UNIVERSITA' DEGLI STUDI DI TORINO
Dipartimento di Scienze Mediche

Scuola di Dottorato in Scienze della Vita e della Salute
Programma in Scienze Biomediche e Oncologia Umana
Indirizzo di Genetica Umana

Ciclo XXIX

DNA methylation as a potential mediator of environmental risks in cancer

Candidato: Francesca FASANELLI

TUTOR: Prof LORENZO RICHIARDI

COORDINATORE DEL DOTTORATO:	COORDINATORE DI INDIRIZZO:
Prof Emilio HIRSCH	Prof Giuseppe MATULLO

ANNI ACCADEMICI: 2014-2017

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA: MED/03

Summary

Introduction 3

1. Background 4

1.1 DNA methylation 4

1.2 DNA methylation and cancer 6

1.3 Environment and DNA methylation 8

1.4 Meet-in-the-middle approach 9

1.5 Mediation analysis in molecular epidemiology 10

2. Present studies 15

2.1 General aims 15

2.2 Summary of the projects 15

2.2.1 Project 1: Smoking, DNA methylation and lung cancer (Chapter 3) 15

2.2.2 Project 2: Mediterranean diet, DNA methylation and colon cancer (Chapter 4) 17

2.2.3 Project 3: methodological paper (Appendix) 19

3. Project 1: Smoking, DNA methylation and lung cancer 22

3.1 Introduction 22

3.2 Results 23

3.3 Discussion 28

3.4 Methods 30

3.5 Tables 40

3.6 Figures 44

3.7 Supplementary material 46

4. Project 2: Mediterranean Diet, DNA methylation and colon cancer 51

4.1 Introduction 51

4.2 Materials and Methods 52

4.3 Results 58

4.4 Discussion 60

4.5 Tables 63

4.6 Figures 70

4.7 Supplementary material 73

5. Conclusions and future perspectives 81

PhD final report 83

Appendix 89

Introduction

Aberrant DNA methylation is an epigenetic modification involved in early stages of tumorigenesis. Methylation represents an adaptive response to external stimuli leading to modulation of gene expression in a temporary or permanent way and to alteration of the functionality of proteins that are part of the methylation machinery. Aberrant methylation induced by long lasting environmental exposures may persist for a long time, providing further support of a possible causal involvement of DNA methylation in carcinogenesis.

The main aim of the present thesis is to study the role of DNA methylation as a potential mediator of the carcinogenic process triggered by specific environmental exposures. Furthermore another goal is to develope new methods for mediation analysis that can also be applied to molecular data.

The thesis is organized as follows. In Chapter 1 there is a concise overview of the main aspects of epigenetics focusing in particular on DNA methylation and its relationship with cancer and environment. A general introduction regarding the meet-in-the-middle approach and the use of mediation analysis in molecular epidemiology is also given. Chapter 2 provides a summary and a short commentary of the three projects carried out during my PhD program.

The detailed descriptions of the two main projects (smoking, DNA methylation and lung cancer, project 1; Mediterranean Diet, DNA methylation and colon cancer, project 2) are reported respectively in Chapter 3 (*Fasanelli et al.* 2015) and Chapter 4 (unpublished). Chapter 5 includes the conclusions and the future perspectives of the work presented. The third project of my PhD program, which responds to the goal of developing new statistical methodologies for mediation analysis, is reported in the Appendix (unpublished).

1. Background

1.1 DNA methylation

It is common that organisms with the same genetics, such as monozygotic twins, have distinct phenotypes and degrees of disease penetrance [1]. This can be partially explained by epigenetics. The term of "epigenetics" (literally "over" or "upon" genetics) has been introduced for the first time in 1942 by Conrad Waddington [2] in order to describe events that could not be wholly explained by traditional genetics. He subsequently defined epigenetics as ''the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being" [3]. This definition initially referred to the role of epigenetics in embryonic development and it has been adjusted and modified several times. Nowadays epigenetics indicates those chemical modifications that lead to the regulation of gene expression and genome function, while the underlying DNA sequence remains intact [4]. Some examples of relevant epigenetic mechanisms are methylation of DNA bases, histone modifications, non-coding RNA and nucleosome remodelling and positioning. The complement of these modifications, collectively referred to as the epigenome, provides a mechanism for cellular diversity by regulating what genetic information can be accessed by the transcriptional machinery [5].

The major form of epigenetic information in mammalian cells is DNA methylation. DNA methylation is vital for normal development of mammals because of the key role it plays in processes such as genomic imprinting [6], X-chromosome inactivation [7], and silencing of transposable elements [8]. It involves the covalent addition of a methyl group ($-CH_3$) to the five position of the cytosine ring (forming 5-methylcytosine). This enzymatic reaction is performed by a number of enzymes known as DNA methyltransferases (DNMTs) and it almost exclusively occurs in cytosines located 5 prime to the guanine base (commonly known as CpG dinucleotides, where the intervening 'p' represents the phosphodiester bond linking cytosine- and guanine-containing nucleotides [6]). The frequency of CpG dinucleotides in humans is 1%, less than one-quarter of the expected frequency based on the GC content in the human genome; this underrepresentation is due to the inherent mutability of methylated cytosine.

In humans, approximately 70% of all CpG sites are methylated, except when they are part of a CpG island, which are usually unmethylated [10] (unless the CpG island is located on the inactive X chromosome, or near imprinted genes [11]). CpG islands are dense clusters of CpG dinucleotides of at least 200bp, with a CG percentage that is greater than 50%, and with an observed-to-expected CpG ratio that is greater than 60% [12]. There are about 29,000 CpG islands in the human genome, and the majority of promoters and/or first exons of genes reside within CpG islands [13]: in particular, the promoters for housekeeping genes are often embedded in CpG islands, as well as a proportion of tissue-specific genes, tumor suppressor and developmental regulator genes.

When these promoter or exon CpG islands are (hyper)methylated, gene expression is usually inhibited [14], [15]. Recent work, however, shows it is mainly methylation of CpGs at the CpG island shores (sequences up to 2kb on either side of the CpG island) that influences gene expression, rather than that at the core of the CpG islands themselves [16]. In addition, it has been reported that even CpGs in intragenic and intergenic regions are associated with promoter function [17],[18], underlying the relationship between DNA methylation of CpG islands and transcription [19].

DNA methylation exerts its biological function of transcription repressor for example interfering with transcription factor binding. In fact the presence of methyl groups bound to the cytosines can physically impede the binding of transcription factors to the gene promoter and, hence, directly interfere with gene activation. A number of transcription factors recognize GC-rich sequence motifs and are unable to bind DNA when there are methylated CpG sites [20]. A second mode of repression involves proteins that are attracted to rather than repelled by methyl-CpG. These proteins, called methyl-CpG-binding domain proteins (MBDs) are characterized by a DNA-binding motifs able to specifically recognize and bind only methylated CpGs. In addition, the MBD proteins can in turns recruit co-repressor complexes leading to the formation of silenced states of chromatin that ensures a stable repression of gene transcription [20].

1.2 DNA methylation and cancer

Given the critical role of DNA methylation in gene expression, it seems obvious that errors in methylation could give rise to a number of devastating consequences, including various diseases. To date, a large amount of research on DNA methylation and disease has focused on cancer. Cancer epigenome is characterized by important epigenetic changes which result in global dysregulation of gene expression profiles and lead to the development and progression of disease states [21]. These changes can induce inappropriate silencing of tumour suppressor genes and/or activation of oncogenes independently or in conjunction with deleterious genetic mutations or deletions. For this reason, Knudson proposed that they could represent the second hit required for cancer initiation (the "two-hit" hypothesis [22]). Above all the epigenetic abnormalities in the cancer cell, aberrant DNA methylation is deeply involved both in cancer development and progression because DNA methylation pattern is accurately transmitted to daughter cells after cell division and so inherited with a high fidelity in somatic cells [5].

In human cancer deregulation of DNA methylation is found in at least two forms: *i)* the gene promoter-associated (mostly CpG island-specific) hypermethylation and *ii)* the overall loss of 5-methyl-cytosine (global hypomethylation) [23].

Hypermethylation of the CpG islands in the promoter regions of tumour suppressor genes was found for the first time in 1993 in [24]. This increased methylation level was reported to be a way of inactivation alternative to genetic alterations [25], thus contributing to tumour development and progression. Besides this direct silencing of tumour suppressor genes, promoter hypermethylation may also lead to inactivation of other cancer-associated genes (including those involved in cell cycle regulation, DNA repair, apoptosis, cell adhesion and angiogenesis). Consequently silencing of these genes by CpG promoter hypermethylation in cancer cells is known to affect a wide range of cellular processes and several cellular pathways.

Although the role of DNA hypermethylation in gene silencing is relatively well understood, much less is known about the importance of aberrant DNA hypomethylation in cancer. It was hypothesized that DNA hypomethylation could lead to the activation and expression of classical oncogenes, but evidence suggests a greater involvement in the activation of developmentally critical genes or genes associated with tumour invasion or metastasis. Loss of global methylation was mainly identified in repeated elements, such as tandem repeats (satellite DNA, minisatellite, and microsatellite) and interspersed repeats (among which short interspersed nuclear elements (SINEs), or long interspersed nuclear elements (LINEs)). For example hypomethylation of normally methylated LINE-1 and Alu repeats has been found to be associated with several cancers, among which breast, ovarian, and colorectal cancers [26]-[31]. This is not surprising, since these DNA elements are highly abundant and comprise most of the CpG islands that are normally methylated in healthy somatic tissues. Other studies suggest that hypomethylation is not only limited to repetitive areas, but also occurs in gene regions [32]-[35]. In a recent study Neri et al. reveals a new function of intragenic DNA methylation that protects the gene body from spurious RNA polymerase II entry, confirming the causal role of global hypomethylation in cancer [36].

Most DNA methylation studies have compared tumour tissue to healthy tissue from the same patients. This approach can lead to the identification of methylation markers that are useful for the sensitive detection of disease or markers associated with disease progression. However these methylation markers may be influenced by disease processes (problem of reverse causation) and they are not usable for risk assessment.

If methylation abnormalities arise early in normal tissues leading to systemic or regional epigenetic defects, then a comparison between histologically normal tissues from cancer patients and healthy controls could lead to the identification of methylation markers that are useful in risk assessment. The problem is that histologically normal tissues from control individuals are difficult to obtain.

An important finding of *Cui et al.* in 1998 opened the way to an exciting potential new approach to risk assessment [37]. In their work they provided the first indication that loss of imprinting might represent a systemic defect that is present in blood cells of individuals with colorectal cancer compared to controls. This suggested the possibility of using blood samples to measure DNA methylation differences between cases and controls years before the cancer onset. The goal is to identify biomarkers derived from peripheral blood for predicting disease risk avoiding the problem of reverse causality. The benefit is that the risk of developing cancer could, potentially, be evaluated in individuals, and, therefore, timely use of the appropriate preventative measures and increased surveillance would be possible.

1.3 Environment and DNA methylation

The term "environment" refers to all the external exposures that include ambient pollution and lifestyle. In particular the concept of "lifestyle" includes different factors such as nutrition, behavior, stress, physical activity, working habits, smoking and alcohol consumption.

Epigenetic studies showed inverse association between global methylation and exposure to air pollution [38], [39]. Furthermore increasing evidence shows that lifestyle factors may influence epigenetic mechanisms, such as DNA methylation, histone acetylation and microRNA expression [40]. The result is a deregulation of key cellular processes and a promotion of oncogenic transformation.

Smoking status and diet are two lifestyle factors that have been shown to affect DNA methylation in many studies. Tobacco smoke contains a complex mixture of organic and inorganic chemicals, many of which have carcinogenic, pro-inflammatory and proaterogenic properties. A lot of recent studies have identified smoking-associated blood DNA methylation biomarkers using the Illumina Infinium HumanMethylation 450K BeadChip array in cord blood and adult blood [41]-[45]. A complete summary of the smoking-associated DNA methylation alterations including novel regionally altered coding and noncoding genes can be found in [46]. In this paper, the authors observed a marked reversibility of methylation changes after smoking cessation and some genes that remained differentially methylated decades after cessation. They also revealed that prediagnostic smoking-related epigenetic alterations in human blood cells are reversible after smoking cessation, consistent with the known cancer risk reduction after smoking cessation.

Multiple investigations have examined a possible role for nutrition in modifying the pattern of DNA methylation either at the global scale or at locus-specific sites [47]-[51]. There are different possible ways through which nutrition influences patterns of DNA methylation. First it gives substrates necessary for DNA methylation such as methionine, folate, choline, betaine and vitamins B_2, B_6 and B_{12}. Second it provides cofactors modulating the enzymatic activity of DNMTs, for example by changing the intracellular concentration of S-adenosylmethionine. Third, it modifies the activity of the enzymes regulating the one-carbon cycle. Importantly, all three mechanisms are mutually compatible and may operate together in time [52]. There is also evidence that an isocaloric balanced diversified diet has a stabilizing effect on the basal patterns of DNA methylation [53].

Other environmental factors that have been found to modify epigenetic patterns are environmental pollutants (such as arsenic, aromatic hydrocarbons and other organic pollutants), obesity, physical activity, alcohol consumption, psychological stress, and working on night shifts [40]. The detailed analysis of these factors does not fall within the scope of the present thesis.

1.4 Meet-in-the-middle approach

Since epidemiological studies are often aimed at assessing the effect of an exposure on disease risk, in order to implement health interventions it is important to verify the causality of this (risk) factor. Vineis and Perera in [54] described an innovative approach (known as “meet-in-the-middle approach”) with the goal to identify the overlap between markers of exposure and predictive markers of disease outcome. The underlying idea is that the finding that preclinical biomarkers related to particular exposures are also modified in certain subclasses of disease would strengthen causal links between these exposures and the disease. The approach is based on a combination, within a prospective study, of a prospective search for biomarkers which are modified in subjects who eventually go on to develop disease and a retrospective search for links of such biomarkers to past environmental exposures. The approach includes as three steps: *i)* an investigation into the association between the exposure and the disease, *ii)* an assessment of the relationship between the disease outcome and the candidate biomarkers and *iii)* a final assessment of the relationship between the exposure and the predictive biomarkers identified at point *ii*. Inference of a causal relationship between exposure and disease is strengthened if associations are documented for each of the three key relationships. Furthermore the finding of an intermediate biomarker has potential to open new avenues for prevention.

In this thesis the focus will be on smoking status and Mediterranean diet as exposures and lung cancer and colon cancer as outcomes. DNA methylation will be the candidate epigenetic biomarker. DNA methylation is a good candidate biomarker for this approach because it is related both to cancer (Section 1.2) and lifestyle exposures (Section 1.3) as required by steps *ii)* and *iii)* of the meet-in-the-middle approach.

1.5 Mediation analysis in molecular epidemiology

The phenomenon whereby a cause affects an intermediate and the change in the intermediate goes on to affect an outcome is what is generally referred to as the phenomenon of “mediation” [55]. Mediation analysis is a set of techniques by which a researcher assesses what proportion of the effect of an exposure on an outcome is operating through a particular intermediate (indirect effect) and what proportion might be through other mechanisms (direct effect).

One of the most widely cited approaches for evaluating mediation in an epidemiological setting is that originally developed by Baron and Kenny in 1986 [56]. This widely implemented approach is known to be problematic because it is highly dependent on a number of strong assumptions, the measurement characteristics of the variables and on reliable identification of causal effects. To overcome these limitations, further methods based on counterfactuals have been developed [57]-

[61]. These new methods offer more flexibility, but they require strong assumptions as the traditional ones (for example no measurement error and no unmeasured confounding).

These mediation approaches (used in conventional epidemiology) have been adapted in molecular epidemiology to understanding the role of molecular intermediates [62]. In this context mediation analysis allows to quantify the magnitude of the indirect effect of the exposure on the outcome through the specific molecular mediators identified for example by the meet-in-the-middle approach. From this point view, mediation analysis can be seen as the fourth step of the meet-in-the-middle approach because it provides the missing information about the magnitude of the direct/indirect effects.

It is important to remember that the molecular intermediates have some limitations that may affect the results of mediation analysis. First they are affected by measurement error, second they may be influenced by both endogenous and exogenous factors and by disease processes causing problems such as confounding, bias and reverse causation. These aspects have to be kept in mind to avoid incorrect conclusions regarding causal effects.

References

1. Wong AH, Gottesman II, Petronis A. Phenotypic differences in genetically identical organisms: the epigenetic perspective. Hum Mol Genet. 2005 Apr 15;14 Spec No 1:R11-8.
2. Waddington CH, The epigenotype, Endeavour, 1, 18-20, 1942.
3. Waddington CH. The epigenotype. 1942. Int J Epidemiol 2012;41:10–13.
4. Azad N, Zahnow CA, Rudin CM, Baylin SB. The Future of Epigenetic Therapy in Solid Tumours--Lessons from the Past. Nat Rev Clin Oncol. 2013 May;10(5):256-66. doi: 10.1038/nrclinonc.2013.42. Epub 2013 Apr 2.
5. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis 2010;31(1):27-36.
6. Reik W, Walter J. Genomic imprinting: parental influence on the genome. Nat Rev Genet 2001;2:21-32
7. Riggs AD. X chromosome inactivation, differentiation, and DNA methylation revisited, with a tribute to Susumu Ohno. Cytogenet Genome Res 2002;99:17-24
8. Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene 2008;27:404-408.
9. Rottach A, Leonhardt H, Spada F. DNA methylation-mediated epigenetic control. J Cell Biochem 2009;108:43–51.
10. Ndlovu MN, Denis H, Fuks F. Exposing the DNA methylome iceberg. Trends Biochem Sci 2011;36:381-387.
11. Riggs AD, Pfeifer GP. X-chromosome inactivation and cell memory. Trends Genet 1992;8:169-174.
12. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol 1987;196:261-282.
13. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 2006;103:1412-1417.
14. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev 2002;16:6-21.
15. Brenet F, Moh M, Funk P et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. PLoS One 2011;6:e14524.
16. Irizarry RA, Ladd-Acosta C, Wen B et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009;41:178-186.
17. Maunakea AK, Nagarajan RP, Bilenky M et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010;466:253-257.
18. Illingworth RS, Bird AP. CpG islands--'a rough guide'. FEBS Lett 2009;583:1713-1720.
19. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev 2011;25:1010-1022.
20. Li and Zhang, 2013 Li E, Zhang Y. DNA methylation in mammals. Cold Spring Harb Perspect Biol 2014;6(5):a019133.
21. Peltomäki P. Mutations and epimutations in the origin of cancer. Exp Cell Res 2012;318(4):299-310.
22. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 1971;68(4):820-3.
23. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004 Feb;4(2):143-53. PMID: 14732866.
24. Ohtani-Fujita N, Fujita T, Aoike A, et al. CpG Methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene. Oncogene. 1993;8:1063–7.
25. Garinis GA, Patrinos GP, Spanakis NE, Menounos PG. DNA hypermethylation: when tumour suppressor genes go silent. Hum Genet 2002;111:115-127.

26. Chalitchagorn K, Shuangshoti S, Hourpai N et al. Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. Oncogene 2004;23:8841-8846.
27. Choi JY, James SR, Link PA et al. Association between global DNA hypomethylation in leukocytes and risk of breast cancer. Carcinogenesis 2009;30:1889-1897.
28. Menendez L, Benigno BB, McDonald JF. L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. Mol Cancer 2004;3:12.
29. Suter CM, Martin DI, Ward RL. Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue. Int J Colorectal Dis 2004;19:95-101.
30. Weisenberger DJ, Campan M, Long TI et al. Analysis of repetitive element DNA methylation by MethyLight. Nucleic Acids Res 2005;33:6823-6836.
31. Kitkumthorn N, Mutirangura A. Long interspersed nuclear element-1 hypomethylation in cancer: biology and clinical applications. 2 2, 315-330. 2011.
32. Grunau C, Brun ME, Rivals I et al. BAGE hypomethylation, a new epigenetic biomarker for colon cancer detection. Cancer Epidemiol Biomarkers Prev 2008;17:1374-1379.
33. Lindsey JC, Lusher ME, Anderton JA, Gilbertson RJ, Ellison DW, Clifford SC. Epigenetic deregulation of multiple S100 gene family members by differential hypomethylation and hypermethylation events in medulloblastoma. Br J Cancer 2007;97:267-274.
34. Wasson GR, McGlynn AP, McNulty H et al. Global DNA and p53 region-specific hypomethylation in human colonic cells is induced by folate depletion and reversed by folate supplementation. J Nutr 2006;136:2748-2753.
35. Ruike Y, Imanaka Y, Sato F, Shimizu K, Tsujimoto G. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. BMC Genomics 2010;11:137.
36. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, Maldotti M, Anselmi F, Oliviero S. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 2017 Mar 2;543(7643):72-77. doi: 10.1038/nature21373. Epub 2017 Feb 22.
37. Cui, H., Horon, I. L., Ohlsson, R., Hamilton, S. R. & Feinberg, A. P. Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability. Nature Med. 4, 1276–1280 (1998).
38. Baccarelli A, Wright RO, Bollati V, Tarantini L, Litonjua AA, Suh HH, Zanobetti A, Sparrow D, Vokonas PS, Schwartz J. Rapid DNA methylation changes after exposure to traffic particles. Am J Respir Crit Care Med. 2009 Apr 1;179(7):572-8. doi: 10.1164/rccm.200807-1097OC. Epub 2009 Jan 8.
39. Sanchez-Guerra M, Zheng Y, Osorio-Yanez C et al. Effects of particulate matter exposure on blood 5-hydroxymethylation: results from the Beijing truck driver air pollution study. Epigenetics. 2015;10(7):633-42. doi: 10.1080/15592294.2015.1050174.
40. Alegría-Torres JA, Baccarelli A, Bollati V. Epigenetics and lifestyle. Epigenomics. 2011 Jun;3(3):267-77. doi: 10.2217/epi.11.22.
41. Joubert BR, Haberg SE, Nilsen RM et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ. Health Perspect. 120(10), 1425–1431 (2012).
42. Shenker NS, Polidoro S, Van Veldhoven K et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum. Mol. Genet. 22(5), 843–851 (2013).
43. Zeilinger S, Kuhnel B, Klopp N et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS ONE 8(5), e63812 (2013).
44. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environ. Health Perspect. 122(7), 673–678 (2014).
45. Markunas CA, Xu Z, Harlid S et al. Identification of DNA methylation changes in newborns

related to maternal smoking during pregnancy. Environ. Health Perspect. 122(10), 1147–1153 (2014).

