002722	429.1 (270.9, 588.7)	7.5 (-5, 18.2)	1.7 (-1.3, 4.1)
Further adjusted for cg19859270, cg16201146, cg01513913 and cg01002722	168.6 (1.1, 325.7)	268 (175.7, 362.1)	61.4 (25.9, 84.6)
Adjusted for risk factors[c] and cg19859270	296.3 (136.2, 450.6)	Ref	Ref
Further adjusted for cg16201146, cg01513913 and cg01002722	168.6 (1.1, 325.7)	127.7 (75.6, 180.7)	43.1 (-7.2, 62.7)
Adjusted for risk factors[c] and cg16201146	400.2 (241.6, 555.9)	Ref	Ref
Further adjusted for cg19859270, cg01513913 and cg01002722	168.6 (1.1, 325.7)	231.6 (144, 317.4)	57.9 (19.1, 81.5)
Adjusted for risk factors[c] and cg01513913	376.7 (224.3, 532.4)	Ref	Ref
Further adjusted for cg19859270, cg16201146 and cg01002722	168.6 (1.1, 325.7)	208.1 (122.7, 293.9)	55.2 (14.5, 78.4)
Adjusted for risk factors[c] and cg01002722	429.1 (270.9, 588.7)	Ref	Ref
Further adjusted for cg19859270, cg16201146 and cg01513913	168.6 (1.1, 325.7)		

Abbreviations: CI, confidence interval; DNAm, DNA methylation; Ref, reference.

[a] Absolute change in cancer cases for current versus never smokers was calculated using additive hazards models.

[b] Effects mediated by DNA methylation were estimated using the 'difference of coefficient method' as the absolute change in cancer incidence of current versus never smokers in the model unadjusted for DNA methylation minus that absolute change in the model further adjusted for DNA methylation, expressed both in absolute terms (difference in change

reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model.

The 95 % CIs in the table were calculated using bootstrap.

[c] Age, sex, former smoking, BMI, study center (Arizona, Oklahoma or North and South Dakota) cell counts (CD8T, CD4T, NK, B cells and monocytes) and five genetic PCs.

Table C10: Differences in lung cancer cases per 100,000 person-years comparing current to never smokers attributable to differences in DNA methylation for each CpG ('mediated effects') excluding cancer cases that happened before 1995 in the Strong Heart Study.

CpG	Difference in change of cancer cases attributable to DNAm (95 % CI)[b]	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases (95 % CI)[a]
cg05575921	218.56 (138.47, 300.13)	72.0 (45.8, 111.2)	85.12 (-25.3, 195.34)
cg14391737	110.86 (62.63, 160.75)	34.8 (19.9, 55.4)	207.62 (103.36, 311.56)
cg21566642	135.24 (71.87, 199.79)	43.8 (23.1, 73.3)	173.28 (60.33, 286.06)
cg03636183	108.97 (55.81, 163.4)	35.5 (18.5, 58.9)	197.58 (91.88, 303.03)
cg09842685	59.06 (32.24, 87.96)	18.9 (10.4, 31)	253.19 (151.46, 354.68)
cg11902777	36.52 (20.52, 54.52)	11.6 (6.5, 19)	278.41 (175.54, 381.05)
cg24859433	84.87 (44.26, 128.23)	27 (14.4, 44.1)	229.0 (126.74, 330.95)
cg03329539	62.09 (26.58, 99.38)	20.1 (8.4, 36.8)	246.69 (137.97, 355.15)
cg23916896	32.94 (15.83, 52.24)	10.6 (5.2, 18)	278.25 (176.72, 379.55)
cg01899089	45.95 (21.06, 73.12)	14.7 (7, 25.2)	265.74 (163.95, 367.3)
cg01940273	87.4 (29.68, 146.08)	28.4 (9.4, 54.5)	220.84 (102.33, 339.16)
cg17739917	57.62 (19.07, 97.02)	18.5 (6.3, 33.7)	254.24 (149.29, 358.84)
cg14624207	34.97 (14.41, 58.15)	11.3 (4.6, 20.9)	273.51 (169.36, 377.42)
cg07251887	29.37 (11.01, 50.38)	9.5 (3.6, 17.8)	279.47 (175.5, 383.19)
cg16201146	28.04 (10.3, 48.32)	8.9 (3.4, 16)	287.02 (185.21, 388.61)
cg14580211	42.87 (12.3, 75.1)	13.6 (4, 25.9)	271.23 (166.59, 375.55)
cg06521527	20.08 (6.34, 36.02)	6.4 (2.1, 12.1)	293.57 (190.83, 396.12)
cg27241845	32 (8.16, 58.19)	10.4 (2.6, 20.9)	276.0 (170.44, 381.19)
cg01513913	26.78 (6.43, 48.95)	8.7 (2, 17.8)	282.31 (176.03, 388.25)
cg24947681	23.27 (4.67, 44.12)	7.5 (1.5, 15.8)	285.96 (180.15, 391.42)
cg04885881	37.33 (4.65, 71.08)	11.9 (1.5, 24.5)	275.58 (170.22, 380.59)
cg16207944	24.16 (3.59, 46.17)	7.8 (1.2, 16.7)	285.1 (179.03, 390.82)
cg24556382	18.96 (3.86, 36.22)	6.1 (1.3, 11.9)	294.37 (192.4, 396.11)
cg23025288	18.8 (3.48, 36.14)	6.1 (1.2, 12.6)	289.46 (185.87, 392.82)
cg03707168	28.74 (3.06, 56.03)	9.3 (1, 19.3)	281.83 (177.56, 385.82)
cg21733098	22.3 (2.63, 43.5)	7.2 (0.9, 15.3)	287.08 (182.46, 391.45)
cg25799109	16.57 (2.23, 32.77)	5.3 (0.7, 11.3)	293.67 (189.89, 397.21)
cg02738868	22.3 (1.4, 44.58)	7.2 (0.5, 15.7)	288.68 (183.25, 393.77)
cg18158149	19.78 (1.79, 39.59)	6.3 (0.6, 12.9)	294.8 (192.86, 396.55)