46. Ambatipudi S, Cuenin C, Hernandez-Vargas H et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016 May;8(5):599-618. doi: 10.2217/epi-2016-0001. Epub 2016 Feb 11.
47. Vucetic Z, Kimmel J, Totoki K, Hollenbeck E, Reyes TM. Maternal high-fat diet alters methylation and gene expression of dopamine and opioid-related genes. Endocrinology 2010;151:4756–64.
48. Bogdarina I, Haase A, Langley-Evans S, Clark AJL. Glucocorticoid effects on the programming of AT1B angiotensin receptor gene methylation and expression in the rat. Plos One 2010;5:e9237.
49. Jousse C, Parry L, Lambert-Langlais S, Maurin AC, Averous J, Bruhat A, et al. Perinatal undernutrition affects the methylation and expression of the leptin gene in adults: implication for the understanding of metabolic syndrome. FASEB J 2011;25:3271–8.
50. Dudley KJ, Sloboda DM, Connor KL, Beltrand J, Vickers MH. Offspring of mothers fed a high fat diet display hepatic cell cycle inhibition and associated changes in gene expression and DNA methylation. Plos One 2011;6:e21662.
51. Altmann S, Murani E, Schwerin M, Metges CC, Wimmers K, Ponsuksili S. Somatic cytochrome c (CYCS) gene expression and promoter-specific DNA methylation in a porcine model of prenatal exposure to maternal dietary protein excess and restriction. Brit J Nutr 2012;107:791–9.
52. Zhang N. Epigenetic modulation of DNA methylation by nutrition and its mechanisms in animals. Animal Nutrition 1 (2015) 144–151.
53. Scoccianti C, Ricceri F, Ferrari P et al. Methylation patterns in sentinel genes in peripheral blood cells of heavy smokers: Influence of cruciferous vegetables in an intervention study. Epigenetics. 2011 Sep 1;6(9):1114-9. doi: 10.4161/epi.6.9.16515. Epub 2011 Sep 1.
54. Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev. 2007 Oct;16(10):1954-65. PMID: 17932342.
55. Vanderweele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction. New York: Oxford University Press, 2015, p.119.
56. Baron RM and KennyD. TheModerator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. J Personal Disord 1986; 51: 1173-1182.
57. Robins JM and Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992; 3: 143-155.
58. Pearl J. Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (ed J Breese and D 26 Koller), San Francisco, CA, 2-5 August 2001, pp.411-420. Burlington: Morgan Kaufmann.
59. VanderWeele TJ and Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. Statistics and Its Interface 2009; 2: 457-468.
60. VanderWeele TJ and Vansteelandt S. Odds ratios for mediation analy- sis for a dichotomous outcome. Am J Epidemiol 2010; 172: 1339-1348.
61. Imai K, Keele L and Tingley D. A general approach to causal mediation analysis. Psychological Methods 2010; 15: 309-334.
62. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. Hum Mol Genet. 2016 Oct 1;25(R2):R149-R156. Epub 2016 Jul 20.

2. Present studies

2.1 General aims

The main aim of the present thesis was to investigate the role of DNA methylation as a potential mediator of the carcinogenic process triggered by specific environmental exposures.

Furthermore another objective was to concentrate on mediation analysis technique developing new methods for data analysis (methodological part).

The specific objectives of the studies were:

1. to understand if methylation levels at some of the sites previously found to be strong markers of smoking also translate into increased risk of lung cancer (project 1, Chapter 3);
2. to investigate the role of DNA methylation as a biological mechanism behind the protection of the Mediterranean Diet against colon cancer (project 2, Chapter 4);
3. to extend a method developed in multiple mediation analysis to survival outcome (project 3, Appendix).

In Section 2.2, a summary and a short comment of the individual projects are presented. In Chapter 3 the final version of *Fasanelli et al.* 2015 [1] is proposed. Pre-print drafts of projects 2 and 3 (unpublished) are included in Chapter 4 and in the Appendix respectively.

2.2 Summary of the projects

2.2.1 Project 1: Smoking, DNA methylation and lung cancer (Chapter 3)

Summary

DNA hypomethylation in certain genes is associated with tobacco exposure but it is unknown whether these methylation changes translate into increased lung cancer risk. In an epigenome-wide study of DNA from pre-diagnostic blood samples from 132 case-control pairs in the Norwegian Women and Cancer (NOWAC) cohort, we observed that the most significant associations with lung cancer risk were for cg05575921 in *AHRR* (OR for 1 SD = 0.37, 95% CI: 0.31-0.54, p-value=3.3x10^{-11}) and cg03636183 in *F2RL3* (OR for 1 SD = 0.40, 95% CI: 0.31-0.56, p-value=3.9x10^{-10}), previously shown to be strongly hypomethylated in smokers. These associations remained significant after adjustment for smoking and were confirmed in additional 664 case-control pairs tightly matched for smoking from MCCS (the Melbourne Collaborative Cohort Study), NSHDS (the Northern Sweden Health and Disease Study) and EPIC HD (the European Prospective Investigation into Cancer and Nutrition, Heidelberg) cohorts. The replication and mediation analyses suggested that residual confounding was unlikely to explain the observed associations and that hypomethylation of these probes may mediate the effect of tobacco on lung cancer risk.

Commentary

This work was the object of my first two years PhD work and was published in *Nature Communications* in 2015 [1].

The mediation analysis applied in this study identified that 37% of the total effect of smoking on lung cancer was mediated by differential methylation in *AHRR* and *F2RL3* region. Subsequent work focused on mediation analysis suggested a likely overestimation of the indirect effect through these two genes [2], [3]. In fact the evaluation of mediation may have been complicated by the fact that the proposed mediators, DNA sites differentially methylated by smoking, are excellent biomarkers of smoking that may better capture the exposure than self-reported smoking (the measurement error in the exposure “self-reported smoking” is more prone to error than DNA methylation, leading to residual confounding of the mediator–outcome association). Despite this methodological problem, this work was useful since it was the starting point for a series of projects aimed at using these smoking-associated methylation markers to improve lung cancer detection and risk stratification. In particular in [4] (a work I was involved in) a gain in discrimination between cases and controls measured by an increase in the area under the ROC curve of at least 8% (p-values>=0.003) was observed in former smokers by adding the methylation of six CpGs as covariates into risk prediction models including smoking status and number of pack-years. Similarly Zhang et al. [5] constructed a multi-loci score based on smoking-associated methylation sites that predicts lung cancer mortality with high accuracy and may thus serve as promising candidate to identify high risk populations for lung cancer screening. All these studies provide convincing evidence that smoking leads to DNA methylation changes measurable in peripheral blood that may improve prediction of lung cancer risk.

A recent paper [6] analyzed the mechanims by which these hypomethylation events arise with the aim to explain how they might increase lung cancer risk. First the authors studied the association between tobacco smoking and epigenome-wide methylation in non-tumour lung tissue identifying seven smoking-associated hypomethylated CpGs. Among these loci, there is the *AHRR* CpG site cg05575921. This result suggests that for this locus the methylation measured in blood faithfully reflects the methylation measured in the target tissue. Furthermore, by studying in detail this CpG in primary alveolar epithelium and in A549 lung adenocarcinoma cells, they found that it borders sequences carrying aryl hydrocarbon receptor (AHR) binding sites and histone modifications typical of enhancers. A549 cell exposure to cigarette smoke condensate was shown to increase these enhancer marks significantly and to stimulate the expression of predicted target xenobiotic response-related genes such as the genes that metabolize procarcinogens (e.g. *CYP1A1* or *CYP1B1*) and *AHRR*, which is a suppressor and feed-back regulator of AHR activity. Hypomethylation may be a byproduct of enhancer activation since transcription factors interact with the enhancer element and presumably protect it from maintenance DNA methyltransferase activity [7], [8].

2.2.2 Project 2: Mediterranean diet, DNA methylation and colon cancer (Chapter 4)

Summary

Adherence to Mediterranean Diet (MD) has a preventive effect on colon cancer. However, the biological mechanisms through which MD protects against colon cancer are poorly understood.

Recent evidence suggests that DNA methylation may be implicated in the pathway between adherence to MD and colon cancer onset.

An agnostic search of DNA methylation signals associated with both colon cancer and MD was carried out using data from two epigenome-wide studies from the EPIC Italy cohort (87 case-control pairs, discovery set; 74 case-control pairs, replication set). In addition considering together the 161 case-control pairs, a hypothesis-driven analysis was performed examining only 995 CpGs located in inflammation genes known from literature to be related to solid human cancer and MD. The DNA methylation signals detected in this analysis were validated in a subgroup of 47 cases and 47 controls among the already analyzed subjects and further replicated (where validated) in a group of 95 new case-controls pairs using pyrosequencing. DNA methylation was assessed in peripheral blood collected at recruitment into the EPIC study.

The genome-wide analysis did not reveal any significant DNA methylation signal. When focusing on inflammation genes, seven CpG sites were found to be associated with colon cancer status and showed also an association with MD in line with its protective effect. Among these seven, two were validated by pyrosequencing (cg17968347-*SERPINE1* and cg20674490-*RUNX3*) and only one of them showed similar associations in an independent sample (cg20674490-*RUNX3*).

This study is a first attempt to identify the biological mechanism behind the protective role of MD against colon cancer investigating the methylation levels of genes in circulating lymphocytes years before the onset of the disease.

Commentary

This work is part of a project funded by the Italian Association for Cancer Research (IG 2013 number 14410). The general aim of the project was to investigate the biological mechanisms behind the protective role of Mediterranean Diet against colon cancer. The proposed mechanisms have been related to the shifts in the plasma-glucose values and weight loss, due to the high amount of cereals with low glycemic index. This aspect has been analyzed from an epidemiological point of view in a work we have recently published on the *International Journal of Cancer* [9]. The conclusion is that abdominal adiposity is not a mediator of the association between Mediterranean Diet and colon cancer.

Another proposed mechanism for cancer prevention associated with the Mediterranean Diet includes the favorable effect of a balanced ratio of omega 6 and omega 3 essential fatty acids and high amounts of fibers, antioxidants and polyphenols found in fruit, vegetables and olive oil that can act through the attenuation of pro-inflammatory mediators. This aspect is analyzed from the epigenetic point of view in the work reported in Chapter 4. The results suggest that DNA methylation of the inflammation gene *RUNX3* may be a potential molecular mediator explaining the protective effect of MD on colon cancer onset. However this finding is not so strong to build a mediation model since independent functional studies are needed to confirm its mediating role in colon carcinogenesis.

Regardless of the results, this work presents some innovative aspects: i) the search for signals associated simultaneously with exposure and outcome and ii) the presence of a validation phase and a replication phase using a different DNA methylation assay. In literature there are very few studies that compare different DNA methylation assays for biomarker development [10], [11] and in general they consider DNA methylation assessed using solid tissues or specific cell lines. Innovatively we validated and replicated our methylation signals analyzing blood lymphocytes. The fact that only one signal was confirmed emphasizes the importance of the validation and replication phases using an alternative technology especially when methylation is measured on blood and differences in methylation percentages are little. This topic will be the subject of a future methodological work.

2.2.3 Project 3: methodological paper (Appendix)

<u>Summary</u>

As said in Section 1.5, the main aim of mediation analysis is to study the direct (not mediated) and indirect (mediated) effects of an exposure on an outcome of interest. To date, the literature on mediation analysis with multiple mediators has mainly focused on continuous and dichotomous outcomes. However, development of methods for multiple mediation analysis of survival outcome is still limited. In this article, we show how to extend a method for multiple mediation analysis based on the computation of appropriate weights to survival outcome. The method is illustrated along with an estimation algorithm, assuming a proportional hazards model conditional on exposure, mediators and covariates and allowing for marginal direct and indirect effects to vary over time. The method is applied to an example from a dataset coming from a published study on mortality for prostate cancer where the interest was to understand to what extent the effect of DNA methyltransferase genotype on mortality was explained by DNA methylation and tumor aggressiveness. The approach described is straightforward and can be used to quantify the marginal time-dependent direct and indirect effects carried by multiple indirect pathways.

Commentary

This paper introduces a methodological work resulting from the collaboration with mediation analysis experts (Linda Valeri, Harvard Medical School; Daniela Zugna, University of Turin) and with the Department of Mathematics “Giuseppe Peano” of the University of Turin. In fact it is part of an interdisciplinary research field that combines mathematical and statistical with genetic, medical and epidemiological skills.

The manuscript will be submitted to a methodological journal and is included in the Appendix of the present thesis. The usefulness of the methodology proposed is illustrated in the paper using a dataset (Section 4 in the Appendix) coming from a published work that studied the relationships among DNA methyltransferase genotype (DNMT, polymorphism rs406193), DNA methylation, tumor aggressiveness and long-term mortality for prostate cancer [12]. Briefly, with the method proposed we were able to quantify how much of the total effect of the variant on the cause-specific mortality was attributable to the indirect effect through tumor tissue methylation and Gleason score.

This work is an example of how integrating different skills can be a useful tool for biomedical research.

References

1. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma M, Naccarati A, Chadeau-Hyam M, De Stavola B, Hodge A, Giles GG, Southey MC, Relton CL, Haycock PC, Lund E, Polidoro S, Sandanger TM, Severi G, Vineis P. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015 Dec 15;6:10192. doi: 10.1038/ncomms10192.
2. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. Hum Mol Genet. 2016 Oct 1;25(R2):R149-R156. Epub 2016 Jul 20.
3. Valeri L, Reese SL, Zhao S, Page CM, Nystad W, Coull BA, London SJ. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight? Epigenomics. 2017 Mar;9(3):253-265. doi: 10.2217/epi-2016-0145. Epub 2017 Feb 21.
4. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, Chung J, Fasanelli F, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J Cancer. 2016 Sep 15. doi: 10.1002/ijc.30431.
5. Zhang Y, Breitling LP, Balavarca Y, Holleczek B, Schöttker B, Brenner H. Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality. Int J Cancer. 2016 Dec 1;139(11):2482-92. doi: 10.1002/ijc.30374. Epub 2016 Aug 22.
6. Stueve TR, Li WQ, Shi J et al. Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. Human Molecular Genetics, Volume 26, Issue 15, 1 August 2017, Pages 3014–3027, https://doi.org/10.1093/hmg/ddx188
7. Xu, J., Watts, J.A., Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S. and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev., 23, 2824–2838.
8. Aran, D., Sabato, S. and Hellman, A. (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol., 14, R21.
9. Fasanelli F, Zugna D, Giraudo MT et al. Abdominal adiposity is not a mediator of the protective effect of Mediterranean diet on colorectal cancer. Int J Cancer. 2017 May 15;140(10):2265-2271. doi: 10.1002/ijc.30653. Epub 2017 Mar 2.
10. Roessler J, Ammerpohl O, Gutwein J, Hasemeier B, Anwar SL, Kreipe H, Lehmann U. Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc. BMC Res Notes. 2012 Apr 30;5:210. doi: 10.1186/1756-0500-5-210.
11. BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol. 2016 Jul;34(7):726-37. doi: 10.1038/nbt.3605. Epub 2016 Jun 27.
12. Gillio Tos A, Fiano V, Zugna D, et al. DNA methyltransferase 3b (DNMT3b), tumor tissue DNA methylation, Gleason score, and prostate cancer mortality: investigating causal relationships. Cancer Causes Control 2012; 23: 1549-1555.

3. Project 1: Smoking, DNA methylation and lung cancer

Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma M, Naccarati A, Chadeau-Hyam M, De Stavola B, Hodge A, Giles GG, Southey MC, Relton CL, Haycock PC, Lund E, Polidoro S, Sandanger TM, Severi G, Vineis P. **Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts.** Nat Commun. 2015 Dec 15;6:10192. doi: 10.1038/ncomms10192.

3.1 Introduction

DNA methylation has recently emerged as an important marker of current and past smoking habits [1]-[9]. Smoking is a leading cause of death worldwide [10], [11] and has been identified as a major risk factor for several diseases including cancer [12], [13] cardiovascular [14], [15] and respiratory diseases [16], [17]. The carcinogenic effect of tobacco smoking persists for decades after smoking cessation, and former smokers remain at increased risk of lung cancer for 20 years or longer [18]-[20].

Using an epigenome-wide methylation study approach we previously demonstrated that tobacco smoking alters DNA methylation patterns, particularly in CpG sites of the *AHRR* and *F2RL3* genes [7]. These results have been extensively replicated by other studies [1]-[6], [8]. In particular our previous study of 1,000 healthy subjects from the EPIC and NOWAC cohorts indicated that smokers had 19% lower methylation levels at the *AHRR* CpG site cg05575921 compared with never-smokers. We also found that one set of specific methylation markers showed a gradual reversal of methylation levels from those typical of current smokers to those of never smokers, whereas other smoking-related CpG sites' methylation markers remained stable more than 30 years after quitting [9]. These findings are also consistent with other recent studies reporting methylation levels in smoking-related CpG loci in former smokers to vary based on their time since quitting [21]-[22].

Whilst these previous studies have provided convincing evidence of an association between tobacco exposure and methylation of specific CpG sites, it is not known whether methylation levels at some of these sites translate into increased risk of smoking related cancers, such as lung cancer. Here, we present the results of an epigenome-wide methylation study based on methylation detection using

Illumina Infinium HM450 on DNA extracted from pre-diagnostic blood of 132 pairs of lung cancer cases and controls from the NOWAC cohort (discovery set). We replicated the findings in three prospective studies, the Melbourne Collaborative Cohort Study (MCCS) (367 cases and 367 matched controls), the Northern Sweden Health and Disease Study (NSHDS) (234 cases and 234 matched controls) and the EPIC Heidelberg Study (EPIC HD) (63 cases and 63 matched controls) (replication sets), with adjustment for smoking habits. To our knowledge, this is the first study performing a genome-wide methylation analysis to evaluate the importance of epigenetic alterations in peripheral blood DNA to lung cancer etiology.

3.2 Results

Discovery set

Incident lung cancer cases in the discovery set (NOWAC) were identified through linkage with the Cancer Registry of Norway, with virtually complete coverage. In the nested case-control study lung cancer cases were diagnosed on average 3.88 years after recruitment (range: 0.29-7.92 years) and the mean age at diagnosis was 56 years (range: 47-64 years). The odds ratio for lung cancer was 7.38 for former and current smokers grouped together (95% confidence interval 3.99-16.66), 6.16 (95% confidence interval 2.65-15.13) for former smokers and 10.13 (95% CI 4.56-24.23) for current smokers.

Table 3.1 shows the top-ranked CpG sites for the locus-by-locus epigenome-wide risk analysis, and includes all CpG sites with Bonferroni-corrected p-values below 0.05. All top-ranked CpGs showed inverse associations with risk, indicating hypomethylation in cancer cases. Supplementary Table 3.1 shows the main information about involvement in cancer pathways for the probes listed in Table 3.1: for all the CpGs except two (cg02451831, cg03898802) there is evidence of involvement in cancer pathways. CpGs in the *AHRR* and *F2RL3* genes displayed the most significant associations with risk consistent with previous observations of smoking being associated with reduced methylation in healthy subjects [1]-[9]. In the following analyses we exclusively focus on these two genes from Table 3.1, because they are the only ones strongly associated with smoking. In particular, the cg05575921 probe in the *AHRR* gene emerged as the CpG site most strongly associated with both tobacco exposure [9] and lung cancer risk (OR for lung cancer per 1 SD of beta: 0.37, 95% CI: 0.31-0.54, p-value=3.33x10^{-11}). Sensitivity analyses excluding cases with time from blood collection to diagnosis of less than 2 years showed no significant differences in effect estimates (OR 0.36, 95% CI: 0.27-0.52 for cg05575921 and OR 0.40, 95% CI: 0.29-0.56 for cg03636183). Supplementary Table 3.2 shows the results of the analyses stratified by time to diagnosis (less and more than 5 years). Associations were slightly stronger for less than 5 years to diagnosis but these were unlikely to reflect reverse causation as they were also evident for more than 5 years to diagnosis.

Table 3.2 shows the results for the probes associated with cancer risk in the *AHRR* and *F2RL3* genes after adjustment for smoking (e.g. smoking status coded as never, former, current): the overall association remained basically unchanged (OR for 1 SD = 0.39, 95% CI: 0.24-0.61, p-value=2.55x10^{-5} for cg05575921 and OR for 1 SD=0.51, 95% CI: 0.35-0.73, p-value=4.19x10^{-4} for cg03636183).

Replication sets

To replicate our results arising from the NOWAC study, we analyzed the cg05575921 and cg03636183 probes in three independent samples: a case-control study nested within MCCS including 367 case-control pairs, a case-control study nested within the NSHDS including 234 case-control pairs and a case-control study nested within the EPIC HD cohort, including 63 case-control pairs, all of which were matched on smoking status (see Methods for details).

Consistent with the results from the NOWAC study, methylation levels in the MCCS, NSHDS and EPIC HD studies were clearly inversely associated with lung cancer risk for both the cg05575921 and cg03636183 CpG sites. The overall OR estimates were slightly weaker in MCCS than in NOWAC (OR 0.62, 95% CI: 0.50-0.78, p=2.91x10^{-5} for cg05575921 and OR 0.70, 95% CI: 0.58-0.85, p=2.21x10^{-4} for cg03636183), but more comparable in NOWAC to NSHDS and EPIC HD

(OR 0.42, 95% CI: 0.30-0.58, $p=2.06\text{x}10^{-7}$ for cg05575921 and OR 0.61, 95% CI: 0.47-0.79, $p=1.56\text{x}10^{-4}$ for cg03636183 in NSHDS; OR 0.45, 95% CI: 0.22-0.92, $p=2.95\text{x}10^{-2}$ for cg05575921 and OR 0.62, 95% CI: 0.38-1.04, $p=7.02\text{x}10^{-2}$ for cg03636183 in EPIC HD) (Table 3.2). We note that some attenuation of the overall NOWAC OR estimates is expected as MCCS, NSHDS and EPIC HD studies were matched by smoking status.

Risk prediction model for lung cancer

We applied to the NOWAC cohort a prediction model including smoking status (coded as never, former, current) and methylation as a covariate. This was not feasible for the other cohorts because of matching by smoking. The area under the curve (AUC) of the model increased from 0.71 to 0.76 when adding *AHRR*-methylation and *F2RL3*-methylation as categorical variables (above or below the median) and to 0.78 when adding the two as continuous variables.

Lung cancer risk by categories of smoking exposure

To further evaluate the associations of the cg05575921 and cg03636183 CpG sites with lung cancer risk, we conducted stratified risk analysis by categories of smoking status. We found little support for an association being present for never smokers for either CpG site, and the associations were clearly influenced by smoking. A notable observation regarding ever smokers was that the association appeared to be stronger for former smokers than for current smokers. For instance, in the NOWAC study the OR for the cg05575921 site was 0.23 (95% CI: 0.10-0.56) for former smokers, and 0.46 (95% CI: 0.24-0.88) for current smokers. This pattern was evident also in MCCS, NSHDS and EPIC HD for both the cg05575921 and cg03636183 CpG sites (Table 3.2).

Methylation of AHRR and F2RL3 genes in former smokers

The associations between smoking cessation and the mean methylation levels in the cg05575921 probe (*AHRR* gene) and the cg03636183 probe (*F2RL3* gene) in NOWAC are shown in Figure 3.1. After smoking cessation, methylation levels increase and after 10 years since quitting appear to approach those of never smokers. This is consistent with the well-documented observation that the risk of lung cancer decreases substantially after smoking cessation.

The effect of smoking (never vs former vs current; time since quitting smoking; smoking duration) on methylation beta levels for cg05575921 and cg03636183 in MCCS and in NSHDS are shown in Figure 3.2. Similarly to what we observed in NOWAC (Figure 3.1), in MCCS and NSHDS methylation levels in current smokers were lower than methylation levels in never smokers and in former smokers the levels approached those of never smokers with increasing time since cessation.