Abbreviations: CI, confidence interval; DNAm, DNA methylation.

Models were adjusted for age, sex, former smoking, BMI, cell counts (CD8T, CD4T, NK,

B cells and monocytes), study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs.

[a] Absolute changes in cancer incidence (per 100,000 person-years) for current versus never smokers were obtained from additive hazards models.

[b] Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for the current versus never smoking comparison from the mediator model by the absolute change in cancer incidence cases for the current versus never smoking comparison (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were derived by simulation from the estimated model coefficients and covariance matrices.

Table C11: Differences in lung cancer cases per 100,000 person-years for a 10 pack-years increase attributable to differences in DNA methylation for each CpG ('mediated effects') excluding cancer cases that happened before 1995 in the Strong Heart Study.

CpG	Difference in change of cancer cases attributable to DNAm (95 % CI)[b]	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases (95 % CI)[a]
cg05575921	18.54 (9.34, 28.23)	21.5 (9.4, 48.8)	67.67 (23.12, 112.12)
cg14391737	13.06 (4.74, 21.74)	15 (5.1, 34.8)	74.2 (30.31, 117.94)
cg24859433	9.37 (4.02, 15.63)	10.6 (4.3, 23.3)	78.7 (35.85, 121.45)
cg09842685	5.71 (2.48, 9.6)	6.5 (2.7, 14.3)	81.88 (39.17, 124.51)
cg11902777	3.77 (1.6, 6.49)	4.3 (1.7, 9.4)	84.47 (41.72, 127.14)
cg21566642	9.85 (2.64, 17.4)	11.3 (2.8, 28.1)	77.46 (33.28, 121.53)
cg03636183	9.29 (2.47, 16.46)	10.6 (2.6, 26.3)	78.1 (34.12, 121.96)
cg03329539	6.62 (1.37, 12.37)	7.5 (1.5, 18.5)	81.47 (38.2, 124.61)
cg14624207	3.86 (1.08, 7.44)	4.4 (1.2, 10.7)	83.85 (40.92, 126.68)
cg23916896	2.97 (0.9, 5.72)	3.3 (1.0, 8.2)	86.27 (43.15, 129.3)
cg07251887	3.97 (0.98, 7.69)	4.5 (1.1, 11)	84.49 (41.5, 127.37)
cg01899089	3.38 (0.95, 6.64)	3.9 (1, 9.9)	83.98 (40.73, 127.14)
cg14580211	4.91 (0.26, 10.0)	5.5 (0.3, 13.7)	83.75 (41.14, 126.27)
cg06521527	2.24 (0.38, 4.76)	2.6 (0.4, 6.4)	85.67 (42.99, 128.26)
cg27241845	3.16 (0.26, 6.87)	3.6 (0.3, 10)	84.89 (41.64, 128.0)
cg24947681	2.89 (0.17, 6.3)	3.3 (0.2, 8.6)	85.91 (42.98, 128.76)

Abbreviations: CI, confidence interval; DNAm, DNA methylation.
Models were adjusted for age, sex, current smoking, BMI and cell counts (CD8T, CD4T, NK, B cells and monocytes), study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs.

[a] Absolute changes in cancer incidence (per 100,000 person-years) for a 10 pack-years change were obtained from additive hazards models.

[b] Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for a 10 pack-years increase from the mediator model by the absolute change in cancer incidence cases for a 10 pack-years increase (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were derived by simulation from the estimated model coefficients and covariance matrices.

Table C12: Differences in smoking-related cancer cases per 100,000 person-years comparing current to never smokers attributable to differences in DNA methylation for each CpG ('mediated effects') excluding cancer cases that happened before 1995 in the Strong Heart Study.

CpG	Difference in change of cancer cases attributable to DNAm (95 % CI)[b]	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases (95 % CI)[a]
cg05575921	282.73 (170.9, 396.48)	69.6 (40.7, 117.2)	123.3 (-48.74, 295.03)
cg14391737	166.05 (96.65, 237.95)	38.9 (22, 65.3)	260.58 (102.54, 418.26)
cg21566642	206.4 (111.42, 303.21)	50.1 (26.1, 88.1)	205.32 (33.97, 376.37)
cg03636183	171.72 (92.6, 252.86)	42.0 (22.0, 73.4)	236.97 (74.94, 398.71)
cg25845814	97.56 (49.11, 148.75)	23.6 (11.8, 41.2)	315.23 (164.35, 465.64)
cg18110140	92.88 (42.72, 145.66)	22.4 (10.2, 40.6)	321.59 (166.55, 476.2)
cg19859270	107.94 (46.21, 172.19)	25.7 (11.3, 45.5)	312.19 (159.96, 463.96)
cg01940273	124.63 (44.64, 206.04)	30.3 (10.5, 59.8)	287.03 (114.56, 459.16)
cg01513913	51.64 (22.1, 84.51)	12.5 (5.2, 23.8)	360.88 (207.23, 514.04)
cg25648203	92.74 (33.54, 153.81)	22.6 (8.2, 43.1)	316.91 (159.49, 473.96)
cg21911711	88.19 (30.69, 147.82)	21 (7.4, 39.5)	331.57 (176.11, 486.6)
cg15310518	49.65 (19.85, 82.52)	12.1 (4.7, 23.4)	360.82 (207.11, 514.05)
cg01899089	53.68 (19.74, 90.37)	12.9 (4.7, 24.6)	363.2 (210.2, 515.79)
cg24859433	73.32 (24.65, 124.5)	17.5 (6.0, 33.2)	345.8 (192.0, 499.1)
cg05934812	37.77 (13.5, 66.31)	8.9 (3.2, 16.9)	387.43 (236.03, 538.58)
cg21322436	43.02 (11.71, 76.59)	10.4 (2.8, 21.1)	372.21 (218.19, 525.74)
cg02738868	45.58 (12.39, 81.48)	10.9 (3, 22.1)	370.74 (216.72, 524.3)
cg19885130	41.28 (11.33, 74.06)	10 (2.9, 19.3)	372.93 (223.47, 522.11)
cg09842685	48.37 (11.16, 87.51)	11.6 (2.6, 24.4)	367.86 (211.08, 524.17)
cg09338374	34.69 (9.02, 63.11)	8.3 (2.2, 16.8)	385.43 (232.29, 538.23)
cg10258214	33.8 (5.42, 64.25)	8.2 (1.3, 17.5)	378.59 (225.72, 531.1)
cg12615852	31.48 (4.53, 60.7)	7.6 (1.1, 16.5)	382.59 (229.26, 535.52)
cg03707168	38.18 (4.62, 73.87)	9.2 (1.1, 19.7)	377.71 (224.58, 530.42)
cg07943658	35.31 (4.31, 68.71)	8.5 (1.1, 18.5)	381.03 (226.93, 534.87)
cg07251887	29.29 (4.08, 57.39)	7.1 (1.0, 15.0)	384.94 (233.81, 535.74)
cg20174472	29.71 (2.01, 59.68)	7.1 (0.5, 16.1)	387.57 (233.09, 541.57)
cg27271698	31.79 (1.04, 64.31)	7.7 (0.3, 17.4)	381.04 (227.91, 533.78)
cg16201146	26.18 (2.06, 53.04)	6.2 (0.5, 13.8)	393.32 (241.12, 545.18)
cg20295214	24.14 (1.83, 49.52)	5.7 (0.5, 12.6)	396.62 (245.4, 547.59)
cg16519923	21.75 (1.8, 44.76)	5.2 (0.4, 11.7)	399.4 (246.94, 551.53)
cg07267541	25.73 (0.72, 53.33)	6.2 (0.2, 14.4)	390.3 (235.71, 544.43)
cg01002722	13.28 (1.81, 28.92)	3.1 (0.5, 7.1)	410.17 (257.92, 562.02)
cg12409728	15.04 (0.8, 32.91)	3.6 (0.2, 8.8)	398.36 (245.93, 550.51)

Abbreviations: CI, confidence interval; DNAm, DNA methylation.