Comparison of the study groups

Supplementary Table 3.3 shows a summary of the key characteristics of the study groups. The limitation to a single gender in NOWAC prevented us from making straightforward comparisons between the estimated associations and from investigating differences in lung cancer risk between genders. On the other hand, matching by smoking in MCCS, NSHDS and EPIC HD did not allow us *i)* to investigate further the role of methylation as a mediator of the association between smoking and cancer in these cohorts and *ii)* to test interactions between smoking variables such as duration or dose. A future goal will be to repeat the analysis in unrestricted population cohorts.

Correlation between methylation and expression

We investigated the correlation between methylation and expression of the two relevant probes using two different sources of data: TCGA (http://cancergenome.nih.gov/) and HapMap (http://hapmap.ncbi.nlm.nih.gov/). In TCGA we focused on expression (RNA-Seq experiments) and methylation (Illumina HumanMethylation450 BeadChip) of samples of normal tissue *i)* from 21 lung adenocarcinoma cases (LUAD - 21 methylation-expression pairs) and *ii)* from 8 lung squamous cell carcinoma cases (LUSC - 8 methylation-expression pairs). *AHRR*-probe methylation seems to be significantly inversely-correlated with *AHRR* expression in LUAD and the same trend was found in LUSC (Pearson's correlation coefficient=-0.66, p value<0.01 in LUAD; Pearson's correlation coefficient=-0.43, p value=0.29 in LUSC). *F2RL3*-probe methylation did not show a statistically significant methylation-expression correlation. Regarding Hap Map, we focused on expression (RNA-Seq experiments) and methylation (Illumina HumanMethylation27 BeadChip) data from lymphoblastoid cell lines of 69 HapMap Yoruba individuals. In this case only the *F2RL3*-

probe is present on the platform and its methylation seems to be significantly inversely-correlated with *F2RL3* expression (Pearson's correlation coefficient=-0.28, p value<0.01).

Mediation analysis

Whilst the results described above from the analysis of a discovery set and three replication sets seem to provide evidence that hypomethylation of the cg05575921 and cg03636183 probes is associated with both tobacco exposure and lung cancer risk, the key question is whether their hypomethylation is involved in the causal pathway, or whether they are simply epiphenomena of smoking habits (i.e. the association of DNA methylation with lung cancer risk is confounded by smoking). To bring some clarity to this question, we used mediation analysis to quantify the amount by which cg05575921 (*AHRR* gene) and cg03636183 (*F2RL3* gene) methylation might mediate the effect of smoking on lung cancer incidence. This was performed for the NOWAC study as such an analysis was not possible for the MCCS, NSHDS or EPIC HD due to matching by smoking status.

We detected statistically significant results for both components of mediation analysis, the natural direct effect of smoking on lung cancer (NDE, i.e. not mediated) and the natural indirect effect (NIE, i.e. the effect mediated by the methylated probe(s)), the two together making up the total causal effect (TCE) (see Methods and Table 3.3, where the underlying identifying assumptions are also stated). The proportion of the smoking-induced risk increase explained by cg05575921 *AHRR*-probe was found to be about 31% (0.31, 95% CI: 0.18-0.46), and 32% (0.32, 95% CI: 0.20-0.53) for the cg03636183 *F2RL3*-probe. Considering the two genes together, their methylation appeared to mediate about 37% (0.37, 95% CI: 0.19-0.66) of the total effect of smoking on lung cancer odds (Figure 3.3 and Table 3.3). The results of mediation analysis were similar when we included the mean methylation of a group of 10 *AHRR* (cg05575921, cg03991871, cg12806681, cg23916896, cg01899089, cg26703534, cg14817490, cg25648203, cg21161138 and cg24090911) and 2 *F2RL3* probes (cg03636183 and cg04259305) located in the body of the gene and significantly associated with lung cancer after false discovery rate (FDR) correction (data not shown). In conclusion, this analysis suggests (a) that methylation of the smoking related *AHRR* and *F2RL3* probes might be relevant to lung cancer etiology, and (b) would explain approximately one third of the risk increase induced by tobacco exposure.

3.3 Discussion

Tobacco smoking is one of the most important carcinogenic exposures, and continuing smokers experience up to 25% lifetime risk of developing a smoking-related cancer – particularly lung cancer – yet the underlying mechanisms by which tobacco carcinogens act on lung cells have been elusive. Mutations, cell proliferation and selection have been hypothesized as complementary mechanisms [23], [24]. Epigenetics has recently emerged as a promising field to illuminate carcinogenetic nechanisms [24] and we have previously shown that smoking is associated with hypomethylation in CpGs of key genes [9]. Here, we present data from four prospective cohort studies that convincingly demonstrate that hypomethylation in specific CpG sites of the *AHRR* and *F2RL3* genes is associated with increased risk of subsequent lung cancer. Although we detected 11 probes in the discovery set that were associated with lung cancer, we selected the *AHRR* and *F2RL3* genes genes because of their strong association with smoking found in previous studies, and because our aim was to test whether methylation may feature in the pathway from smoking to lung cancer. *AHRR* is the repressor of the aryl hydrocarbon receptor, a key regulator of the relationships between the cell and the external environment, including the effects of stressors such as dioxins and polycyclic aromatic hydrocarbons (that are contained in tobacco smoke) [25]. *AHRR* is expressed in all tissues, where it controls cell proliferation and apoptosis; it is upregulated and epigenetically modified in lung alveolar macrophages of smokers [1]. We have previously investigated the lung tissue of smokers and non-smokers: methylation levels in the *AHRR* gene probes were significantly lower ($p<0.001$) with a concurrent increase in *AHRR* expression ($p=0.005$) in the lung tissue of current smokers compared with non-smokers [7]. This was further validated in a mouse model of smoke exposure [7].

F2RL3 is also a functionally relevant gene. It encodes for the protease-activated receptor-4 (PAR-4), which has been suggested to be involved in the pathophysiology of both cardiovascular and neoplastic diseases [26]. A recent paper reported that hypomethylation of *F2RL3* is predictive of total mortality and the authors suggested that the adverse health effects of smoking might be mediated in part by pathways related to *F2RL3* methylation [26].

The main question arising from our previous studies of healthy subjects was whether methylation changes in the *AHRR* and *F2RL3* genes are causally involved in lung cancer etiology by mediating the risk induced by tobacco smoking. Whilst it is not possible to fully answer this question based on our data, our results are consistent with the notion of a mediating role. We have observed *i)* that data from multiple independent study populations have conclusively established an association between tobacco smoking and *AHRR* and *F2RL3* methylation, and *ii)* that these methylation sites are also associated with lung cancer risk after adjustment for smoking habits and with careful mediation analysis. Whilst it is possible that residual confounding from tobacco smoking might still explain the association with risk, we note that the attenuation in OR estimates when adjusting for smoking is negligible in all three studies. Should residual confounding from tobacco smoking explain our observed associations, we would expect a notable attenuation of OR estimates in adjusted risk models. Additionally, the observation that smoking associated hypomethylation in these specific CpG sites is reversible following smoking cessation is compatible with the gradual decrease in lung cancer risk that former smokers experience. A full evaluation of the causal relevance of *AHRR* and *F2RL3* methylation in lung cancer etiology requires additional investigations, such as a Mendelian randomization analysis of a sufficiently powered study [27]. Hypomethylation of certain probes/genes which extends beyond smoking cessation for several years, as observed for the two probes identified in this study, might be more closely associated with lung cancer risk than transient hypomethylation. In previous analyses of healthy subjects [9] we generally observed a relatively rapid reversal of smoking-related methylation changes, but for a group of probes including cg05575921 and cg03636183 reversal is slower or not apparent even after decades. A larger study is required to evaluate whether reversal of methylation alterations in cg05575921 and cg03636183 occurs at the same rate as the decrease in the risk of lung cancer in former smokers. Also, future prediction models will be built based on a larger number of cohorts not matched by smoking habits (work in preparation). In the present study we were able to build such a model only for the NOWAC cohort, and there was a modest increase in prediction (AUC changing from 0.71 to 0.78 when methylation information was added).

Hypomethylation persists in some probes for much longer than the average half-life of circulating white-blood cells suggesting that stem cells (in the bone marrow in the case of white blood cells, and hypothetically also in the lung [1]) may preserve a "memory" of past exposures in the form of a greater proportion of unmethylated CpG sites vs methylated CpG sites. We speculate that exposure to toxic agents leads to clonal expansion of cells that are hypomethylated in CpGs of genes involved in activation of a pathway reactive to environmental insults, and this imbalance in the proportion of methylated DNA in stem cells persists, remaining mitotically stable through subsequent cell divisions.

The association of hypomethylation at the two selected probes with lung cancer was nominally stronger for former than for current smokers in all our studies but this observation could be due to chance or residual confounding by factors related or unrelated to smoking.

In conclusion, our study shows that smoking induced hypomethylation in the *AHRR* and *F2RL3* genes is associated with important risk increases of subsequent lung cancer, and indicates that these specific methylation alterations may mediate the carcinogenic effect of tobacco exposure in lung cancer aetiology.

3.4 Methods

Discovery set

Lung cancer cases and matched controls were identified within the Norwegian *NOWAC* longitudinal cohort. The biobank of the NOWAC cohort was collected in the years 2003-2006.

Random samples of Norwegian women were mailed a letter of information with an invitation to receive equipment for blood sampling at the local doctor or other institutions. Those who filled in the eight-page questionnaire and accepted the invitation to donate blood received some months later equipment for blood drawing together with a two-page questionnaire with information on date, lifestyle factors etc. Around 50,000 women returned by over-night mail two tubes of blood to the Institute of Community Medicine at UiT The Artic University of Norway. Upon arrival, the citrate glass tube was centrifuged and buffy-coat and plasma frozen immediately at –80 degrees together with a PAXgene tube. All participants gave informed consent. The study was approved by the Regional Committee for Medical and Health Research Ethics in North Norway. Data storage and linkage to the National Cancer Registry of Norway was approved by the Norwegian Data Inspectorate; follow-up identified 132 eligible cases of lung cancer by 2011. For each case one control with adequate blood samples was selected matched on time since blood sampling and year of birth in order to control for effects of storing time and ageing. The cases and the controls were kept together through all later laboratory procedures in order to reduce any batch effects.

Replication sets

The Melbourne Collaborative Cohort Study (MCCS) is a prospective cohort study of 41,514 volunteers (24,469 women) aged between 27 and 76 years at baseline (99.3% of whom were aged 40-69) [28]. The MCCS study protocol was approved by the Cancer Council Victoria's Human Research Ethics. At baseline attendance, in 1990-1994, participants completed questionnaires that measured demographic characteristics and lifestyle factors including diet. Height and weight were directly measured and a blood sample was collected and stored. For a large proportion of individuals (75%) only dried blood spots on Guthrie cards were available while for others buffy coat or lymphocyte samples were available. A total of 533 incident cases of lung cancer identified through linkage with the State and National Cancer Registry wasdiagnosed during follow-up up to the end of 2011. A total of 367 cases remained available after excluding cases 1) diagnosed after the age of 80 years; 2) with no biospecimen available; 3) with a diagnosis of any cancer before blood draw; or 4) with no information on smoking status. The MCCS sample included 367 cases (159 adenocarcinomas, 33 large cell cancers, 73 squamous cancers and 49 small cell cancers) and 367 matched controls selected with a density sampling procedure. Matching variables included sex, date of blood collection (within 6 months), date of birth (within 1 year), country of birth (Australia and UK versus Southern Europe), type of biospecimen (lymphocyte, buffy coat and dried blood spot) and smoking status (never smokers; short-term former smokers: quitting smoking less than 10 years before blood draw; long-term former smokers: quitting smoking 10 years or more before blood draw; current light smokers: less than 15 cigarettes per day at blood draw; and current heavy smokers: 15 cigarettes or more at blood draw). In the sample, the mean time between blood draw and diagnosis was 9.38 years (SD, 5years).

The Northern Sweden Health and Disease Study (NSHDS) is an ongoing prospective cohort and intervention study intended for health promotion in the population of Västerbotten County in northern Sweden. The study was approved by the Umeå University Ethical Committee; details of the study population have been published previously [29]. Briefly, study participants were recruited to the NSHDS in the context of the Västerbotten Intervention Project (VIP), which was initiated in 1985 to advocate a healthy diet and lifestyle. All residents in the Västerbotten County were invited to participate in the project by attending a health check-up at 40, 50 and 60 years of age. At the health check-up, which was held at the local health care centre, participants were asked to complete a self-administered questionnaire including various demographic factors such as education, smoking habits, physical activity and diet. In addition, height and weight were measured and participants were asked to donate a blood sample of 20mL for future research. Incident lung cancer cases were identified through linkage to the regional cancer registry. Lung cancer cases were defined on the basis of the International Classification of Diseases for Oncology, Second Edition (ICD-O-2), and included all primary malignant cancers that are coded as C34.0-C34.9 with pre-diagnostic blood samples. One control was chosen at random for each lung cancer case from appropriate risk sets consisting of all cohort members alive and free of cancer (except nonmelanoma skin cancer) at the time of diagnosis of the index case. Matching criteria included: date of birth (± 1 year, relaxed up to ± 5 years for cases without available controls), ethnicity, gender, date of blood collection (± 1 month, relaxed up to ± 3 months, and further to ± 6 months for cases without available controls), and detailed smoking status: never smokers, short-term former smokers (quitting smoking less than 10 years before blood draw), long term former smokers (quitting smoking over 10 years before blood draw), current light smokers (<15 cigarettes/day at blood draw) and current heavy smokers (≥15 cigarettes per day at blood draw). After quality control, a total of 234 incident lung cancer cases (111 adenocarcinomas, 6 large cell cancers, 47 squamous cancers, and 29 small cell cancers) and 234 individually matched controls were available for this analysis. In the sample, the mean time from blood draw to diagnosis was 9.6 years (range: 1.1-17.5).

The *European Investigation into Cancer and Nutrition (EPIC)* is a large multicenter cohort study of diet and chronic diseases. The study rationale has been published previously [30], [31]. In brief, in the EPIC Heidelberg cohort study (EPIC HD) 25,500 study participants from the general population were recruited from June 1994 to October 1998. Inhabitants of Heidelberg and of the surrounding region who met the age criteria of the EPIC study design (men: 40–64, women: 35–64) were randomly invited by mail to take part in the study. Study subjects were asked to complete questionnaires and were interviewed about their individual health, diet and lifestyle such as life history of tobacco smoking and alcohol intake. Additionally, anthropometric measurements were taken and a blood sample of 30ml was collected which was fractionated and stored in aliquots in liquid nitrogen for future research. Up to six follow up questionnaires were sent to the participants, at 2 to 3-year intervals, to ask about incident diseases and changes in lifestyle and diet. All self-reported incident cases of cancer were systematically verified against clinical and pathology records. The present study was based on 211 incident lung cancer cases identified by July 2015. Cases with less than one year from blood draw to diagnosis were excluded. Of the remaining cases those with the shortest follow-up times to diagnosis and who were either current or former smokers at the baseline recruitment were selected for this study (n=66). EPIC controls without any neoplastic disease were randomly matched to the lung cancer cases using an incidence density protocol. Matching was done on the basis of age at baseline (± 5 years), gender, smoking status (current and former), and pack years (± 1 PY). After initial quality control 63 incident lung cancer cases (25 adenocarcinomas, 15 squamous cell carcinoma, 19 small cell lung cancer and 4 uncharacterized lung cancers) with a mean interval between blood draw and diagnosis of 4.8 years (range: 1.1-8.6) and 63 individually matched controls remained for further analysis. This study was approved by the ethics committee of the Medical Faculty of the University of Heidelberg (S-627/2013).

DNA methylation measurement, data pre-processing and quality control

Genome-wide DNA methylation analyses were performed on pre-diagnostic blood samples using the Illumina Infinium HumanMethylation450 platform.

NOWAC laboratory procedures were carried out at the Human Genetics Foundation (Turin, Italy), using the Illumina Infinium HumanMethylation450 (HM450). Buffy coats stored in liquid nitrogen were thawed, and genomic DNA was extracted using the QIAGEN QIAsymphony DNA Midi Kit. 500 ng of DNA were bisulphite-converted using the Zymo Research EZ-96 DNA Methylation-Gold™ Kit, and hybridised to Illumina Infinium HumanMethylation450 BeadChips. These were subsequently scanned using the Illumina HiScanSQ system, and sample quality was assessed using control probes present on the micro-arrays. Finally, raw intensity data were exported from Illumina GenomeStudio (version 2011.1).

MCCS laboratory procedures were carried out at the Genetic Epidemiology Laboratory, the University of Melbourne according to manufacturers' protocols. DNA extraction from lymphocytes and buffy coats was performed using Qiagen mini spin columns (Hilden, Germany) while dried blood spot DNA was extracted using a method developed in-house [32] and the quality and quantity of DNA was assessed using the Quant-iT™ Picogreen® dsDNA assay measured on the Qubit®

Fluorometer (Life Technologies, Grand Island, NY). Samples were distributed into 96-well plates and processed in chips of 12 arrays (8 chips per plate) with case-control pairs arranged randomly on the same chip. All subsequent steps were performed as described above for NOWAC.

NSHDS laboratory procedures were carried out on two sites. DNA extraction from the buffy coat of EDTA-venous blood samples was conducted at Umeå University, Sweden, using FlexiGene DNA Kit (QIAGEN GmbH, Hilden, Germany). Illumina Infinium HumanMethylation450 BeadChip analysis was conducted at the ALSPAC/IEU Laboratory at the University of Bristol, according to the protocol described above for NOWAC.

EPIC HD laboratory procedures were carried out at the German Cancer Research Center (DKFZ; Heidelberg, Germany) and at LGC Bioscience (United Kingdom). Buffy coat DNA was isolated at LGC Bioscience by the company's standardized protocols and returned to DKFZ. DNA methylation profiling with the Illumina Infinium HumanMethylation450 BeadChip array was performed according to the manufacturer's instructions at the DKFZ Genomics and Proteomics Core Facility. Quality control of genomic DNA included three independent measurements with Quant-iT™ Picogreen® dsDNA assay and all samples were tested on 1% agarose gels for DNA integrity. The Zymo Research EZ-96 DNA Methylation™ Kit was used for bisulfite conversion of DNA. All subsequent steps were performed as described for NOWAC.

NOWAC data pre-processing was carried out using in-house software written for the R statistical computing environment. For each sample and each probe, measurements were set to missing if obtained by averaging intensities over less than three beads, or if averaged intensities were below detection thresholds estimated from negative control probes. Background subtraction (to remove background noise) and dye bias correction (for probes using the Infinium II design) were also performed. The resulting subset of 473,929 probes targeting autosomal CpG loci was selected for further analyses, and among these, probes with missing values in more than 20% of the samples were excluded from the analyses, leaving 450,890 probes. Samples with more than 5% of non-detected probes were also excluded from the analysis (14 samples excluded).

For the MCCS, methylation data were normalised to the internal built-in controls as provided by the standard Illumina software and subset-quantile within array normalization (SWAN) for type I and II probe bias correction [33]. The 65 CpGs corresponding to single nucleotide polymorphisms were excluded. Methylation measures were assigned as missing for CpG sites with a detection p-value higher than 0.01. No samples failed (a sample was considered as "failed" if more than 5% of the CpG measures were missing) and 182 (0.04%) CpG sites where excluded because values were missing for more than 20% of the samples, thus leaving 485,330 CpGs suitable for the analysis. Only the 458 male samples were considered when filtering probes in the Y chromosome.

In the NSHDS, methylation data were normalized using a functional normalization procedure that uses the built-in control probes to remove unwanted technical variation [34]. CpG sites that mapped to multiple genomic regions were excluded [35]. CpG sites with a detection p-value >0.01 were set to missing. CpG sites were excluded if they were missing in more than 20% of samples. Samples were excluded if more than 5% of their CpG sites were missing or if their average detection p-value was >0.01. Samples were also dropped if their case-control pair was missing. Of 490 samples initially available, 22 were excluded on the basis of the aforementioned procedures, leaving a total of 234 matched case-control pairs for analysis. Methylation levels at each locus were quantified using the beta-values [36].

In EPIC HD, the quality control measures included removal of SNP-containing probes, removal of CpGs not analysed in all samples or those in non-CpG context, correction for batch effects and normalization with beta quantile dilation method: 63 sample pairs entered the final differential methylation analysis.

Statistical analysis

Association study

In the NOWAC study, unconditional logistic regression models were used for all analyses, with DNA methylation levels included as an independent variable and standardized to 1 standard deviation. To account for residual technical confounding, all models were adjusted for micro-array and position of the sample on the micro-array. All analyses were additionally adjusted for blood cell composition differentials estimated using the algorithm developed by Houseman et al. [37] by including in the model the percentage of each cell type. The Houseman prediction model was calibrated using DNA methylation profiles of purified human leukocytes from six healthy male blood donors, and predictions were obtained using the subset of 89,490 probes found to be differentially methylated across cell types at a stringent Bonferroni-corrected significance threshold ensuring a family wise error rate below 0.01. Further adjustment included matching variables (year of birth, date of blood collection). Multiple testing was accounted for by using a stringent strategy: Bonferroni correction with control of the family wise error rate below 0.05.

In NOWAC we also built a predictive model based on smoking status and methylation of *AHRR* and *F2RL3,* and estimated the areas under the curve (AUC) with and without gene methylation. This was not feasible for the other cohorts because of matching by smoking.

In MCCS, conditional logistic regression was applied to estimate ORs of lung cancer. A stratified analysis by smoking status (never/former/current smokers) was also performed with further adjustment for number of cigarettes smoked (<15, 15-24, 25 or more per day), duration of smoking (less than 30 years, 30-39, 40 or more) and time since quitting (less than 5 years, 5-14, 15 or more). Associations between smoking and methylation levels were assessed by fitting linear mixed effect models with random intercepts to the M-values of methylation (M=log2(beta/(1-beta)) [36]) with three levels of clustering due to matching sets being within batch and these within plate. The model was also controlled for the fixed effects of age at blood collection, gender and the smoking variables.

In the NSHDS, ORs for lung cancer were estimated by conditional logistic regression. Due to the case-control matching, all models were adjusted for age, sex, smoking status (never/former/current smokers) and smoking quantity (1-14 versus >14 cigarettes per day) by design. To estimate the separate effects by smoking status, models were run separately for never, former and current smokers, with adjustment for time since quit smoking (in former smokers only) and smoking duration (in former and current smokers).

In EPIC HD, blood cell type composition of every sample was estimated [37] using Granulocytes, CD4+ and CD8+ T-cells, Natural Killer cells (NK) and Monocytes. A principal component (PC) analysis of the cell types was performed and the first two PCs were included in a linear regression model of methylation differences for every CpG. The risk of lung cancer was then modeled using conditional logistic regression on standardized residuals obtained from the cell type regression, adjusting for the average number of cigarettes smoked, duration of smoking and time since smoking cessation (for former smokers). Lung cancer risk was investigated using the overall study population as well as for subgroups determined by smoking status at baseline (current or former) and odds ratios and 95% confidence intervals were computed.

Mediation analysis

We performed mediation analysis to assess whether methylation of cg05575921 (*AHRR*) and cg03636183 (*F2RL3*) probes mediated the effect of smoking (ever smoking versus never smoking) on lung cancer risk using parametric G-computation [38] achieved by Monte Carlo simulations [39] and adapted to deal with the case-control design following VanderWeele and Vamsteelandt [40]. This requires the specification of a model for the mediator and one for the outcome. Linear regression was used to model methylation levels as a function of smoking status, age and their interaction, and logistic regression to model lung cancer status as a function of age, smoking status, methylation and their interactions. The linear regressions for methylation were weighted to account for the study design; cases were weighted by the prevalence of lung cancer and controls were weighted by 1 minus the prevalence.