Models were adjusted for age, sex, former smoking, BMI and cell counts (CD8T, CD4T,

NK, B cells and monocytes), study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs.

[a] Absolute changes in cancer incidence (per 100,000 person-years) for current versus never smokers were obtained from additive hazards models.

[b] Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for the current versus never smoking comparison from the mediator model by the absolute change in cancer incidence cases for the current versus never smoking comparison (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were derived by simulation from the estimated model coefficients and covariance matrices.

Table C13: Differences in smoking-related cancer cases per 100,000 person-years for a 10 pack-years increase attributable to differences in DNA methylation for each CpG ('mediated effects') excluding cancer cases that happened before 1995 in the Strong Heart Study.

CpG	Difference in change of cancer cases attributable to DNAm (95 % CI)[b]	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases (95 % CI)[a]
cg05575921	23.46 (10.73, 36.83)	18.6 (7.7, 39.8)	102.58 (43.27, 161.75)
cg14391737	20.99 (8.53, 34.01)	16.5 (6.2, 35.8)	106.06 (46.88, 165.09)
cg03636183	15.69 (5.77, 26.21)	12.4 (4.3, 26.7)	111.27 (53.39, 168.96)
cg21566642	16.6 (5.49, 28.26)	13.1 (4.0, 29.8)	110.47 (51.07, 169.74)
cg25845814	8.47 (3.23, 14.72)	6.7 (2.4, 14.5)	118.78 (61.62, 175.8)
cg19859270	11.65 (3.14, 20.92)	9.0 (2.4, 19.8)	117.27 (60.12, 174.28)
cg18110140	9.39 (2.66, 16.91)	7.4 (2, 16.6)	118.17 (60.61, 175.56)
cg01513913	6.14 (1.78, 11.49)	4.8 (1.3, 11.3)	122.04 (64.26, 179.63)
cg15310518	5.77 (1.37, 11.07)	4.5 (1.0, 10.9)	122.47 (64.54, 180.22)
cg21911711	9.02 (1.08, 17.56)	7.0 (0.9, 16.2)	119.71 (62.56, 176.75)
cg25648203	6.9 (0.91, 13.6)	5.4 (0.7, 13.2)	121.33 (63.62, 178.87)
cg24859433	7.07 (0.84, 14.12)	5.5 (0.6, 13.6)	121.39 (63.41, 179.19)
cg05934812	4.8 (1.12, 9.68)	3.7 (0.9, 8.9)	123.62 (66.44, 180.66)
cg01899089	3.85 (0.78, 8.02)	3.0 (0.6, 7.6)	123.75 (66.27, 181.1)
cg21322436	3.7 (0.37, 7.94)	2.9 (0.3, 7.7)	125.22 (67.18, 183.09)
cg19885130	4.08 (0.22, 8.84)	3.2 (0.2, 8.4)	123.6 (65.68, 181.38)
cg09338374	3.13 (0.19, 7.02)	2.4 (0.1, 6.7)	125.11 (67.2, 182.93)
cg02738868	3.3 (0.17, 7.49)	2.6 (0.1, 7.1)	123.85 (66.18, 181.4)

Models were adjusted for age, sex, current smoking, BMI and cell counts (CD8T, CD4T, NK, B cells and monocytes), study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs.

Absolute changes in cancer incidence (per 100,000 person-years) for a 10 pack-years change were obtained from additive hazards models.

Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for a 10 pack-years increase from the mediator model by the absolute change in cancer incidence cases for a 10 pack-years increase (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were derived by simulation from the estimated model coefficients and covariance matrices.

Table C14: Differences in smoking-related cancer cases (excluding liver cancer) per 100,000 person-years comparing current to never smokers attributable to changes in DNA methylation for each CpG ('mediated effects') in the Strong Heart Study.