We quantified the amount by which either or both of the two methylation probes mediated the effect of smoking on lung cancer incidence by partitioning the total causal effect (TCE) of smoking into a natural indirect effect (NIE) and a natural direct effect (NDE) [41], [42]. We expressed these quantities on the log odds ratio scale because of the case-control design, although they can be interpreted as log rate ratios (because cases are incident lung cancers).

The natural direct effect (NDE) is the effect of smoking on lung cancer (on the log OR scale) when methylation takes the natural value it would have taken in the absence of smoking; while the natural indirect effect (NIE) quantifies the change that would be found in log odds of lung cancer for smokers if we could change their methylation level to be that of never smokers. The total causal effect (TCE) is the sum of these effects. The proportion of the total effect explained by the hypothesized mechanism (proportion mediated) is given by the ratio between NIE and TCE (on the log scale). Identification of the mediated proportion required structural and parametric assumptions, namely: no unmeasured exposure-mediator, mediator-outcome, and exposure-outcome confounding; correct model specification for each of the outcome and the mediator(s) [41], [42].

In our analysis it is possible that unmeasured confounders could lead to inaccurate estimates of the effects: in particular, regarding exposure-mediator confounders, information such as smoking intensity, duration of smoking and passive smoking would probably affect the final estimates. The ideal situation would be to create an exposure variable that summarizes all this information and to repeat mediation analysis using the new variable as the exposure variable. In our case, the presence of several missing values in NOWAC data prevented us from performing this type of analysis. Air pollution might be a confounder of the mediator-outcome relationship, but we assumed that it would be a negligible factor in Norway.

3.5 Tables

Table 3.1 - Top-ranked CpG sites for the locus-by-locus risk analysis in NOWAC data (discovery set): CpGs in the *AHRR* and *F2RL3* genes display the most significant inverse associations with risk (hypomethylation in cases). Unconditional logistic regression models were used with DNA methylation levels included as an independent variable and were adjusted for matching variables, micro-array, position of the sample on the micro-array and blood cell composition differentials.

Probe Name	Gene Name	Chromosome	Position	Region	OR for 1 SD	95% CI	P-value	P-value Bonferroni
cg05575921	***AHRR***	5	373378	N_Shore	0.37	0.31-0.54	3.33x10^{-11}	1.36x10^{-5}
cg03636183	***F2RL3***	19	17000585	N_Shore	0.40	0.31-0.56	3.86x10^{-10}	1.58x10^{-4}
cg21566642		2	233283329	Island	0.36	0.23-0.48	1.33x10^{-9}	5.43x10^{-4}
cg06126421		6	233284934		0.41	0.25-0.49	1.52x10^{-9}	6.21x10^{-4}
cg25305703	*CASC21*	8	233284402		0.45	0.35-0.60	3.28x10^{-8}	1.34x10^{-2}
cg21161138	*AHRR*	5	399360		0.46	0.36-0.62	5.01x10^{-8}	2.04x10^{-2}
cg01940273		2	26578098	Island	0.44	0.33-0.60	5.21x10^{-8}	2.13x10^{-2}
cg02451831	*KIAA0087*	7	30720080		0.43	0.29-0.57	6.55x10^{-8}	2.67x10^{-2}
cg05951221		2	233284661	Island	0.41	0.30-0.58	8.59x10^{-8}	3.51x10^{-2}
cg04884171	*BOLA2*	16	128378218	S_Shelf	0.33	0.15-0.41	1.18x10^{-8}	4.82x10^{-2}
cg03898802	*DOPEY2*	21	37617652	Island	0.37	0.29-0.57	1.20x10^{-7}	4.90x10^{-2}

Table 3.2 - Results of the lung cancer risk analysis for the *AHRR* and *F2RL3* gene probes after strict adjustment for smoking in the discovery set and in the replication sets (ca=cases; co=controls)

.	**NOWAC**					**MCCS**				
	ca	**co**	**OR**	**95% CI**	**p-value**	**ca**	**co**	**OR**	**95% CI**	**p-value**
AHRR **cg05575921**										
Unadjusted	125	125	0.37	0.31-0.54	3.33x10^{-11}					
Adjusted*	124	122	0.39	0.24-0.61	2.55x10^{-05}	367	367	0.62	0.50-0.78	2.91x10^{-5}
Never	11	54	0.90	0.26-3.10	8.70x10^{-01}	43	43	0.63	0.24-1.64	3.47x10^{-1}
Former**	41	33	0.23	0.10-0.56	1.00x10^{-02}	153	153	0.48	0.31-0.75	1.45x10^{-3}
Current***	72	35	0.46	0.24-0.88	1.90x10^{-02}	164	164	0.75	0.56-0.99	4.13x10^{-2}
F2RL3 **cg03636183**										
Unadjusted	125	125	0.40	0.31-0.56	3.86x10^{-10}					
Adjusted*	124	122	0.51	0.35-0.73	4.19x10^{-04}	367	367	0.70	0.58-0.85	2.21x10^{-4}
Never	11	54	1.07	0.29-4.00	9.20x10^{-01}	43	43	0.78	0.44-1.36	3.73x10^{-1}
Former**	41	33	0.25	0.35-0.55	1.00x10^{-03}	153	153	0.70	0.50-0.98	3.81x10^{-2}
Current***	72	35	0.55	0.32-0.94	3.00x10^{-02}	164	164	0.81	0.61-1.06	1.18x10^{-1}
	NSHDS					**EPIC HEIDELBERG**				
	ca	**co**	**OR**	**95% CI**	**p-value**	**ca**	**co**	**OR**	**95% CI**	**p-value**
AHRR **cg05575921**										
Unadjusted										
Adjusted*	234	234	0.42	0.30-0.58	2.06x10^{-7}	63	63	0.45	0.22-0.92	2.95x10^{-02}
Never	26	26	1.96	0.40-9.68	1.10x10^{-1}	.	.	.	.	.
Former**	70	70	0.27	0.12-0.61	1.70x10^{-3}	16	16	0.06	0.00-2.23	1.26x10^{-01}
Current***	120	120	0.47	0.31-0.72	5.40x10^{-4}	47	47	0.56	0.27-1.16	1.16x10^{-01}
F2RL3 **cg03636183**										
Unadjusted										
Adjusted*	234	234	0.61	0.47-0.79	1.56x10^{-4}	63	63	0.62	0.38-1.04	7.02x10^{-02}
Never	26	26	1.38	0.51-3.73	5.20x10^{-1}	.	.	.	.	.
Former**	70	70	0.45	0.26-0.80	6.60x10^{-3}	16	16	0.29	0.03-3.24	3.17x10^{-01}
Current***	120	120	0.70	0.49-0.98	3.70x10^{-2}	47	47	0.64	0.36-1.12	1.17x10^{-01}

* In NOWAC the estimates are from the unconditional logistic regression models adjusted for smoking status coded as never, former, current; in MCCS the estimates are from the conditional logistic regression models where controls were matched on age, sex, date of blood collection, country of birth, type of biospecimen and smoking status as described in the text; in NSHDS, estimates are from conditional logistic regression models where cases and controls were matched on age, sex, smoking status and smoking quantity; in EPIC HD the estimates are from conditional regression models where cases and controls where matched on smoking status and packyears of smoking

** In MCCS and EPIC HD the estimates are also adjusted for number of cigarettes smoked, duration of smoking and time since quitting smoking; in NSHDS estimates are also adjusted for duration of smoking and time since quitting smoking

*** In MCCS and EPIC HD the estimates are also adjusted for number of cigarettes smoked and duration of smoking; in NSHDS estimates are also adjusted for duration of smoking

Table 3.3 Mediation analysis of the NOWAC cohort based on g-formula. Total causal effect (TCE), natural direct effect (NDE) and natural indirect effect (NIE) for the cg05575921 probe in *AHRR*, for the cg03636183 probe in *F2RL3* and for the two probes combined: 31% of the total effect of smoking on lung cancer risk is mediated by *AHRR* site-specific methylation, 32% of the total effect of smoking on lung cancer risk is mediated by *F2RL3* site-specific methylation and 37% of the total effect of smoking on lung cancer risk is mediated by the combined contribution of *AHRR* and *F2RL3* methylation (separate pathways for the two probes).

<table>
<tr><th colspan="5"><i>AHRR</i>-cg05575921</th></tr>
<tr><th></th><th>log OR</th><th>Std. Err.</th><th>pvalue</th><th>95% CI</th></tr>
<tr><td><b>TCE</b></td><td>1.83</td><td>0.29</td><td><0.001</td><td>(1.37 – 2.64)</td></tr>
<tr><td><b>NDE</b></td><td>1.26</td><td>0.31</td><td><0.001</td><td>(0.75 – 2.08)</td></tr>
<tr><td><b>NIE</b></td><td>0.56</td><td>0.08</td><td><0.001</td><td>(0.39 – 0.73)</td></tr>
<tr><td><b>effect mediated</b></td><td>0.31</td><td>0.08</td><td><0.001</td><td>(0.18-0.46)</td></tr>
<tr><th colspan="5"><i>F2RL3</i>-cg03636183</th></tr>
<tr><th></th><th>log OR</th><th>Std. Err.</th><th>pvalue</th><th>95% CI</th></tr>
<tr><td><b>TCE</b></td><td>1.82</td><td>0.30</td><td><0.001</td><td>(1.29 – 2.48)</td></tr>
<tr><td><b>NDE</b></td><td>1.23</td><td>0.33</td><td><0.001</td><td>(0.63 – 1.93)</td></tr>
<tr><td><b>NIE</b></td><td>0.59</td><td>0.09</td><td><0.001</td><td>(0.43 – 0.80)</td></tr>
<tr><td><b>effect mediated</b></td><td>0.32</td><td>0.08</td><td><0.001</td><td>(0.20-0.53)</td></tr>
<tr><th colspan="5"><i>AHRR</i>-cg05575921 and <i>F2RL3</i>-cg03636183</th></tr>
<tr><th></th><th>log OR</th><th>Std. Err.</th><th>pvalue</th><th>95% CI</th></tr>
<tr><td><b>TCE</b></td><td>1.79</td><td>0.30</td><td><0.001</td><td>(1.28 – 2.53)</td></tr>
<tr><td><b>NDE</b></td><td>1.13</td><td>0.34</td><td>0.001</td><td>(0.49 – 1.86)</td></tr>
<tr><td><b>NIE</b></td><td>0.66</td><td>0.15</td><td><0.001</td><td>(0.42 – 1.09)</td></tr>
<tr><td><b>effect mediated</b></td><td>0.37</td><td>0.11</td><td>0.001</td><td>(0.19 - 0.66)</td></tr>
</table>

3.6 Figures

Figure 3.1 - NOWAC cohort: associations between smoking cessation (years since quitting on horizontal axis) and methylation levels (vertical axis).

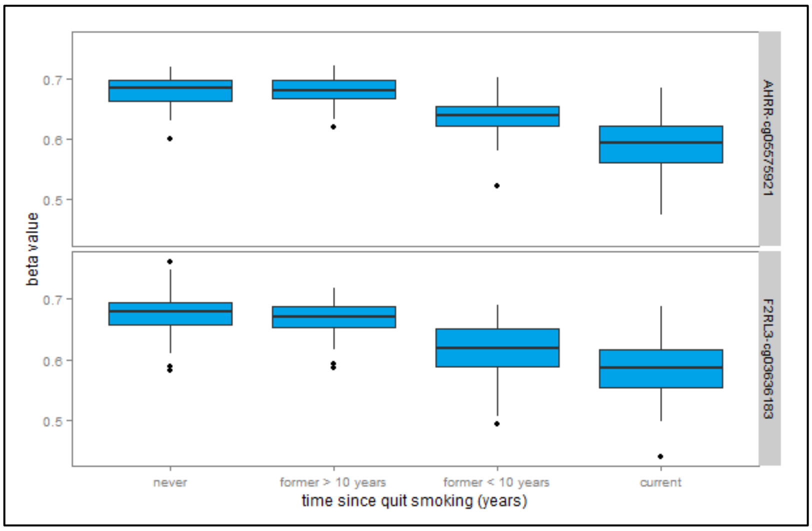


Figure 3.2 – MCCS and NSHDS cohorts: associations between duration of smoking and time since smoking cessation and methylation levels in *AHRR*-cg05575921 and *F2RL3*-cg03636183.

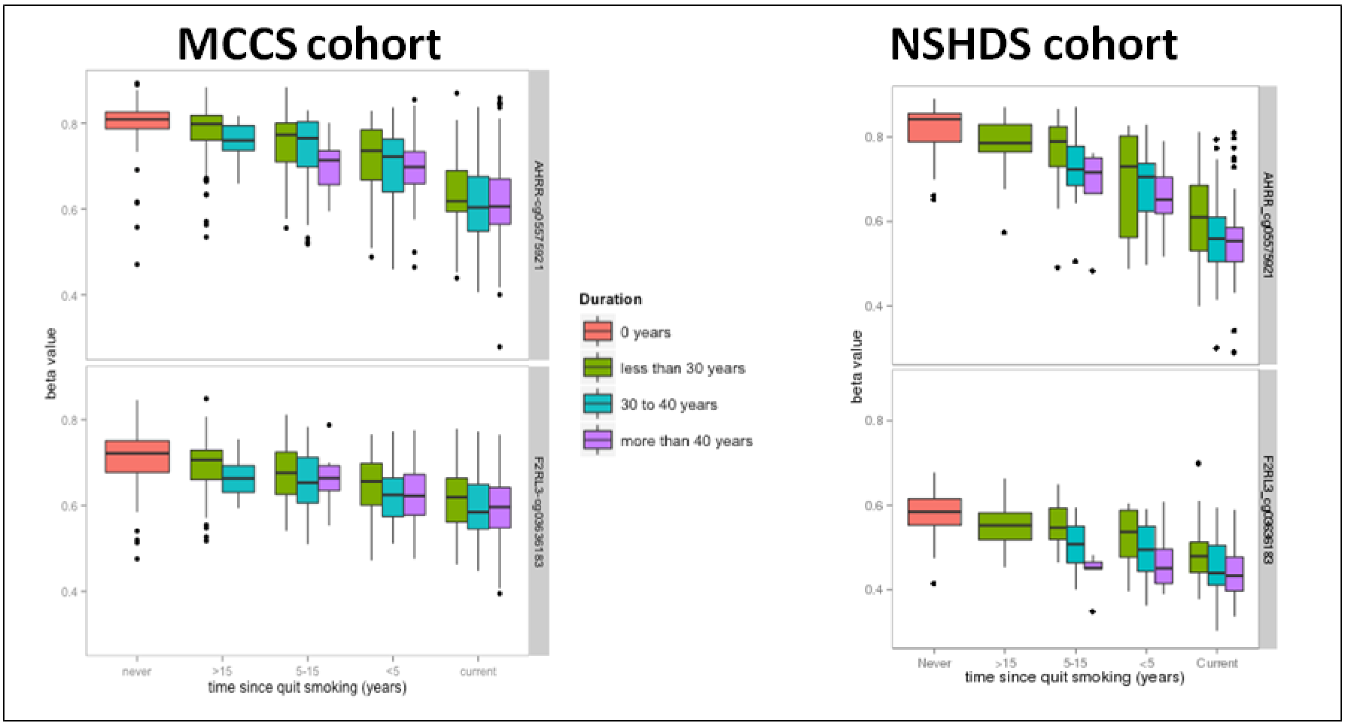
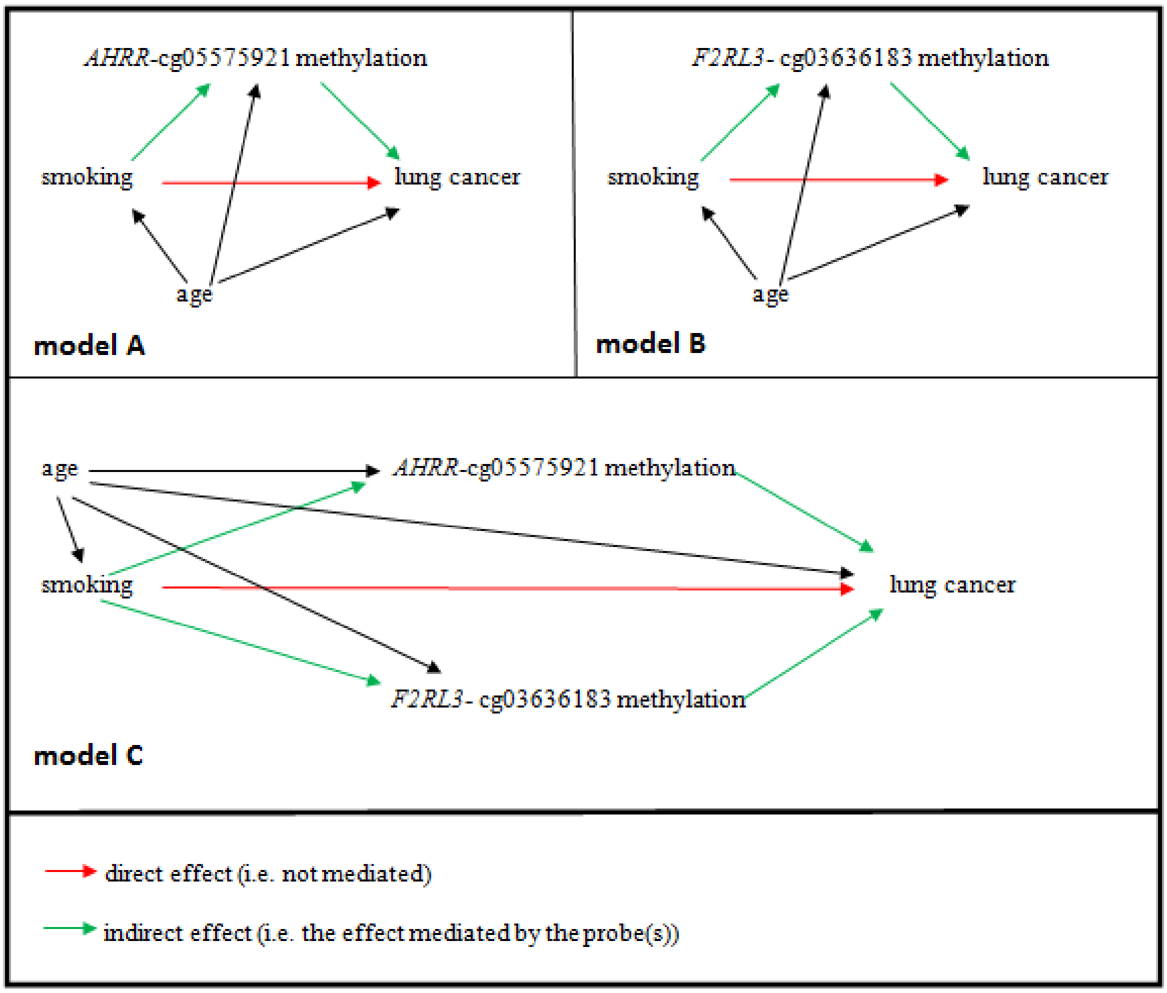


Figure 3.3 - Mediation analysis: graphical representation. In **model A** the percentage of the effect mediated by *AHRR*-cg05575921 is about 31% of the total effect of smoking on lung cancer risk, while in **model B** the percentage mediated by *F2RL3*-cg03636183 is about 32%. The joint mediation effect of these two CpGs is 37% if the two mediators are included together in the model with separate pathways (**model C**).


AHRR-cg05575921 methylation
smoking
lung cancer
age
model A
F2RL3- cg03636183 methylation
smoking
lung cancer
age
model B
age
AHRR-cg05575921 methylation
smoking
lung cancer
F2RL3- cg03636183 methylation
model C
direct effect (i.e. not mediated)
indirect effect (i.e. the effect mediated by the probe(s))

3.7 Supplementary material

Supplementary Table 3.1 - Main information about involvement in cancer pathways for the top-ranked CpGs found after the locus-by-locus risk analysis in NOWAC data (discovery set).

Probe Name	**Gene Name**	**Information**	**References**
cg05575921 cg21161138	*AHRR*	*AHRR* mediates dioxin toxicity and is involved in regulation of cell growth and differentiation.	Zudaire E et al. *The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers.*
cg03636183	*F2RL3*	*F2RL3* codes for the thrombin PAR-4 The function of PAR-4 is not fully clear yet but there is emerging evidence that it might be involved in the pathophysiology of several malignant tumors including lung cancer	Zhang Y et al. *F2RL3 methylation, lung cancer incidence and mortality.*
cg21566642 cg01940273 cg05951221		the closest gene is *ALPPL2* *ALPPL2* is a protein coding gene whose expression is strongly correlated with that of Heme Oxygenase-1 gene (it is expressed in many cancers and promotes growth and survival of neoplastic cells)	Tauber S et al. *Transcriptome analysis of human cancer reveals a functional role of heme oxygenase-1 in tumor cell adhesion.*
cg06126421		the closest gene is *FLOT1* *FLOT1* seems to have a role in non-small cell lung cancer tumorigenesis	Li H et al. *Abnormal expression of FLOT1 correlates with tumor progression and poor survival in patients with non-small cell lung cancer.*
cg25305703	*CASC21*	*CASC21* (cancer susceptibility candidate 21) has an oncogenic function	Kim T et al. *Long-range interaction and correlation between MYC enhancer and oncogenic long noncoding RNA CARLo-5*
cg02451831	*KIAA0087*	*KIAA0087* is a RNA Gene, and belongs to non coding RNA class currently no evidence of involvement in cancer tumorigenesis	
cg04884171	*BOLA2*	*BOLA2* encodes the BolA-like protein 2; this protein is conserved from prokaryotes to eukaryotes and seems to be involved in cell proliferation or cell-cycle regulation	Hunecke D et al. *MYC-regulated genes involved in liver cell dysplasia identified in a transgenic model of liver cancer*
cg03898802	*DOPEY2*	*DOPEY2* is a protein coding gene currently no evidence of involvement in cancer tumorigenesis	

Supplementary Table 3.2 - Analysis stratified by time to diagnosis.

		ca	co	OR for 1 SD
cg05575921-AHRR	all	125	125	0.37(0.31-0.54)
	time to diagnosis <5 years	84	125	0.20(0.10-0.37)
	time to diagnosis >=5 years	41	125	0.42(0.30-0.56)
		heterogeneity		p=0.021
cg03636183-F2RL3	all	125	125	0.40(0.31-0.56)
	time to diagnosis <5 years	84	125	0.32(0.19-0.54)
	time to diagnosis >=5 years	41	125	0.42(0.30-0.57)
		heterogeneity		p=0.375

Supplementary Table 3.3 - Summary of the key characteristics of the study groups.

	cohorts			
	NOWAC	**MCCS**	**NSHDS**	**EPIC-HEIDELBERG**
number of eligible cases	132	367	245	66
number of cases considered in the analysis*	125*	367*	234*	63
	nested case-control studies			
age at baseline (years)	47 (range: 34 -61)	59 (range: 39 - 70)	55 (range: 29-64)	56 (range: 39-65)
age at diagnosis (years)	56 (range: 47 - 64)	69 (range: 48 - 80)	64 (range: 42-81)	61 (range: 45-70)
time from blood draw to diagnosis (years)	3.88 (range: 0.29 - 7.92)	9.38 (range: 0.01 - 18.67)	9.6 (range: 1.1-17.5)	4.8 (range: 1.1-8.6)
women (N)	250	276	230	22
men (N)	0	458	238	104

References

1. Monick, M.M. et al., Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers, Am J Med Genet B Neuropsychiatr Genet 159B (2), 141 (2012).
2. Elliott, H.R. et al., Differences in smoking associated DNA methylation patterns in South Asians and Europeans, Clin Epigenetics 6 (1), 4 (2014).
3. Dogan, M.V. et al., The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women, BMC Genomics 15, 151 (2014).
4. Zeilinger, S. et al., Tobacco smoking leads to extensive genome-wide changes in DNA methylation, PLoS One 8 (5), e63812 (2013).
5. Philibert, R.A., Beach, S.R., Lei, M.K., and Brody, G.H., Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking, Clin Epigenetics 5 (1), 19 (2013).
6. Besingi, W. and Johansson, A., Smoke-related DNA methylation changes in the etiology of human disease, Hum Mol Genet 23 (9), 2290 (2014).
7. Shenker, N.S. et al., Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking, Hum Mol Genet 22 (5), 843 (2013).
8. Philibert, R.A., Beach, S.R., and Brody, G.H., Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers, Epigenetics 7 (11), 1331 (2012).
9. Guida and al., e., Dynamics of Smoking-Induced Genome-Wide Methylation Changes with Time Since Smoking Cessation, Human Molecular Genetics (in press) (2015).
10. Ezzati, M. and Lopez, A.D., Estimates of global mortality attributable to smoking in 2000, Lancet 362 (9387), 847 (2003).
11. Mathers, C.D. and Loncar, D., Projections of global mortality and burden of disease from 2002 to 2030, PLoS Med 3 (11), e442 (2006).
12. Newcomb, P.A. and Carbone, P.P., The health consequences of smoking. Cancer, Med Clin North Am 76 (2), 305 (1992).
13. Vineis, P. et al., Tobacco and cancer: recent epidemiological evidence, J Natl Cancer Inst 96 (2), 99 (2004).
14. Conen, D. et al., Smoking, smoking cessation, [corrected] and risk for symptomatic peripheral artery disease in women: a cohort study, Ann Intern Med 154 (11), 719 (2011).
15. Kawachi, I. et al., Smoking cessation and decreased risk of stroke in women, JAMA 269 (2), 232 (1993).
16. Vernooy, J.H. et al., Local and systemic inflammation in patients with chronic obstructive pulmonary disease: soluble tumor necrosis factor receptors are increased in sputum, Am J Respir Crit Care Med 166 (9), 1218 (2002).
17. Willemse, B.W. et al., Effect of 1-year smoking cessation on airway inflammation in COPD and asymptomatic smokers, Eur Respir J 26 (5), 835 (2005).
18. Ebbert, J.O. et al., Lung cancer risk reduction after smoking cessation: observations from a prospective cohort of women, J Clin Oncol 21 (5), 921 (2003).
19. Vermeulen, R. and Chadeau-Hyam, M., Dynamic aspects of exposure history-do they matter?, Epidemiology 23 (6), 900 (2012).
20. Vlaanderen, J. et al., Effect modification of the association of cumulative exposure and cancer risk by intensity of exposure and time since exposure cessation: a flexible method applied to cigarette smoking and lung cancer in the SYNERGY Study, Am J Epidemiol 179 (3), 290 (2014).
21. Zhang, Y., Yang, R., Burwinkel, B., Breitling, L.P., and Brenner, H., F2RL3 methylation as a biomarker of current and lifetime smoking exposures, Environ Health Perspect 122 (2), 131 (2014).

22. Wan, E.S. et al., Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome, Hum Mol Genet 21 (13), 3073 (2012).
23. Talikka, M. et al., Genomic impact of cigarette smoke, with application to three smoking-related diseases, Crit Rev Toxicol 42 (10), 877 (2012).
24. Vineis, P., Schatzkin, A., and Potter, J.D., Models of carcinogenesis: an overview, Carcinogenesis 31 (10), 1703 (2010).
25. Hankinson, O., The aryl hydrocarbon receptor complex, Annu Rev Pharmacol Toxicol 35, 307 (1995).
26. Zhang, Y. et al., F2RL3 methylation in blood DNA is a strong predictor of mortality, Int J Epidemiol 43 (4), 1215 (2014).
27. Relton, C.L. and Davey Smith, G., Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease, Int J Epidemiol 41 (1), 161 (2012).
28. Giles, G. and Enghlish, D., The Melbourne Collaborative Cohort Study, IARC scientific publications. 156, 69 (2002).
29. Hallmans, G. et al., Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions, Scand J Public Health Suppl 61, 18 (2003).
30. Riboli E, Kaaks R, The EPIC Project: Rationale and Study Design, Int J Epidemiol. 1997;26 Suppl 1:S6-14
31. Boeing H, Wahrendorf J, Becker N, EPIC-Germany--A source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition, Ann Nutr Metab. 1999;43(4):195-204
32. Joo, J.E. et al., The use of DNA from archival dried blood spots with the Infinium HumanMethylation450 array, BMC Biotechnol 13, 23 (2013).
33. Maksimovic, J., Gordon, L., and Oshlack, A., SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips, Genome Biol 13 (6), R44 (2012).
34. Fortin, J.P. et al., Functional normalization of 450k methylation array data improves replication in large cancer studies, Genome Biol 15 (12), 503 (2014).
35. Naeem, H. et al., Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array, BMC Genomics 15, 51 (2014).
36. Du, P. et al., Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, BMC Bioinformatics 11, 587 (2010).
37. Houseman, E.A. et al., DNA methylation arrays as surrogate measures of cell mixture distribution, BMC Bioinformatics 13, 86 (2012).
38. Robins, J., A new approach to causal inference in mortality studies with a sustained exposure period — application to control of the healthy worker survivor effect, Mathematical Modelling 7, 1393 (1986).
39. Daniel, R., De Stavola, B., and Cousens, S., gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula, The Stata Journal 11, 479 (2011).
40. Vanderweele, T.J. and Vansteelandt, S., Odds ratios for mediation analysis for a dichotomous outcome, Am J Epidemiol 172 (12), 1339 (2010).
41. Robins, J.M. and Greenland, S., Identifiability and exchangeability for direct and indirect effects, Epidemiology 3 (2), 143 (1992).
42. Pearl, J.,Direct and indirect effects presented at the Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, San Francisco, CA, 2001

4. Project 2: Mediterranean Diet, DNA methylation and colon cancer

4.1 Introduction

The Mediterranean Diet (MD) is a dietary scheme that is recognized to be relevant in cancer prevention. Epidemiological cohort studies conducted in different countries revealed an association between a greater adherence to MD and a reduced risk of mortality and incidence of cancer and other major chronic diseases [1], [2]. In the Italian branch of the European Prospective Investigation into Cancer and Nutrition (EPIC) study, increasing adherence to the Italian Mediterranean Index was shown to be associated with a significantly decreased risk of colorectal cancer in men and women [3]. A pooled analysis of three Italian case-control studies [4] as well as two recent cohort studies [5], [6] confirmed a favourable role of MD on colorectal cancer.

To date the biological mechanisms through which MD protects against colorectal cancer remain poorly understood. To clarify the protection role of MD, in a previous study [7] we analyzed abdominal adiposity as a potential biological mediator of the association between adherence to MD and colon cancer onset, concluding that abdominal adiposity does not explain this relationship.

Another possible explanation of the MD influences into carcinogenesis is the action through epigenetic mechanisms. In fact diet, as other environmental factors, can perturb the way genes are controlled by DNA methylation, noncoding RNAs and histone modifications, resulting in deregulation of key cellular processes and promotion of oncogenic transformation. Epigenetic events can affect many steps in tumour development; therefore, better understanding of epigenetic mechanisms is fundamental for our ability to successfully prevent, diagnose and treat cancer [8].

Chronic inflammation may also explain the association between MD and colon cancer. Indeed in some randomized control trials [9] and observational studies [10] it has been shown that MD can attenuate the level of the systemic inflammation. Inflammation is a complex stereotypical reaction of the body expressing the response to a possible damage of its cells and tissues. A number of various mediators are implicated/involved in this phenomenon and there is an increasing evidence for a specific epigenetic modulation [11], [12]. Chronic inflammation has been shown in turn to be a possible causative factor in a variety of cancer types, among which colon cancer. Indeed people with chronic inflammatory bowel diseases, such as ulcerative colitis and Crohn's disease, have an increased risk of colon cancer [13] and chronic aspirin use seems to reduce the risk of colon cancer [14].

In particular, we hypothesized that MD may protect against colon cancer through a change in the methylation pattern of genes, in particular of inflammation-related genes.

To test our hypotheses, we used two different approaches: an aprioristic analysis on genome-wide methylation and a candidate-genes analysis on inflammation-related genes.

4.2 Materials and Methods

Study samples

Data from the Italian component of the EPIC study [15] (EPIC-Italy) were considered.

EPIC-Italy includes 47,745 volunteers from the centers of Turin, Varese, Ragusa, Florence, and Naples aged 35–70 years at the time of recruitment (1993–1998). Anthropometric measurements and lifestyle variables, including detailed information on diet, were collected at recruitment through standardized questionnaires, together with a blood sample that was sent to local laboratories for processing and aliquot preparation. Blood was separated into 0.5 ml fractions and stored in liquid nitrogen at −196°C. All participants signed an informed consent form; the ethical review boards of the International Agency for Research on Cancer and of each local participating centre approved the study protocol.

A nested case-control study (CACO1) was first conducted within EPIC-Turin utilizing 95 incident colon cases diagnosed within follow-up and 95 individually matched controls selected at random from the participants at risk of colon cancer at the time of the diagnosis of the cases. Controls were matched to cases by gender, date of birth (within 5 years) and seasonality of blood sampling (autumn-winter versus spring-summer). A second nested case-control study (CACO2) was conducted on 74 case-control pairs from EPIC-Varese and EPIC-Ragusa subjects. Matching variables were the same as CACO1 plus the study center (Varese or Ragusa).

Illumina HM450 epigenome-wide studies

Two epigenome-wide studies were conducted considering CACO1 and CACO2.

The laboratory procedures were carried out at the Human Genetics Foundation (Turin, Italy). Buffy coats stored in liquid nitrogen were thawed, and genomic DNA was extracted using the QIAsymphony DNA Midi Kit (Qiagen).

500 ng of DNA were bisulphite-converted using the EZ-96 DNA Methylation-Gold™ Kit (Zymo), and hybridised to Infinium HumanMethylation450 BeadChips. Samples were processed with Infinium HumanMethylation450 BeadChips (Illumina) with case-control pairs arranged randomly on the same chip. These were subsequently scanned using the Illumina HiScanSQ system, and sample quality was assessed using control probes present on the micro-arrays (11 samples, 4 matched pairs and 3 single individuals, were excluded for low bisulfite quality control resulting in a total of 179 remaining samples in CACO1; no sample excluded in CACO2). Finally, raw intensity data were exported from Illumina GenomeStudio (version 2011.1).

Illumina data pre-processing and quality control

Data pre-processing was carried out using an in-house software written for the R statistical computing environment. For each sample and each probe, measurements were set to missing if obtained by averaging intensities over less than three beads, or if averaged intensities were below detection thresholds estimated from negative control probes. Background subtraction (to remove background noise) and dye bias correction (for probes using the Infinium II design) were also performed. DNA methylation was expressed as a ratio of the intensities of methylated cytosines over the total intensities (ß values). In CACO1 only autosomal CpG loci were considered and probes with missing values in more than 20% of the samples were excluded. Furthermore two samples with more than 5% of non-detected probes and three decoupled samples were excluded from the analysis (87 case-control pairs analyzed in CACO1). The data were finally pruned eliminating probes with SNP in probe body (N=81,246), probes with SNP in target CpG (N=5,214) and probes with cross hybridization on XY and autosomal (N=28,724). No sample and no probe were excluded in CACO2.

Main exposure definition and other variables

The Italian Mediterranean Index (IMI) [16] was used as a measure of the exposure (MD). Briefly, this index is a score from 0 to 11 where higher scores indicate better adherence to MD. For details on the definition of IMI see [16]. In the statistical analyses IMI was categorized in the three following categories: 0-2 (low adherence to MD), 3-4 (middle adherence to MD) and 5-11 (high adherence to MD).

Other variables considered were: age, sex, center, smoking status (never, former and current), total physical activity (inactive, moderately inactive, moderately active and active [17]), level of education (tertiles of the relative index of inequality RII [18]), body mass index (BMI), seasonality, fasting status, year of recruitment and cell composition. In particular, the proportions of cell counts were estimated on the entire methylation data set (485,512 probes) according to the method suggested by Houseman [19].

<u>*EWAS approach*</u>

Agnostic analysis on genome-wide methylation

An agnostic search of signals associated to colon cancer and MD was conducted considering the sample CACO1 as a discovery set and the sample CACO2 as a replication set.

In CACO1 the association between DNA methylation of each probe and the risk of colon cancer was investigated using conditional logistic regression model adjusted for the exposure (IMI), BMI and differential cell counts with DNA methylation included as an independent variable (standardized to 1 standard deviation computed on the control group). Among the first 50 top signals, only those located in genes involved in colon carcinogenesis were selected for replication in CACO2.

The signals related to MD were searched in CACO1 using linear regression models for each probe adjusted for age, sex, differential cell counts and disease status (here the logarithmic transformation of ß values M=log2(ß /1- ß) was used [20] and the categorical variable IMI was taken as continuous). Among the first 50 top signals, only those located in genes involved in human diet metabolism were selected for replication in CACO2. In both the association analyses, we searched GeneCards (www.genecards.org) to find information and functional data concerning these genes and Pubmed (www.ncbi.nlm.nih.gov/pubmed, from January 2000 to June 2017) for information on the involvement of these genes in colon carcinogenesis or human diet metabolism (using as a search algorithm only the name of the gene in order to be more inclusive).

<u>*Gene candidate approach*</u>

Selection of CpG sites

To study the role of methylation in inflammation-related genes, data of the two epigenome-wide studies (87+74=161 case-control pairs, CACO1+CACO2) were pruned selecting only the CpG sites located in a set of inflammation genes known from literature to be related to solid human cancer and/or Mediterranean Diet (*IL-6, IL1B, NF-kB1, NF-kB2, TNFalfa, IL-10, TLR4, TLR2, PCK1, STAT3, PPARG, H1ST1H1A, JUN, NFATC1, NFE2L2, REL, RELA, RELB, IL-17a, IFNgamma, PTX3, IL22RA2, PGE2, SLIT2, RUNX1, RUNX2, RUNX3, TGFB2, IL-12B, IL-8, SERPINE1, PLA2G1B, PLA2G2A, IL-18, CRP, KLK10, LMO2, GPR21, GPR65, GPR81, GPR84, TRIM63, AQP3, SOCS3, BCL3, IRS2, MAL2, BIRC3*).

The purpose of this location-based pruning was to select CpG sites that may be inflammation mediators of the protective effect of MD on colon cancer. Among the 995 inflammation-related CpG sites, only those with a mean difference in methylation percentage between cases and controls higher than 1% were considered in the association analysis (32 CpGs). The limit of 1% was chosen to increase the probability that the difference in methylation reflected an effective biological change. A conditional logistic regression model with elastic net penalties [21] was used to select the most important CpG sites related to colon cancer. The DNA methylation levels of the CpGs were standardized to 1 standard deviation computed on the control group.

Elastic net (EN) [22] is a regularization and variable selection method, which retains the parsimony property of Lasso regression method [23] (for any given constraint value, only a subset of the covariates have non-zero coefficients), but at the same time encourages the grouping effect as Ridge regression [24]. We applied the cyclic coordinate descent algorithm [25] and we set the parameter which controls the trade off between Lasso and Ridge penalties equal to 0.5 (at value 1 pure Lasso penalty; at value 0 pure Ridge penalty). We used 10-fold cross validation (CV) for the choice of the regularization parameter lambda that characterizes the best model (that is the one with the minimum CV error). To assess whether the associations found were stable in random subsets of the sample, one thousand EN models were fitted using each time 63.4% of the initial data. At the end we obtained a ranked list of probes based on how many times they were included in each model based on data subset. The CpG sites considered for further detailed analyses had to satisfy at the same time the two following criteria: i) CpGs selected by EN (applied to the entire dataset), with a coefficient higher than 0.15 (median of the distribution of the estimated coefficients) in absolute value and ii) CpGs in the list of the most associated sites with a frequency higher than 50%.

Detailed analysis of the selected signals

For each selected CpG, a conditional logistic regression model was fitted to estimate odds ratio (OR) of colon cancer with DNA methylation levels included as an independent variable and standardized to 1 standard deviation (model A). Owing to the case–control matching, all models were adjusted for study center, gender, age and seasonality by design. The possible effect of cell composition on the results was assessed by adding to the models the proportions of cell counts (model B). Another model (model C) was fitted adjusting also for the additional covariates fasting status, year of recruitment, BMI, smoking status, physical activity, level of education and IMI. Sensitivity analyses were performed excluding cases with time elapsed between blood collection and diagnosis of colon cancer lower than 2 years or higher than 10 years.

The association between M-values of methylation and IMI was evaluated in the control group by fitting a linear mixed effect model with chip fitted as random effect and IMI, sex, age, center, seasonality, fasting status, year of recruitment, BMI, smoking status, physical activity, level of education and the differential cell types as fixed effects. Linearity of trends across categories of IMI was tested by treating the categorical variable as continuous in the linear mixed effect model.

Validation and replication

A random selection of case-control pairs from CACO1 was performed for laboratory validations with pyrosequencing (CACO3). The CpG sites selected for validation were those for which the effect of adherence to MD conferred methylation levels that were protective on colon cancer. This means that among the CpGs whose hypermethylation was protective on colon cancer (OR<1 for 1 standard deviation increase in methylation percentage) only those for which at higher adherence to MD corresponded higher methylation levels were validated. At the opposite among the CpGs whose hypermethylation was harmful to colon cancer (OR>1 for 1 standard deviation increase in methylation percentage) were validated only those for which at higher adherence to MD corresponded lower methylation levels.

Another nested case-control study (CACO4) was finally conducted employing 95 independent case-control pairs from EPIC-Italy (centers of Varese, Ragusa, Turin and Naples; matching variables: age, gender, center, seasonality and year of recruitment). The validated signals were replicated in this sample using pyrosequencing.

Both in validation and in replication analyses, the associations with DNA methylation levels obtained by pyrosequencing were analyzed using models similar to those employed in the discovery phase (see details in the Tables). Details on the pyrosequencing methodology are provided in the supplementary material.

All statistical analyses were performed using R Statistical Software (The R Foundation for Statistical Computing, Vienna, Austria) version 3.2.3 (2015-12-10) and Stata version 13 (StataCorp, College

Station, TX, USA). A diagram illustrating the data sets used in the different stages of the analysis is shown in Figure 1.

4.3 Results

EWAS approach

Table 4.1 shows the baseline characteristics of CACO1 set (Table 4.1a) and of CACO2 set (Table 4.1b) according to the case-control status. In the discovery set, cases and controls showed differences for BMI (p-value=0.0255) and for educational level (p-value=0.013). In the replication set cases and controls did not show any difference for all the variables considered. Importantly, the proportions of differential cell types estimated were not different between the two groups in both sets. For completeness, the features of the two samples together (CACO1+CACO2) are reported in Supplementary Table 4.1.

Table 4.2 shows the results of the association analyses with colon cancer for the subset of the 50 top-ranked CpG sites found in CACO1 and involved in colon carcinogenesis. Only two CpG sites (cg06287951-*DLX4*, cg01331191-*CBLL1*) showed a coherent association with the outcome in CACO2 set (OR obtained in CACO1 set = 15.04 (95% CI 3.23-69.98), OR obtained in CACO2 set = 2.45 (95% CI 1.17-5.12) for cg06287951-*DLX4*; OR obtained in CACO1 set = 2.80 (95% CI 1.54-5.09), OR obtained in CACO2 set = 1.57 (95% CI 1.02-2.41) for cg01331191-*CBLL1*).

The results of the association analyses with IMI for the subset of the 50 top-ranked CpG sites found in CACO1 set and involved in human diet metabolism are reported in Table 4.3. No CpG site showed a coherent association in CACO2 set.

Candidate-genes approach

The correlations among the 32 inflammation-related CpG sites considered are shown in Figure 4.2. Five groups with two or more CpG sites with positive correlations were evident, but the majority of the CpGs seemed to be uncorrelated between each other.

Figure 4.3 and Figure 4.4 show respectively the profile of parameter estimates plotted against the value of the regularization parameter lambda and the CV curve (multiplied by -1, so we look for a minimum) of the conditional logistic regression model with elastic net penalties. The algorithm starts at the value of lambda for which all parameter estimates are null (far right), proceeds at decreasing lambda and re-computes the estimates until an unconstrained maximum conditional likelihood estimate is reached (lambda=0). In particular it can be seen from Figure 4.4 that the CV error was minimized for a model with 26 predictors (lambda=4.53, log(lambda)=1.51).

The sixteen predictors having a coefficient higher than 0.15 in absolute value selected by the model are reported in Table 4.4. The mean differences in methylation percentages between cases and controls were small for the majority of the CpGs selected (< 2%), except for cg12195446-*IRS2* (7.6%) and cg12252547-*MAL2* (5.1%).

The detailed analysis of the CpG sites selected for validation is reported in Table 4.5. The results for the remaining CpG sites are reported in Supplementary Table 2.

In the association analyses between IMI and the M-values of methylation, increasing trends for higher IMI categories were observed for cg18773937-*IL1B*, cg17968347-*SERPINE1* and cg01265860-*RUNX1*. For CpG site cg24312520-*STAT3*, only the third category of IMI showed an increase in methylation levels with respect to the first one (coeff: 0.11, 95% CI: -0.08,0.30); for CpG sites cg15363134-*NFATC1* and cg20674490-*RUNX3* the second category of IMI showed a more pronounced increase in methylation levels with respect to the first category of IMI with respect to the increase of the third category compared to the first (coeff: 0.24 versus 0.07 for *NFATC1*-probe; coeff: 0.17 versus 0.08 for *RUNX3*-probe). A downward trend was observed for the CpG site cg08053846-*SERPINE1* (its hypermethylation is associated with colon cancer).

Considering CACO3 and DNA methylation measured by pyrosequencing, only two of the seven probes selected for validation showed coherent associations with both colon cancer and IMI (cg17968347-*SERPINE1* and cg20674490-*RUNX3,* Table 4.6). In fact although these associations were at the limit of significance, their direction and magnitude were essentially the same. Of the two validated CpG sites, only for cg20674490-*RUNX3* the association with colon cancer showed coherent direction and similar magnitude in the CACO4 sample (OR=0.80 95% CI 0.60-1.07 in CACO4 (Table 4.7, crude model) versus OR=0.74 95% CI 0.47-1.16 in CACO3 (Table 4.6, crude model)). For this CpG the association with MD showed an increasing trend in CACO4 not present in CACO3 and CACO1+CACO2 (Tables 4.5, 4.6, 4.7).