CpG	Difference in change of cancer cases attributable to DNAm (95 % CI)[b]	Percentage of difference in cancer cases attributable to DNAm (95 % CI)	Absolute difference in cancer cases (95 % CI)[a]
cg05575921	278.7 (170.1, 388.6)	65.9 (38.9, 109.6)	143.9 (-28.5, 315.9)
cg14391737	212.5 (135.0, 293.1)	47.2 (29.1, 75.7)	237.7 (77.3, 397.4)
cg21566642	219.2 (124.3, 315.8)	51.0 (28.0, 87.0)	210.7 (39.3, 381.4)
cg03636183	176.8 (95.9, 259.6)	41.5 (22.1, 70.8)	249.5 (87.4, 411.1)
cg25845814	100.9 (51.4, 153.1)	23.5 (11.9, 40.3)	329.4 (176.7, 481.7)
cg18110140	97.3 (46.2, 151.2)	22.5 (10.6, 40.1)	334.6 (177.7, 490.9)
cg01940273	143.7 (63.5, 225.4)	33.5 (14.3, 62)	285.4 (114.4, 456.0)
cg19859270	118.1 (50.7, 187.9)	26.8 (11.9, 46.4)	323.0 (169.4, 476.2)
cg21911711	86.7 (35.3, 140.2)	19.9 (8.1, 36.2)	348.7 (192.8, 504.1)
cg25648203	96.9 (38.5, 157.2)	22.7 (9.0, 41.9)	330.6 (172.1, 488.5)
cg01899089	62.2 (26.9, 100.5)	14.3 (6.1, 26.4)	371.6 (215.9, 526.9)
cg24859433	71.4 (27.7, 117.5)	16.4 (6.3, 30.5)	363.8 (207.6, 519.5)
cg01513913	52.8 (21.8, 87.2)	12.3 (4.9, 23.2)	377.2 (221.8, 532.3)
cg15310518	49.9 (18.8, 84.0)	11.7 (4.3, 22.7)	378.2 (221.8, 534.2)
cg09842685	59.6 (19.99, 101.32)	13.7 (4.5, 26.6)	374.9 (218.1, 531.4)
cg02738868	53.9 (19.39, 91.58)	12.4 (4.4, 24)	379.7 (223.6, 535.4)
cg05934812	40.9 (15.91, 70.23)	9.2 (3.7, 17.2)	401.8 (248.3, 555.1)
cg07943658	49.5 (14.91, 87.26)	11.4 (3.4, 22.6)	384.7 (227.9, 541.3)
cg00073090	39.7 (10.39, 71.73)	9.2 (2.4, 18.5)	391.3 (237.4, 544.9)
cg16201146	33.1 (9.26, 60.32)	7.6 (2.2, 15.1)	403.5 (249.6, 557.1)
cg03707168	49.0 (11.12, 89.77)	11.3 (2.6, 22.5)	386.3 (232.3, 540.1)
cg19885130	39.1 (8.81, 71.94)	9.1 (2.2, 17.7)	392.3 (241.9, 542.5)
cg12615852	37.6 (7.78, 69.97)	8.7 (1.8, 17.9)	394.2 (239.7, 548.3)
cg20174472	36.0 (6.24, 68.5)	8.3 (1.5, 17.6)	397.8 (242.5, 552.8)
cg27271698	36.6 (5.41, 69.73)	8.5 (1.3, 18.2)	393.1 (237.7, 548.1)
cg07251887	31.9 (5.82, 61.13)	7.4 (1.4, 15.4)	399.4 (245.8, 552.7)
cg10258214	34.6 (4.84, 66.54)	8.1 (1.1, 17.2)	395.4 (240.4, 549.9)
cg25189904	52.2 (3.98, 101.93)	12.1 (1, 26.1)	378.5 (221.6, 535.1)
cg07267541	28.3 (3.39, 55.98)	6.5 (0.8, 14.4)	404.6 (249.2, 559.7)
cg23025288	26.3 (2.13, 53.09)	6.1 (0.5, 13.6)	403.9 (249.3, 558.3)
cg16519923	21.7 (1.38, 45.04)	5 (0.3, 11.2)	416.5 (261.8, 570.9)
cg18158149	26.9 (0.04, 56.19)	6.1 (0, 13.6)	411.3 (258.1, 564.1)
cg01002722	13.4 (2.01, 29.02)	3 (0.5, 6.9)	427.3 (273.0, 581.2)

Models were adjusted for age, sex, former smoking, BMI and cell counts (CD8T, CD4T, NK, B cells and monocytes), study center (Arizona, Oklahoma or North and South Dakota) and five genetic PCs.

Absolute changes in cancer incidence (per 100,000 person-years) for a 10 pack-years change were obtained from additive hazards models.

Effects mediated by DNA methylation were estimated with the 'product of coefficients method' that multiplies the coefficient for the mean change in DNA methylation for a 10 pack-years increase from the mediator model by the absolute change in cancer incidence cases for a 10 pack-years increase (difference in change reflecting the number of attributable cancer cases per 100,000 person-years) and relative to the adjusted changes in cancer cases before adding DNA methylation to the model. The 95 % CIs in the table were derived by simulation from the estimated model coefficients and covariance matrices.