4.4 Discussion

The general aim of the present work was to assess whether the adherence to MD is associated with changes in the methylation status that could explain its protective effect on colon cancer onset.

We performed first an agnostic search for associations of DNA methylation with case-control status and with MD using an epigenome-wide methylation study based on methylation detection using Illumina Infinium HM450 on DNA extracted from pre-diagnostic blood of 87 pairs of colon cancer cases and controls from the EPIC-Italy cohort. We tried to replicate the top-signals in another epigenome-wide study from the same population (74 case-control pairs) detecting two CpG sites with an indication of a coherent association with colon cancer. The first CpG maps in *DLX4,* a gene whose expression was found to be related with colorectal carcinogenesis in [26]. The second one maps in *CBLL1,* a gene with multiple function in tumorigenesis and found to be highly up-regulated in human colon and gastric adenocarcinomas compared to normal tissues [27]. All the ten top-signals associated with MD in the discovery set were not confirmed in the replication set. Therefore the results of our agnostic search did not reveal any significant DNA methylation signals that could link MD to colon cancer.

Since it has been previously shown that MD may attenuate the level of chronic inflammation and chronic inflammation has been related to an increased risk of colon cancer, we hypothesized that the protective effect of MD adherence on colon cancer may be mediated by specific DNA methylation profiles of genes that are regulators of inflammation.

MD has been related to a decreased inflammatory activity in a large number of studies [9], [10]: it probably exerts an anti-inflammatory effect through the intake of monounsaturated fatty acids (such as oleic acid), polyphenols and fibers. MD might leave an epigenetic mark in cells, which would attenuate inflammatory responses and, in the long term, eventually protect from the development of colon cancer. To test this hypothesis, we considered CpG sites located in inflammation genes known from literature to be related to solid human cancer and/or MD. The results of our hypothesis-driven analysis seemed to support only in part our a priori hypothesis since, among the seven probes selected for validation, only two were confirmed by pyrosequencing (cg17968347-*SERPINE1* and cg20674490-*RUNX3*) and only one of the two showed similar associations in an independent sample (cg20674490-*RUNX3*). Gene *RUNX3* has important functions in innate and adaptive immune cell types, in particular in inactivating IL23A transcription, and has been associated with several immune-related diseases [28]. Gene *RUNX3* is also a tumor suppressor gene whose hypermethylation of the promoter was shown to be a key mechanism of its inactivation [29]. Therefore the CpG site located in this gene may be considered an epigenetic mediator of the protective effect of the MD on colon cancer.

Genome-wide assays are inherently imprecise and noisy [30]. For this reason we performed a validation phase and a replication phase using a locus-specific methylation technique (pyrosequencing). In literature there are very few studies that compare different DNA methylation assays for biomarker development [31], [32]. In general these studies show a good concordance between the measurements of the two arrays, but they consider DNA methylation assessed in solid tissues or specific cell lines. A study in which the CpG pyrosequencing-based validation of Illumina 450K array results based on DNA methylation of blood leukocyte are shown is [33], but only five samples are analyzed. In our study we considered a sample of 94 subjects and 7 CpG sites for validation and a sample of 190 subjects and 2 CpG sites for replication. The fact that only one signal was confirmed emphasizes the importance of the validation and replication phases using an alternative technology. These phases are essential in order to exclude technical errors and false positive findings especially when the differences in methylation percentages between cases and controls are little (about 1%) and so more likely due to background noise.

. Furthermore it is important to note that the interpretation of our data is challenging because the extent to which variations in DNA methylation translate into variation in gene expression levels is unknown and because we do not know which CpG sites are associated with the regulation of the expression of a given gene [34], [35]. In particular in this study, due to the biological sample collection and preservation techniques, it was not possible to investigate directly the relationship between DNA methylation and gene expression level.

. Because of the study design, it was also not possible to monitor DNA methylation level changes at different time points. This was an important limitation since only 17% of Illumina 450K probes were considered as stable variable methylated probes, i.e., as markers that vary in the population but are stable over time [36].

. We assessed DNA methylation patterns using blood samples since peripheral blood is a tissue of interest and in particular a valuable source of information for low-grade inflammation. However we were aware that heterogeneity in white blood cells could potentially confound DNA methylation measurements [37]. To address this problem we applied Houseman correction for cell composition verifying the stability of the associations after the adjustment [19]. We also adjusted for the major lifestyle-related risk factors, but we could not control for other factors potentially implicated in DNA methylation, such as environmental or psychosocial exposures, as they were not available in the study.

In conclusion, our study is a first attempt to identify the biological mechanism behind the protection of the MD against colon cancer investigating the methylation levels of genes in circulating lymphocytes years before the onset of the disease. The results of the study suggest that DNA methylation of *RUNX3* gene may be a potential molecular mediator explaining the protective effect of MD on colon cancer onset. However this finding is still uncertain since independent functional studies are needed to confirm its role in colon carcinogenesis.

4.5 Tables

Table 4.1: Descriptive statistics of the samples

a. **Discovery set (epigenome-wide analysis) (CACO1)**

Variables	Controls	Cases	all	p-value
N	87	87	174	
Median of (IQR)				
age, years*	56 (7)	56 (7)	56 (7)	
BMI, kg/m^2	25.16 (5)	26.42 (3)	26.09 (6)	0.0255
Counts of				
*Gender**				
men	65 (75%)	65 (75%)	130 (75%)	
women	22 (25%)	22 (25%)	44 (25%)	
*Centre**				
Turin	87 (100%)	87 (100%)	174 (100%)	
Educational level				0.013
1°tertile RII	28 (35%)	26 (33%)	54 (34%)	
2°tertile RII	31 (38%)	16 (20%)	47 (29%)	
3°tertile RII	22 (27%)	37 (47%)	59 (37%)	
Total physical activity				0.648
inactive	17 (21%)	16 (20%)	33 (20%)	
moderately inactive	32 (39%)	34 (42%)	66 (40%)	
moderately active	23 (28%)	17 (21%)	40 (25%)	
active	10 (12%)	14 (17%)	24 (15%)	
Smoking status				0.451
never smokers	28 (34%)	29 (36%)	57 (35%)	
former smokers	30 (36%)	35 (43%)	65 (40%)	
current smokers	24 (29%)	17 (21%)	41 (25%)	
Fasting status				0.444
yes	47 (54%)	52 (60%)	99 (57%)	
no	40 (46%)	35 (40%)	75 (43%)	
Cell Types				
CD8T	6.7%	7.6%	7.1%	0.307
CD4T	14%	13%	13%	0.053
NK	6.9%	6.9%	6.9%	0.699
Bcell	5.8%	5.3%	5.6%	0.724
Mono	6.6%	7.0%	6.7%	0.199
Gran	65.2%	65.1%	65.1%	0.583

b. Replication set (epigenome-wide analysis) (CACO2)

Variables	Controls	Cases	all	p-value
N	74	74	148	
Median of (IQR)				
age, years*	54.5 (12)	53.5 (12)	54 (12)	
BMI, kg/m^2	24.71 (5)	26.24 (6)	25.64 (5)	0.0554
Counts of				
*Gender**				
men	23 (31%)	23 (31%)	46 (31%)	
women	51 (69%)	51 (69%)	109 (69%)	
*Centre**				
Varese	64 (86%)	64 (86%)	128 (86%)	
Ragusa	10 (14%)	10 (14%)	20 (14%)	
Educational level				0.808
1°tertile RII	23 (33%)	27 (39%)	50 (36%)	
2°tertile RII	23(33%)	21 (30%)	44 (32%)	
3°tertile RII	23 (33%)	22 (31%)	45 (32%)	
Total physical activity				0.679
inactive	17 (23%)	21 (23%)	38 (26%)	
moderately inactive	33 (45%)	35 (47%)	68 (46%)	
moderately active	13 (18%)	12 (16%)	25 (17%)	
active	10 (14%)	6 (8%)	16 (11%)	
Smoking status				0.472
never smokers	43 (59%)	38 (51%)	81 (55%)	
former smokers	18 (25%)	18 (24%)	36 (24%)	
current smokers	12 (16%)	18 (24%)	30 (21%)	
Fasting status				0.154
yes	72 (97%)	74 (100%)	146 (99%)	
no	2 (3%)	0 (0%)	2 (1%)	
Cell Types				
CD8T	7.4%	8.9%	8.1%	0.253
CD4T	13%	15%	14%	0.299
NK	8.5%	8.7%	8.6%	0.951
Bcell	5.3%	5.1%	5.2%	0.942
Mono	7.9%	7.9%	7.9%	0.724
Gran	58.8%	59.4%	59.2%	0.378

Table 4.2: Subset of the 50 top-ranked CpG sites associated with colon cancer in the discovery set involved in colon carcinogenesis: odds ratio (OR) of the association in the discovery set (CACO1 174 samples) and in the replication set (CACO2 148 samples).

Associations between DNA methylation and colon cancer are assessed using a conditional logistic regression model with DNA methylation levels included as an independent variable and standardized to 1 standard deviation adjusting for the exposure (IMI), BMI and differential cell counts.

The replicated CpG sites are indicated in grey.

Probe Name	Gene	References	OR (discovery set)	p-value (discovery set)	OR (replication set)	p-value (replication set)
cg24041799	*MTOR*	Francipane MG et al. *Oncotarget 2014*	2.88 (1.60-5.19)	0.0004	1.45 (0.86-2.45)	0.1611
cg25333181	*PLXND1*	Rehman M et al. *Plos One 2016*	0.24 (0.10-0.54)	0.0005	1.20 (0.73-1.99)	0.4663
cg06287951	*DLX4*	Hollington P et al. *Anticancer Res. 2004*	15.04 (3.23-69.98)	0.0005	2.45 (1.17-5.12)	0.0168
cg15254238	*TET1*	Rawluszko-Wieczorek AA et al. *J Cancer Res Clin Oncol 2015*	2.94 (1.58-5.49)	0.0006	1.00 (0.65-1.53)	0.9885
cg27620871	*PRSS22*	Solmi R et al. *BMC Cancer 2006*	2.48 (1.46-4.19)	0.0007	0.89 (0.54-1.44)	0.6271
cg18592365	*ELTD1*	Pekow J et al. *Inflamm Bowel Dis 2013*	2.86 (1.55-5.25)	0.0007	1.07 (0.76-1.53)	0.6860
cg02351179	*RAB1B*	Zhai H et al. *Oncogene 2013*	0.32 (0.16-0.62)	0.0007	1.25 (0.85-1.85)	0.244
cg01331191	*CBLL1*	Aparicio LA et al. *Cancer Metastasis Rev. 2012*	2.80 (1.54-5.09)	0.0007	1.57 (1.02-2.41)	0.038
cg06100227	*MAPK13*	Del Reino P et al. *Cancer Research 2014*	2.64 (1.50-4.66)	0.0008	1.25 (0.83-1.88)	0.270

Table 4.3: Subset of the 50 top-ranked CpG sites associated with IMI in the discovery set involved in human diet metabolism: parameter estimate of the association in the discovery set (174 samples CACO1) and in the replication set (148 samples CACO2).

Associations between IMI and DNA methylation are assessed using a linear regression model adjusting for age, sex, differential cell counts and disease status. The logarithmic transformation of beta values (M-values) is used and the categorical variable IMI is treated as continuous.

Probe Name	**Gene**	**References**	**parameter estimate (discovery set)**	**p-value (discovery set)**	**parameter estimate (replication set)**	**p-value (replication set)**
cg02172492	OSR1	Davies M et al. *Am J Physiol Renal 2014*	0.15	3.3x10^-6	-0.02	0.4563
cg23466166	*PTK2*	Iorio V et al. *Cell Death Dis 2015*	0.37	4.2x10^-5	0.18	0.1021
cg21793437	*CACNA1C*	Ojo OO et al. *Biol Chem 2016*	0.12	4.8x10^-5	-0.01	0.6961
cg11469321	*BDH2*	O'Shea E et al. *J Anim Sci 2016*	0.06	9.5x10^-5	-0.04	0.0266
cg21428833	*LMBRD1*	Constantinou P et al. *Mol Syndromol 2016*	0.52	0.0001	-0.07	0.7809
cg18631798	*ACOX3*	Pivovarova O et al. *J Clin Endocrinol Metab 2015*	-0.40	0.0001	0.03	0.8676
cg22081905	*TRPM5*	Riper SD. *Handb Exp Pharmacol 2014*	-0.21	0.0001	-0.05	0.3861
cg10541332	*FHIT*	Le Roy CI et al. *Gut Microbes 2017*	0.30	0.0001	-0.11	0.0384
cg12034757	*NAGLU*	Fu CP et al. *Clin Chem Lab Med 2015*	0.08	0.0001	-0.002	0.9042
cg13545297	*HOXC8*	Yamamoto Y et al. *Obesity (Silver Spring) 2010*	0.22	0.0001	-0.14	0.0059

Table 4.4: CpG sites selected by EN with 10-fold CV with a coefficient higher than 0.15 in absolute value in CACO1+CACO2.

Delta= mean absolute difference in methylation percentage between cases and controls
Frequency= percentage of times in which the CpGs are selected when EN is applied on a random subset of the initial dataset
The CpG sites with a frequency higher than 50% are indicated in red.
The CpG sites analyzed in Table 4.5 (those with a frequency higher than 50% and an association with MD in the control group in line with its protective effect) are indicated in grey.

Probe Name	Parameter Estimate	Gene	Chromosome	Location	MAPINFO	Delta	Frequency
cg13104385	0.56	IL6	7	Body	22767384	0.020	62.4%
cg18773937	-0.37	*IL1B*	2	Promoter	113594611	0.016	72.1%
cg12195446	0.37	*IRS2*	13	Body	110424497	0.076	64.6%
cg05265849	-0.34	*IL6*	7	Body	22767390	0.010	9.3%
cg02749784	-0.25	*MAL2*	8	Promoter	120219927	0.018	72.4%
cg17968347	-0.23	*SERPINE1*	7	Body	100777740	0.012	55.7%
cg12252547	0.23	*MAL2*	8	Promoter	120220032	0.051	57.3%
cg01265860	-0.23	*RUNX1*	21	Body	36256316	0.016	56.7%
cg24312520	-0.22	*STAT3*	17	Body	40489584	0.017	57.9%
cg16308790	-0.22	*NFATC1*	18	Body	77225973	0.011	51.2%
cg08053846	0.21	*SERPINE1*	7	Promoter	100769605	0.018	64.4%
cg15363134	-0.20	*NFATC1*	18	Body	77161214	0.015	62.7%
cg20674490	-0.17	*RUNX3*	1	Body	25240932	0.019	50.9%
cg27026615	-0.17	*PTX3*	3	Body	157156326	0.021	43.6%
cg08510264	0.16	*IRS2*	13	First Exon	110438288	0.012	32.9%
cg06493806	0.15	*NFATC1*	18	Body	77278806	0.031	44.9%

Table 4.5: Detailed analysis of the selected CpG sites that show an association with MD in the control group in line with its protective effect in CACO1+CACO2.

	Association with colon cancer*					Association with MD (only control group)**			
IL1B	**ca**	**co**	**OR**	**95% CI**	**p-value**	**IMI category**	**coef**	**95% CI**	**p-value**
cg18773937									
model A	157	157	0.69	(0.53,0.89)	0.005	1 (20)	reference		
model B	157	157	0.66	(0.50,0.88)	0.004	2 (71)	0.27	(-0.11,0.64)	0.159
model C	134	134	0.60	(0.43,0.86)	0.005	3 (51)	0.48	(0.07,0.88)	0.021
SERPINE1								p-trend	0.019
cg17968347									
model A	161	161	0.78	(0.62,0.98)	0.030	1 (22)	reference		
model B	161	161	0.72	(0.56,0.93)	0.011	2 (71)	0.04	(-0.09,0.18)	0.530
model C	138	138	0.54	(0.37,0.78)	0.001	3 (52)	0.13	(-0.02,0.28)	0.090
RUNX1								p-trend	0.061
cg01265860									
model A	161	161	0.77	(0.61,0.98)	0.036	1 (22)	reference		
model B	161	161	0.70	(0.52,0.95)	0.021	2 (71)	0.10	(-0.06,0.26)	0.225
model C	138	138	0.53	(0.34,0.81)	0.004	3 (52)	0.12	(-0.05,0.29)	0.176
STAT3								p-trend	0.226
cg24312520									
model A	161	161	0.78	(0.62,0.98)	0.037	1 (22)	reference		
model B	161	161	0.75	(0.58,0.98)	0.034	2 (71)	-0.02	(-0.20,0.15)	0.804
model C	138	138	0.76	(0.54,1.08)	0.121	3 (52)	0.11	(-0.08,0.30)	0.254
NFATC1								p-trend	0.127
cg15363134									
model A	161	161	0.82	(0.67,0.99)	0.043	1 (22)	reference		
model B	161	161	0.79	(0.63,0.98)	0.032	2 (71)	0.24	(0.05,0.42)	0.012
model C	138	138	0.68	(0.44,0.93)	0.017	3 (52)	0.07	(-0.13,0.27)	0.490
RUNX3								p-trend	0.884
cg20674490									
model A	160	160	0.70	(0.55,0.91)	0.008	1 (22)	reference		
model B	160	160	0.68	(0.52,0.90)	0.007	2 (71)	0.17	(-0.08,0.43)	0.190
model C	137	137	0.62	(0.44,0.88)	0.007	3 (51)	0.08	(-0.20,0.36)	0.568
SERPINE1								p-trend	0.839
cg08053846									
model A	134	134	1.28	(1.02,1.60)	0.037	1 (22)	reference		
model B	134	134	1.33	(1.05,1.70)	0.018	2 (71)	-0.13	(-0.46,0.19)	0.424
model C	114	114	1.25	(0.93,1.70)	0.145	3 (51)	-0.21	(-0.56,0.14)	0.232
								p-trend	0.239

*Associations between DNA methylation and colon cancer are assessed using a conditional logistic regression model with DNA methylation levels included as an independent variable and standardized to 1 standard deviation

model A= crude model (adjusted for study center, sex, age and seasonality by design)

model B= model A adjusted for differential cell types

model C= model B adjusted also for fasting status, year of recruitment, BMI, smoking status, physical activity, level of education and IMI

** Associations between the M-values of DNA methylation sites and IMI are assessed using multivariate linear mixed effects models adjusting for various confounding variables (see text for details)

Table 4.6: Validation analysis: associations between colon cancer/MD and DNA methylation levels at *SERPINE1* and *RUNX3* CpG sites obtained by pyrosequencing in CACO3.

	Association with colon cancer*					Association with MD (only control group)**			
SERPINE1	**ca**	**co**	**OR**	**95% CI**	**p-value**	**IMI category**	**coef**	**95% CI**	**p-value**
cg17968347									
model A	47	47	0.82	(0.59,1.15)	0.257	1 (6)	reference		
model B	44	44	0.67	(0.44,1.00)	0.054	2 (23)	0.04	(-0.31,0.38)	0.835
						3 (15)	0.13	(-0.23,0.51)	0.444
RUNX3								p-trend	0.358
cg20674490									
model A	47	47	0.74	(0.47,1.16)	0.196	1 (6)	reference		
model B	44	44	0.55	(0.31,0.99)	0.048	2 (23)	0.44	(-0.09,0.97)	0.102
						3 (15)	0.13	(-0.43,0.97)	0.643
								p-trend	0.830

*Associations between DNA methylation and colon cancer are assessed using a conditional logistic regression model with DNA methylation levels included as an independent variable and standardized to 1 standard deviation

model A= crude model (adjusted for study center, sex, age and seasonality by design)

model B= model A adjusted also for BMI and IMI

** Associations between the M-values of DNA methylation sites and IMI are assessed using multivariate linear models adjusting for age and sex

Table 4.7: Replication analysis: associations between colon cancer/MD and DNA methylation levels at *SERPINE1* and *RUNX3* CpG sites obtained by pyrosequencing in CACO4.

	Association with colon cancer*					Association with MD (only control group)**			
SERPINE1	**ca**	**co**	**OR**	**95% CI**	**p-value**	**IMI category**	**coef**	**95% CI**	**p-value**
cg17968347									
model A	93	93	1.08	(0.82,1.42)	0.573	1 (17)	reference		
model B	87	87	1.12	(0.79,1.56)	0.522	2 (44)	0.01	(-0.22,0.23)	0.961
						3 (27)	0.10	(-0.17,0.36)	0.431
RUNX3								p-trend	0.377
cg20674490									
model A	92	92	0.80	(0.60,1.07)	0.132	1 (17)	reference		
model B	86	86	0.80	(0.57,1.13)	0.210	2 (44)	0.29	(-0.12,0.70)	0.162
						3 (27)	0.37	(-0.09,0.85)	0.114
								p-trend	0.133

*Associations between DNA methylation and colon cancer are assessed using a conditional logistic regression model with DNA methylation levels included as an independent variable and standardized to 1 standard deviation

model A= crude model (adjusted for study center, sex, age, year of recruitment and seasonality by design)

model B= model A adjusted also for BMI, smoking status, physical activity, level of education and IMI

** Associations between the M-values of DNA methylation sites and IMI are assessed using multivariate linear models adjusting for age, sex, center, smoking status, physical activity and BMI

4.6 Figures

Figure 4.1: Diagram illustrating the data sets used in the different stages of the analysis

ca=case

co=control

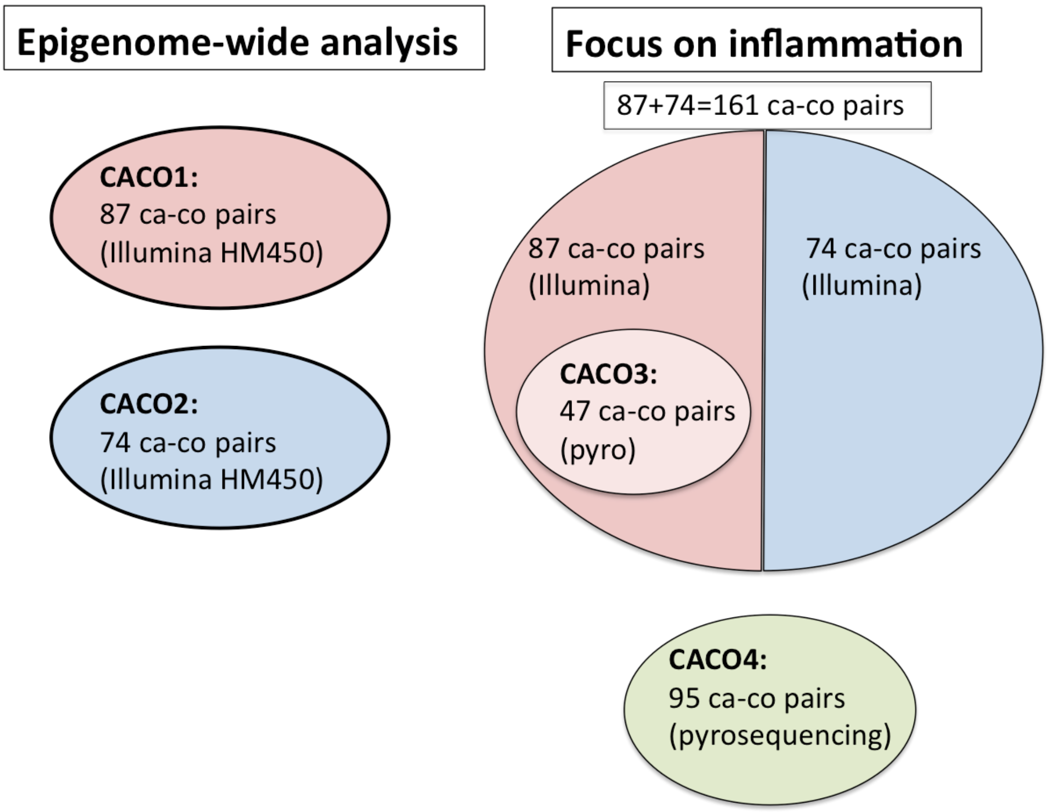


Figure 4.2: Correlations between the 32 CpGs sites with a mean difference in methylation percentage between cases and controls higher than 1%.

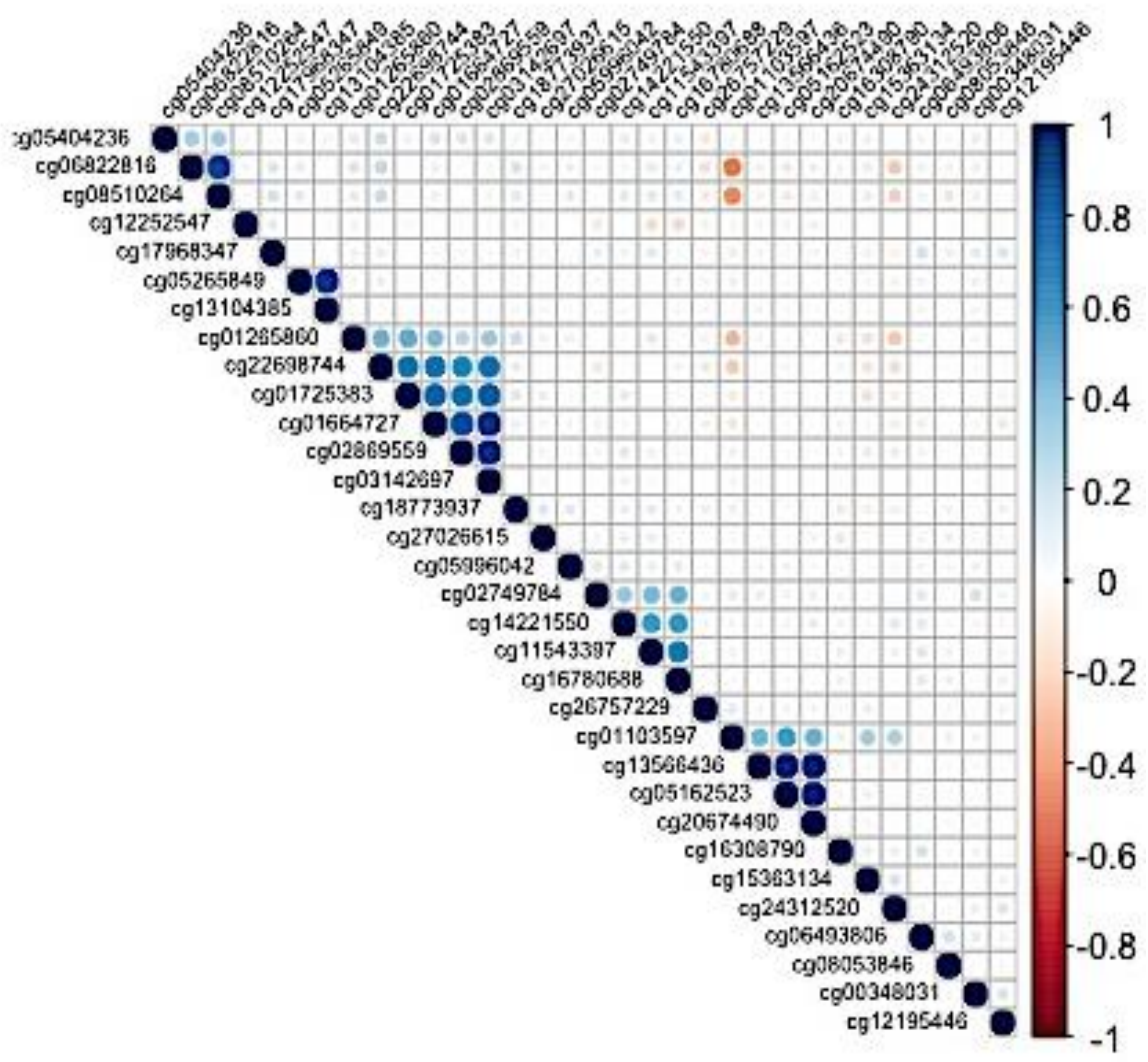


Figure 4.3: Elastic net solution path.

Parameter estimates are plotted against the regularization parameter lambda.
The black line indicates the parameter estimates of the best model (chosen by CV; see Figure 4).

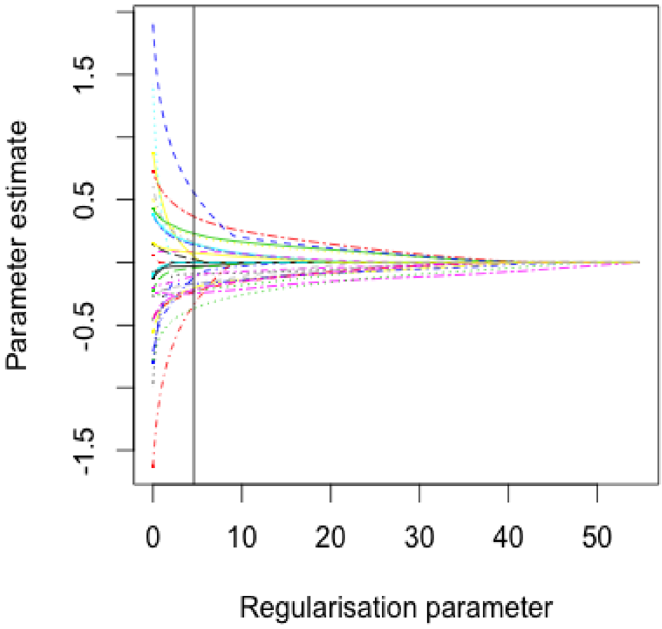
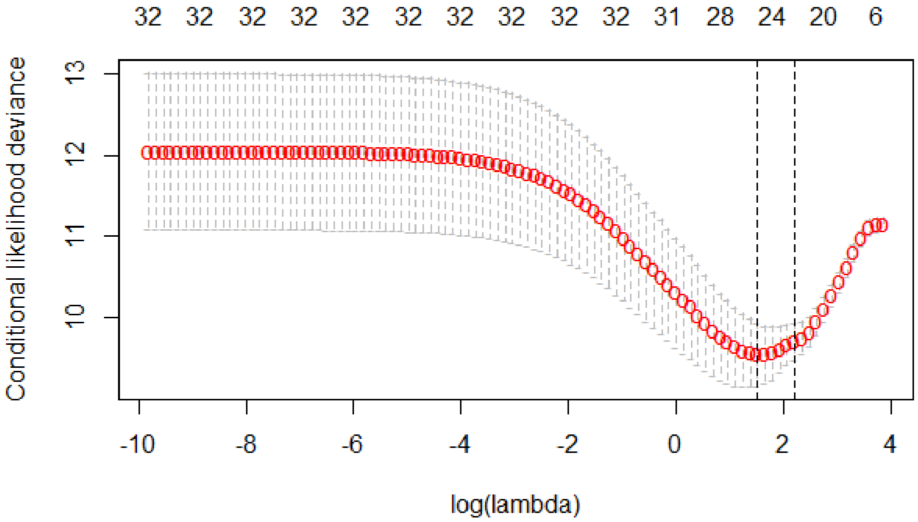


Figure 4.4: Cross Validation deviance curve (with standard error bands).

There are two vertical lines: the leftmost is at the minimizing log(lambda), while the other is drawn at the smallest lambda with CV one standard deviation away from the minimum CV error. CV error seems to be minimized for a model with 26 predictors (log(lambda)=1.51)



4.7 Supplementary material

Supplementary Table 4.1: Descriptive statistics of CACO1+CACO2 (analysis of inflammation-related genes)

Variables	Controls	Cases	all	p-value
N	161	161	322	
Median of (IQR)				
age, years	55 (9)	55 (9)	55 (9)	
BMI, kg/m^2	25.01 (5)	26.42 (5)	25.85 (5)	0.0019
Counts of				
Gender				
men	88 (55%)	88 (55%)	176 (55%)	
women	73 (45%)	73 (45%)	146 (45%)	
Centre				
Varese	64 (40%)	64 (40%)	128 (40%)	
Ragusa	10 (6%)	10 (6%)	20 (6%)	
Turin	87 (54%)	87 (54%)	174 (54%)	
Educational level				0.078
1°tertile RII	51 (34%)	53 (36%)	104 (35%)	
2°tertile RII	54 (36%)	37 (25%)	91 (30%)	
3°tertile RII	45 (30%)	59 (39%)	104 (35%)	
Total physical activity				0.801
inactive	34 (22%)	37 (24%)	71 (23%)	
moderately inactive	65 (42%)	69 (44%)	134 (43%)	
moderately active	36 (23%)	29 (19%)	75 (21%)	
active	20 (13%)	20 (13%)	40 (13%)	
Smoking status				0.828
never smokers	71 (46%)	67 (43%)	138 (44%)	
former smokers	48 (31%)	53 (34%)	101 (33%)	
current smokers	36 (23%)	35 (23%)	71 (23%)	
Fasting status				0.360
yes	119 (74%)	126 (78%)	245 (76%)	
no	42 (26%)	35 (22%)	77 (24%)	
Cell Types				
CD8T	8.1%	7.1%	7.5%	0.129
CD4T	13%	14%	14%	0.528
NK	7.4%	7.5%	7.5%	0.851
Bcell	5.2%	5.4%	5.3%	0.777
Mono	7.4%	7.1%	7.2%	0.543
Gran	63.0%	63.9%	63.2%	0.791

Supplementary Table 4.2: Detailed analysis of the selected CpG sites that show an association with MD in the control group not in line with its protective effect.

	Association with colon cancer*					**Association with MD (only control group)****			
IL6	**ca**	**co**	**OR**	**95% CI**	**p-value**	**IMI category**	**coef**	**95% CI**	**p-value**
cg13104385									
model A	161	161	1.30	(1.03,1.64)	**0.027**	1 (22)	reference		
model B	161	161	1.30	(1.03,1.65)	**0.030**	2 (71)	0.02	(-0.21,0.25)	0.866
model C	138	138	1.32	(0.98,1.78)	0.066	3 (52)	0.10	(-0.15,0.35)	0.434
IRS2								p-trend	0.361
cg12195446									
model A	161	161	1.40	(1.08,1.81)	**0.010**	1 (22)	reference		
model B	161	161	1.41	(1.08,1.83)	**0.010**	2 (71)	0.90	(-0.31,2.11)	0.146
model C	138	138	1.72	(1.22,2.44)	**0.002**	3 (52)	0.47	(-0.83,1.78)	0.478
MAL2								p-trend	0.762
cg02749784									
model A	161	161	0.67	(0.52,0.88)	**0.002**	1 (22)	reference		
model B	161	161	0.68	(0.52,0.89)	**0.005**	2 (71)	-0.12	(-0.33,0.08)	0.249
model C	138	138	0.71	(0.52,0.98)	**0.037**	3 (52)	-0.04	(-0.27,0.19)	0.726
MAL2								p-trend	0.996
cg12252547									
model A	161	161	1.35	(1.06,1.73)	**0.016**	1 (22)	reference		
model B	161	161	1.38	(1.07,1.78)	**0.014**	2 (71)	0.41	(-0.36,1.17)	0.295
model C	138	138	1.55	(1.13,2.13)	**0.007**	3 (52)	0.39	(-0.43,1.21)	0.354
NFATC1								p-trend	0.458
cg16308790									
model A	157	157	0.86	(0.70,1.06)	0.155	1 (22)	reference		
model B	157	157	0.86	(0.70,1.07)	0.188	2 (71)	-0.04	(-0.27,0.19)	0.726
model C	138	138	0.96	(0.74,1.26)	0.795	3 (52)	-0.08	(-0.33,0.16)	0.503

*Associations between DNA methylation and colon cancer are assessed using a conditional logistic regression model with DNA methylation levels included as an independent variable and standardized to 1 standard deviation

model A= crude model (adjusted for study center, sex, age and seasonality by design)

model B= model A adjusted for differential cell types

model C= model B adjusted also for fasting status, year of recruitment, BMI, smoking status, physical activity, level of education and IMI

** Associations between the M-values of DNA methylation sites and IMI are assessed using multivariate linear mixed effects models adjusting for various confounding variables (see text for details)

Supplementary Table 4.3: Descriptive statistics of CACO3

Variables	Controls	Cases	all	p-value
N	47	47	94	
Median of (IQR)				
age, years*	56 (10)	55 (8)	56 (8)	
BMI, kg/m^2	25 (5)	26.41 (3)	25.90 (4)	0.0167
Counts of				
*Gender**				
men	35 (74%)	35 (74%)	70 (74%)	
women	12 (26%)	12 (26%)	24 (26%)	
*Centre**				
Turin	47 (100%)	47 (100%)	94 (100%)	
Educational level				0.034
1°tertile RII	12 (28%)	14 (31%)	26 (30%)	
2°tertile RII	18 (42%)	8 (18%)	26 (30%)	
3°tertile RII	13 (30%)	23 (51%)	36 (40%)	
Total physical activity				0.908
inactive	8 (18%)	9 (19%)	17 (19%)	
moderately inactive	20 (45%)	20 (43%)	40 (44%)	
moderately active	9 (20%)	8 (17%)	17 (19%)	
active	7 (16%)	10 (21%)	17 (19%)	
Smoking status				0.768
never smokers	10 (23%)	13 (28%)	23 (25%)	
former smokers	16 (36%)	18 (38%)	34 (37%)	
current smokers	18 (41%)	16 (34%)	34(37%)	
Fasting status				0.533
yes	25 (53%)	28 (60%)	53 (56%)	
no	22 (47%)	19 (40%)	41 (44%)	

Supplementary Table 4.4: Descriptive statistics of CACO4

Variables	Controls	Cases	all	p-value
N	95	95	190	
Median of (IQR)				
age, years*	57 (10)	57 (10)	57 (10)	
BMI, kg/m^2	26.11 (5)	26.99 (5)	26.70 (6)	0.2152
Counts of				
*Gender**				
men	41 (43%)	41 (43%)	46 (43%)	
women	54 (57%)	54 (57%)	108 (57%)	
*Centre**				
Varese	29 (30%)	29 (30%)	58 (30%)	
Ragusa	13 (14%)	13 (14%)	26 (14%)	
Torino	46 (49%)	46 (49%)	92 (49%)	
Napoli	7 (7%)	7 (7%)	14 (7%)	
Educational level				0.320
1°tertile RII	35 (39%)	38 (40%)	73 (39%)	
2°tertile RII	27 (30%)	20 (21%)	47 (25%)	
3°tertile RII	28 (31%)	37 (39%)	65 (35%)	
Total physical activity				0.483
inactive	26 (28%)	28 (29%)	54 (29%)	
moderately inactive	36 (38%)	44 (46%)	80 (42%)	
moderately active	18 (19%)	14 (15%)	32 (17%)	
active	14 (15%)	9 (9%)	23 (12%)	
Smoking status				0.178
never smokers	25 (27%)	19 (20%)	44 (23%)	
former smokers	19 (20%)	30 (32%)	49 (26%)	
current smokers	50 (53%)	46 (48%)	96 (51%)	

Supplementary Information:

Laboratory methods: Pyrosequencing

Pyrosequencing assay was performed for CACO3 and CACO4 samples on a PyroMark Q24 MDx system using PyroMark Gold Q24 Advanced reagents (Qiagen, Hilden Germany). Primers were designed according to PyroMark Assay Design software version 2.0 (Qiagen). PCR reaction was performed in a total volume of 35 µl using the PyroMark PCR kit (Qiagen) containing 1X PCR Master Mix, 1X CoralLoad Concentrate, , 0.2 µM of each primer, and 1 µl of bisulfite-converted DNA with the following cycling profile: 95°C for 10 min followed by 45 cycles of denaturation at 95°C for 30 sec, annealing at specific temperature for each gene (55°C for *RUNX3*; 50°C for *SERPINE1*) for 30 sec, extension at 72°C for 1 min. Extension at 72°C for 10 min was finally performed. The PCR product (15 µl) was added to 19 µl of distilled water and incubated under shaking with 40 µl of binding buffer pH 7.6, containing 10mM Tris-HCl, 2 M NaCl, 1mM EDTA, and 1 µl of sepharose beads covered by streptavidin. The PCR product was washed with ethanol 70%, denatured with NaOH 0.2 M and re-washed with Tris-Acetate 10 mM pH 7.6. Pyrosequencing reaction was performed in a total volume of 20 µl, including 19.85 µl of 20 mM Tris-Acetate, 5 mM MgAc2 and 0.15 µl of 50 µM sequencing primer. Assays were created according to manufacturer's instruction. The nucleotide dispensation order was suggested by the software PyroMark Q24 Advanced version 3.0.0.

Methylation quantification was achieved using the provided software, and expressed for each DNA locus as percentage of methylated cytosines divided by the sum of methylated and unmethylated cytosines. Positive controls for methylated [EpiTect Control DNA (human), methylated (Qiagen)] and unmethylated status [EpiTect Control DNA (human), unmethylated (Qiagen)] were included in each pyrosequencing run. Each sample was analyzed twice in different runs and the average of the two results was computed. Adequacy of the results for each sample was achieved when difference in methylation percentage between runs was ≤2% and pyrograms resulted as "passed".

References

1. Trichopoulou A, Costacou T, Bamia C, Trichopoulos D. Adherence to a Mediterranean diet and survival in a Greek population. N Engl J Med. 2003 Jun 26;348(26):2599-608. PMID: 12826634
2. Sofi F, Cesari F, Abbate R, Gensini GF, Casini A. Adherence to Mediterranean diet and health status: meta-analysis. BMJ. 2008 Sep 11;337:a1344. PMID: 18786971.
3. Agnoli C, Grioni S, Sieri S, Palli D, Masala G, Sacerdote C, Vineis P, Tumino R, Giurdanella MC, Pala V, Berrino F, Mattiello A, Panico S, Krogh V. Italian Mediterranean Index and risk of colorectal cancer in the Italian section of the EPIC cohort. Int J Cancer. 2013 Mar 15;132(6):1404-11. PMID: 22821300.
4. Rosato V et al. Mediterranean diet and colorectal cancer risk: a pooled analysis of three Italian case-control studies. Br J Cancer. 2016 Sep 27;115(7):862-5. doi: 10.1038/bjc.2016.245. Epub 2016 Aug 18.
5. Torres Stone RA, Waring ME, Cutrona SL, et al. The association of dietary quality with colorectal cancer among normal weight, overweight and obese men and women: a prospective longitudinal study in the USA. BMJ Open 2017;7:e015619. doi:10.1136/ bmjopen-2016-015619
6. Park SY, Boushey CJ, Wilkens LR, Haiman CA, Le Marchand L. High-Quality Diets Associate With Reduced Risk of Colorectal Cancer: Analyses of Diet Quality Indexes in the Multiethnic Cohort. Gastroenterology. 2017 Aug;153(2):386-394.e2. doi: 10.1053/j.gastro.2017.04.004. Epub 2017 Apr 17.
7. Fasanelli F, Zugna D, Giraudo MT, Krogh V, Grioni S, Panico S, Mattiello A, Masala G, Caini S, Tumino R, Frasca G, Sciannameo V, Ricceri F, Sacerdote C. Abdominal adiposity is not a mediator of the protective effect of Mediterranean Diet on colorectal cancer. Int J Cancer. 2017 May 15;140(10):2265-2271. doi: 10.1002/ijc.30653. Epub 2017 Mar 2.
8. Herceg Z. Epigenetics and cancer: towards an evaluation of the impact of environmental and dietary factors. Mutagenesis. 2007 Mar;22(2):91-103. PMID: 17284773.
9. Casas R, Sacanella E, Urpí-Sardà M, Corella D, Castañer O, Lamuela-Raventos RM, Salas-Salvadó J, Martínez-González MA, Ros E, Estruch R. Long-Term Immunomodulatory Effects of a Mediterranean Diet in Adults at High Risk of Cardiovascular Disease in the PREvención con DIeta MEDiterránea (PREDIMED) Randomized Controlled Trial. J Nutr. 2016 Sep;146(9):1684-9
10. Koloverou E, Panagiotakos DB, Pitsavos C, Chrysohoou C, Georgousopoulou EN, Grekas A, Christou A, Chatzigeorgiou M, Skoumas I, Tousoulis D, Stefanadis C; ATTICA Study Group. Adherence to Mediterranean diet and 10-year incidence (2002-2012) of diabetes: correlations with inflammatory and oxidative stress biomarkers in the ATTICA cohort study. Diabetes Metab Res Rev. 2016 Jan;32(1):73-81
11. Chiba T, Marusawa H, Ushijima T. Inflammation-associated cancer development in digestive organs: mechanisms and roles for genetic and epigenetic modulation. Gastroenterology. 2012 Sep;143(3):550-63.
12. Ligthart S, Marzi C, Aslibekyan S et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biol. 2016 Dec 12;17(1):255.
13. Cannon J. Colorectal Neoplasia and Inflammatory Bowel Disease. Surg Clin North Am. 2015 Dec;95(6):1261-9
14. Santilli F, Boccatonda A, Davì G. Aspirin, platelets, and cancer: The point of view of the internist. Eur J Intern Med. 2016 Oct;34:11-20
15. Palli D at al. A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. Tumori. 2003 Nov-Dec;89(6):586-93

16. Agnoli C, Krogh V, Grioni S, et al. A priori-defined dietary patterns are associated with reduced risk of stroke in a large Italian cohort. J Nutr 2011;141:1552-8.
17. Wareham NJ, Jakes RW, Rennie KL et al. Validity and repeatability of a simple index derived from the short physical activity questionnaires used in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Public Health Nutr 2003;6:407-13
18. Mackenbach JP, Kunst AE. Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from Europe. Doc Sci Med 1997;44:757-71.
19. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13, 86 (2012)
20. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
21. Reid S, Tibshirani R. Regularization Paths for Conditional Logistic Regression: The clogitL1 Package. J Stat Softw. 2014 Jul;58(12). pii: 12
22. Zou H and Hastie T. Regularization and variable selection via the elastic net. J. R. Statist. Soc. B (2005) 67, Part 2, pp. 301–320
23. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288.
24. Hoerl, A. and Kennard, R. (1988) Ridge regression. In Encyclopedia of Statistical Sciences, vol. 8, pp. 129–136. New York: Wiley
25. Freidman HJ, Hastie T and Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 10.18637/jss.v033.i01
26. Hollington P, Neufing P, Kalionis B, Waring P, Bentel J, Wattchow D, Tilley WD. Expression and localization of homeodomain proteins DLX4, HB9 and HB24 in malignant and benign human colorectal tissues. Anticancer Res. 2004 Mar-Apr;24(2B):955-62.
27. Aparicio LA, Valladares M, Blanco M, Alonso G, Figueroa A. Biological influence of Hakai in cancer: a 10-year review. Cancer Metastasis Rev. 2012 Jun;31(1-2):375-86. doi: 10.1007/s10555-012-9348-x.
28. Lotem J, Levanon D, Negreanu V, Bauer O, Hantisteanu S, Dicken J, Groner Y. Runx3 at the interface of immunity, inflammation and cancer. Biochim Biophys Acta. 2015 Apr;1855(2):131-43.
29. Goel A, Arnold CN, Tassone P, Chang DK, Niedzwiecki D, Dowell JM, Wasserman L, Compton C, Mayer RJ, Bertagnolli M, Boland CR. Epigenetic inactivation of RUNX3 in microsatellite unstable sporadic colon cancers. Int J Cancer. 2004;112:754–759
30. Michels KB et al. Recommendations for the design and analysis of epigenome-wide association studies. Nature Methods 10, 949-955 (2013) doi:10.1038/nmeth.2632
31. Roessler J, Ammerpohl O, Gutwein J, Hasemeier B, Anwar SL, Kreipe H, Lehmann U. Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc. BMC Res Notes. 2012 Apr 30;5:210. doi: 10.1186/1756-0500-5-210.
32. BLUEPRINT consortium. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol. 2016 Jul;34(7):726-37. doi: 10.1038/nbt.3605. Epub 2016 Jun 27.
33. Milenkovic D, Vanden Berghe W, Boby C, Leroux C, Declerck K, Szarc vel Szic K, Heyninck K, Laukens K, Bizet M, Defrance M, Dedeurwaerder S, Calonne E, Fuks F, Haegeman G, Haenen GR, Bast A, Weseler AR. Dietary flavanols modulate the transcription of genes associated with cardiovascular pathology without changes in their DNA methylation state. PLoS One. 2014 Apr 24;9(4):e95527. doi: 10.1371/journal.pone.0095527. eCollection 2014.
34. Tung J, Gilad Y. Social environmental effects on gene regulation. Cell Mol Life Sci 2013;70:4323–39.]

35. Heijmans BT, Mill J. Commentary: The seven plagues of epigenetic epidemiology. Int J Epidemiol 2012;41:74–78.
36. Flanagan JM, Brook MN, Orr N, Tomczyk K, Coulson P, Fletcher O, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow A, Brown R, Garcia-Closas M. Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. Cancer Epidemiol Biomarkers Prev. 2015 Jan;24(1):221-9. doi: 10.1158/1055-9965.EPI-14-0767. Epub 2014 Nov 4.
37. Adalsteinsson BT, Gudnason H, Aspelund T et al. Heterogeneity in white blood cells has potential to confound DNA methylation measurements. PLoS One 2012;7:e46705.

5. Conclusions and future perspectives

The main focus of this thesis was to analyze the role of DNA methylation as a potential mediator of the carcinogenic process triggered by specific environmental exposures.

In details we studied the etiology of lung cancer and colon cancer focusing on smoking and Mediterranean Diet exposures respectively. These environmental exposures can influence the biology of the tumors through epigenetic alterations. Aberrant DNA methylation is one of the most important epigenetic modifications involved in early stages of tumorigenesis. Principally methylation represents an adaptive response to external stimuli leading to modulation of gene expression in a temporary or permanent way and to alteration of the functionality of proteins that are part of methylation machinery. Aberrant methylation induced by long lasting environmental exposures may persist for a long time, providing further support of a possible causal involvement of DNA methylation in carcinogenesis. Persistence of altered methylation was found in blood cells up to 17 years after smoking cessation. Similarly, stable epigenetic marks of nutritional factors at key life stages may persist over decades.

The link between DNA methylation and cancer is well documented in literature. It is commonly known that inactivation of certain tumor-suppressor genes occurs as a consequence of hypermethylation within the promoter regions. A series of studies have shown that a broad range of genes are silenced by DNA methylation in different cancer types. On the other hand, global hypomethylation, inducing genomic instability, also contributes to cell transformation. Apart from DNA methylation alterations in promoter regions and repetitive DNA sequences, this phenomenon is associated also with the regulation of expression of noncoding RNAs such as microRNAs that may play a role in tumor suppression. For all these reasons, DNA methylation is a good mediator candidate that might explain the link between smoking and lung cancer and between Mediterranean Diet and colon cancer.

The study of DNA methylation as a molecular intermediate is fundamental for a lot of reasons. First of all, the finding of a biological mediator strengthen causal links between the exposure and the disease providing the opportunity to understand the flow of information that underlies disease. Second, knowledge of biological mediators may be useful for the planning of prevention strategies since a biological intermediate is often a potentially modifiable risk factor lying between an exposure and an outcome which, when intervened upon, will block the causal pathway between the exposure and the outcome. In fact, unlike genetic alterations, DNA methylation is reversible what makes it extremely interesting for therapy approaches.

Lastly the molecular intermediates may be used to improve the prediction of disease risk as seen in the case of the CpG sites found to be related to both smoking and lung cancer.

In the future we would like to integrate the information about methylation as mediator by studying also the other related mechanisms such as gene expression levels, metabolite concentrations and proteomics. All these mechanisms interact with each other having an intrinsic hierarchical structure that should be known. The comprehensive understanding of this structure will simultaneously require the study, use and development of new analytical methodologies. In fact the increasing availability of molecular data represents both an opportunity for advancing knowledge in clinical field and public health, and a methodological challenge for data analysis. The main future perspective is to work in this methodological framework by focusing on big data analysis techniques, in particular on pattern identification, development of predictive models and causal inference methods for estimating cause-effect relationships.

To conclude the results of the present thesis are important in the elucidation of the pathways involved in lung cancer and colon cancer pathogenesis, and their potential clinical implications, i.e., using methylation in personalized prevention medicine. Nevertheless, the interpretation of these findings is still uncertain since independent functional studies are needed to investigate the effect of methylation on gene expression. The paper reported in the Appendix is a first work that is part of the methodological project that we would like to follow in our future research.

PhD final report

Attività generale di ricerca

Ambito di ricerca interdisciplinare, che abbina le competenze matematiche e statistiche con quelle genetiche, mediche ed epidemiologiche per lo studio di un ampio spettro di problemi che implichino l'analisi avanzata di dati biomedici. In particolare: analisi del ruolo della metilazione del DNA come potenziale mediatore del processo di carcinogenesi innescato da specifiche esposizioni ambientali.

Attività di formazione (PhD program)

- partecipazione ai seminari organizzati dal dottorato; in particolare esposizione orale dei seguenti seminari satellite:
  - 9 Ottobre 2017: "1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function";
  - 14 Febbraio 2017: "Neurocognition across the spectrum of mucopolysaccharidosis type I: Age, severity, and treatment";
  - 16 Dicembre 2016: "A microRNA biomarker of hepatocellular carcinoma recurrence following liver transplantation accounting for within-patient heterogeneity";
  - 17 Maggio 2016: "Accounting for Population Stratification in DNA Methylation Studies*";*
  - 21 Dicembre 2015 "*FOXP2* gene and language impairment in schizophrenia: association and epigenetic studies";
  - 19 Maggio 2015: "A rare functional cardioprotective *APOC3* variant has risen in frequency in distinct population isolates"
  - 9 Dicembre 2014: "Gestational diabetes mellitus epigenetically affects genes predominantly involved in metabolic diseases"
  - 10 Giugno 2014: "Successful identification of rare variants using oligogenic segregation analysis as a prioritizing tool for whole-exome sequencing studies"
- partecipazione al "D-day 2017", Centro di Biotecnologie Molecolari M.B.C., Torino 22 settembre 2017 (esposizione e presentazione del poster: "Mediation analysis in Molecular Epidemiology.")
- partecipazione alla "Giornata dedicata alla valorizzazione delle competenze dei dottori di ricerca e al postdoc, Workshop 3, Introduzione al Fundraising", Torino 4 ottobre 2016
- partecipazione al "D-day 2016", Centro di Biotecnologie Molecolari M.B.C., Torino 15 settembre 2016
- Corso d'Inglese Scientifico livello B1/B2, Torino, MBC, novembre 2015 – marzo 2016 (votazione finale 29/30)
- partecipazione al seminario promosso dalla Common Strategic Task Force/CSTF di Ateneo "Essere giovani protagonisti in H2020: il CV e opportunità di finanziamento attraverso la mobilità internazionale", Torino, 2 luglio 2014

Altre attività di formazione

- Partecipazione come uditrice al "Master Biennale Universitario di II livello in Epidemiologia (2015-2016)" (10 moduli: 1) Principi di epidemiologia 2) Metodi Statistici I 3) Metodi Statistici II 4) Design, conduction and analysis of cohort studies 5) Disegno, conduzione ed analisi di studi caso-controllo 6) Modelli di regressione in epidemiologia 7)

Disegno ed analisi di studi clinici e di intervento Revisioni sistematiche e metanalisi 8) Statistical methods for survival analysis 9) Principi dello screening / Interpretazione epidemiologica degli studi e comunicazione del rischio 10) Metodi avanzati in statistica ed epidemiologia)

- Corso "Analisi dati next-generation sequencing (NGS)", Forlì, 28 settembre - 2 ottobre 2015
- Corso pre-congressuale della Società Italiana di Statistica Medica ed Epidemiologia Clinica (SISMEC) "Mediation Analysis in Epidemiology", Torino, 16 settembre 2015
- Corso "Statistical approaches to characterize the exposome from OMICS platforms", Imperial College, Londra, 08-12 dicembre 2014
- Corso di aggiornamento in "Evoluzione delle funzionalità grafiche in SAS System 9.2", Grugliasco, 21 febbraio 2014

Partecipazione a Convegni, Workshop, Meeting, Seminari

- XLI convegno dell'Associazione Italiana di Epidemiologia, Mantova, 25-27 ottobre 2017 (presentazione orale dal titolo: "Analisi di mediazione multipla per l'associazione tra depressione materna e sibili e fischi al torace nei primi 18 mesi di vita del bambino")
- Convegno di Primavera Associazione Italiana di Epidemiologia (AIE) 2017, Roma, 5-6 giugno 2017 (docente del corso precongressuale "Introduzione all'epidemiologia molecolare"; presentazione orale dal titolo: "L'analisi delle componenti principali applicata ad uno studio epigenome-wide innestato nella coorte EPIC Italia")
- XIX Congresso Nazionale SIGU, Torino, 23-26 novembre 2016
- XL riunione annuale dell'Associazione Italiana di Epidemiologia, Torino, 19-21 ottobre 2016 (presentazione orale dal titolo: "Weighting approach per mediatori multipli nell'analisi di sopravvivenza.", vincitrice del III posto del Premio Maccacaro 2016)
- XXXIX riunione annuale dell'Associazione Italiana di Epidemiologia, Milano, 28-30 ottobre 2015 (presentazione orale dal titolo: "Dieta mediterranea e rischio di cancro del colon-retto: un'analisi di mediazione.")
- EPIC meeting "Statistical Methods in Nutritional Epidemiology", IARC, Lione, 24-25 settembre 2015 (presentazione orale dal titolo: "Mediation analysis: application to the study of the relationship between diet and colorectal cancer in EPIC Italy")
- VIII Congresso Nazionale SISMEC 2015, Torino, 17-19 settembre 2015
- XXXVIII riunione annuale dell'Associazione Italiana di Epidemiologia, Napoli, 5-7 novembre 2014
- Workshop "The nine months that change your life", Torino, 28 ottobre 2014
- "Methylation analyses in EPIC Meeting", Torino, 16 ottobre 2014
- "Exposomic Meeting", Imperial College, Londra, 25 settembre 2014
- Riunione del gruppo EPIC Italy, Torino, 13 giugno 2014

Pubblicazioni

Ricceri F, Giraudo MT, Fasanelli F, Milanese D, Sciannameo V, Fiorini L, Sacerdote C. **Diet and endometrial cancer: a focus on the role of fruit and vegetable intake, Mediterranean diet and dietary inflammatory index in the endometrial cancer risk.** BMC Cancer. 2017 Nov 13;17(1):757. doi: 10.1186/s12885-017-3754-y.

Popovic M, Fasanelli F, Fiano V, Biggeri A, Richiardi L. **Increased correlation between methylation sites in epigenome-wide replication studies: impact on analysis and results.** Epigenomics. 2017 Nov 6. doi: 10.2217/epi-2017-0073. [Epub ahead of print]

Sieri S, Agnoli C, Pala V, Grioni S, Brighenti F, Pellegrini N, Masala G, Palli D, Mattiello A, Panico S, Ricceri F, Fasanelli F, Frasca G, Tumino R, Krogh V. **Dietary glycemic index, glycemic load, and cancer risk: results from the EPIC-Italy study.** Sci Rep. 2017 Aug 29;7(1):9757. doi: 10.1038/s41598-017-09498-2. PMID: 28851931 Free PMC Article

Trajkova S, d'Errico A, Ricceri F, Fasanelli F, Pala V, Agnoli C, Tumino R, Frasca G, Masala G, Saieva C, Chiodini P, Mattiello A, Sacerdote C, Panico S. **Impact of preventable risk factors on stroke in the EPICOR study: does gender matter?** Int J Public Health. 2017 Jun 22. doi: 10.1007/s00038-017-0993-2. [Epub ahead of print] PMID: 28643029

Jakszyn P, Fonseca-Nunes A, Lujan-Barroso L, Aranda N, Tous M, Arija V, Cross A, Bueno de Mesquita B, Weiderpass E, Kühn T, Kaaks R, Sjöberg K, Ohlsson B, Tumino R, Palli D, Ricceri F, Fasanelli F, Krogh V, Mattiello A, Jenab M, Gunter M, Perez-Cornago A, Khaw KT, Tjønneland A, Olsen A, Overvad K, Trichopoulou A, Peppa E, Vasilopoulou E, Boeing H, Sánchez-Cantalejo E, Huerta JM, Dorronsoro M, Barricarte A, Quirós JM, Peeters PH, Agudo A. **Hepcidin levels and gastric cancer risk in the EPIC-EurGast Study.** Int J Cancer. 2017 Sep 1;141(5):945-951. doi: 10.1002/ijc.30797. Epub 2017 Jun 21. PMID: 28543377

Hüsing A, Fortner RT, Kühn T, Overvad K, Tjønneland A, Olsen A, Boutron-Ruault MC, Severi G, Fournier A, Boeing H, Trichopoulou A, Benetou V, Orfanos P, Masala G, Pala V, Tumino R, Fasanelli F, Panico S, Bueno de Mesquita HB, Peeters PH, van Gills CH, Quirós JR, Agudo A, Sánchez MJ, Chirlaque MD, Barricarte A, Amiano P, Khaw KT, Travis RC, Dossus L, Li K, Ferrari P, Merritt MA, Tzoulaki I, Riboli E, Kaaks R. **Added Value of Serum Hormone Measurements in Risk Prediction Models for Breast Cancer for Women Not Using Exogenous Hormones: Results from the EPIC Cohort.** Clin Cancer Res. 2017 Aug 1;23(15):4181-4189. doi: 10.1158/1078-0432.CCR-16-3011. Epub 2017 Feb 28. PMID: 28246273

Fasanelli F, Zugna D, Giraudo MT, Krogh V, Grioni S, Panico S, Mattiello A, Masala G, Caini S, Tumino R, Frasca G, Sciannameo V, Ricceri F, Sacerdote C. **Abdominal adiposity is not a mediator of the protective effect of Mediterranean Diet on colorectal cancer.** Int J Cancer. 2017 May 15;140(10):2265-2271. doi: 10.1002/ijc.30653. Epub 2017 Mar 2.

Zamora-Ros R, Barupal DK, Rothwell JA, Jenab M, Fedirko V, Romieu I, Aleksandrova K, Overvad K, Kyrø C, Tjønneland A, Affret A, His M, Boutron-Ruault MC, Katzke V, Kühn T, Boeing H, Trichopoulou A, Naska A, Kritikou M, Saieva C, Agnoli C, Santucci de Magistris M, Tumino R, Fasanelli F, Weiderpass E, Skeie G, Merino S, Jakszyn P, Sánchez MJ, Dorronsoro M, Navarro C, Ardanaz E, Sonestedt E, Ericson U, Maria Nilsson L, Bodén S, Bueno-de-Mesquita HB, Peeters PH, Perez-Cornago A, Wareham NJ, Khaw KT, Freisling H, Cross AJ, Riboli E, Scalbert A. **Dietary flavonoid intake and colorectal cancer risk in the European prospective investigation into cancer and nutrition (EPIC) cohort.** Int J Cancer. 2017 Apr 15;140(8):1836-1844. doi: 10.1002/ijc.30582. Epub 2017 Jan 19.

Ricceri F, Sacerdote C, Giraudo MT, Fasanelli F, Lenzo G, Galli M, Sieri S, Pala V, Masala G, Bendinelli B, Tumino R, Frasca G, Chiodini P, Mattiello A, Panico S. **The Association between Educational Level and Cardiovascular and Cerebrovascular Diseases within the EPICOR Study: New Evidence for an Old Inequality Problem.** PLoS One. 2016 Oct 6;11(10):e0164130. doi: 10.1371/journal.pone.0164130.

Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, Chung J, Fasanelli F, Guida F, Campanella G, Chadeau-Hyam M, Grankvist K, Johansson M, Ala U, Provero P, Wong EM, Joo J, English DR, Kazmi N, Lund E, Faltus C, Kaaks R, Risch A, Barrdahl M, Sandanger

TM, Southey MC, Giles GG, Johansson M, Vineis P, Polidoro S, Relton CL, Severi G. **DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk.** Int J Cancer. 2016 Sep 15. doi: 10.1002/ijc.30431.

Critelli R, Fasanelli F, Oderda M, Polidoro S, Assumma MB, Viberti C, Preto M, Gontero P, Cucchiarale G, Lurkin I, Zwarthoff EC, Vineis P, Sacerdote C, Matullo G, Naccarati A. **Detection of multiple mutations in urinary exfoliated cells from male bladder cancer patients at diagnosis and during follow-up.** Oncotarget. 2016 Sep 7. doi: 10.18632/oncotarget.11883.

Lassale C, Gunter MJ, Romaguera D, Peelen LM, Van der Schouw YT, Beulens JW, Freisling H, Muller DC, Ferrari P, Huybrechts I, Fagherazzi G, Boutron-Ruault MC, Affret A, Overvad K, Dahm CC, Olsen A, Roswall N, Tsilidis KK, Katzke VA, Kühn T, Buijsse B, Quirós JR, Sánchez-Cantalejo E, Etxezarreta N, Huerta JM, Barricarte A, Bonet C, Khaw KT, Key TJ, Trichopoulou A, Bamia C, Lagiou P, Palli D, Agnoli C, Tumino R, Fasanelli F, Panico S, Bueno-de-Mesquita HB, Boer JM, Sonestedt E, Nilsson LM, Renström F, Weiderpass E, Skeie G, Lund E, Moons KG, Riboli E, Tzoulaki I. **Diet Quality Scores and Prediction of All-Cause, Cardiovascular and Cancer Mortality in a Pan-European Cohort Study.** PLoS One. 2016 Jul 13;11(7):e0159025. doi: 10.1371/journal.pone.0159025.

Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma M, Naccarati A, Chadeau-Hyam M, De Stavola B, Hodge A, Giles GG, Southey MC, Relton CL, Haycock PC, Lund E, Polidoro S, Sandanger TM, Severi G, Vineis P. **Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts.** Nat Commun. 2015 Dec 15;6:10192. doi: 10.1038/ncomms10192.

Saieva C, Caini S, Ceroti M, Fasanelli F, Ricceri F, Agnoli C, Grioni S, Mattiello A, Santucci de Magistris M, Tumino R, Martorana C, Masala G. Alcohol consumption and epithelial cancer risk in the EPIC-Italy cohort. Epidemiol Prev. 2015 Sep-Dec;39(5-6):345-9.

Mattiello A, Chiodini P, Santucci de Magistris M, Krogh V, Grioni S, Fasanelli F, Vineis P, Saieva C, Bendinelli B, Frasca G, Giurdanella MC, Panico S. **Dietary habits and cardiovascular disease: the experience of EPIC Italian collaboration.** Epidemiol Prev. 2015 Sep-Dec;39(5-6):339-44. Italian.

Ricceri F, Fasanelli F, Giraudo MT, Sieri S, Tumino R, Mattiello A, Vagliano L, Masala G, Quirós JR, Travier N, Sánchez MJ, Larranaga N, Chirlaque MD,Ardanaz E, Tjonneland A, Olsen A, Overvad K, Chang-Claude J, Kaaks R, Boeing H, Clavel-Chapelon F, Kvaskoff M, Dossus L, Trichopoulou A, Benetou V,Adarakis G, Bueno-de-Mesquita HB, Peeters PH, Sund M, Andersson A, Borgquist S, Butt S, Weiderpass E, Skeie G, Khaw KT, Travis RC, Rinaldi S, Romieu I,Gunter M, Kadi M, Riboli E, Vineis P, Sacerdote C. **Risk of second primary malignancies in women with breast cancer: Results from the European prospective investigation into cancer and nutrition (EPIC).** Int J Cancer. 2015 Feb 3. doi: 10.1002/ijc.29462

Ricceri F, Trevisan M, Fiano V, Grasso C, Fasanelli F, et al. (2014) **Seasonality Modifies Methylation Profiles in Healthy People**. PLoS ONE 9(9): e106846. doi:10.1371/journal.pone.0106846

Poster

Fasanelli F. **Mediation analysis in molecular epidemiology.** poster esposto al D-Day della Scuola di Dottorato in Scienze della Vita e della Salute (Torino, 19 settembre 2017)

Giraudo MT, Fasanelli F, Ricceri F, Sacerdote C, Zugna D. **Weighting approach for multiple mediators in survival analysis. -** poster esposto al I First Italian Meeting on Probability and Mathematical Statistics (Torino, 19-22 Giugno 2017)

Trevisan M, Fiano V, Grasso C, De Marco L, Sacerdote C, Fasanelli F, Gillio Tos A. **Stato di metilazione in geni umani selezionati come marcatore di aggressività in lesioni pre-neoplastiche della cervice uterina. -** poster esposto alla XL riunione annuale dell'Associazione Italiana di Epidemiologia (Torino, 19-21 ottobre 2016)

Fasanelli F, Ricceri F, Zugna D, Giraudo MT, Krogh V, Grioni S, Mattiello A, Panico S, Masala G, Caini S, Tumino R, Frasca G, Vineis P, Sacerdote C. **Mediterranean Diet and Colorectal Cancer: a mediation analysis in the EPIC Italy cohort.** - poster esposto alla IARC 50th Anniversary Conference: "Global Cancer, Occurrence, Causes and Avenues to Prevention" (Lyon, 8-10 June 2016)

Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma M, Naccarati A, Chadeau-Hyam M, De Stavola B, Hodge A, Giles GG, Southey MC, Relton CL, Haycock PC, Lund E, Polidoro S, Sandanger TM, Severi G, Vineis P. **Hypomethylation of smoking-related genes is associated with future lung cancer in three prospective cohorts.** - poster esposto all'VIII Congresso Nazionale SISMEC 2015 (Torino)

Critelli R, Fasanelli F, Assumma MB, Zwarthoff EC, Oderda M, Polidoro S, Sacerdote C, Vineis P, Matullo G, Naccarati A. **Mutation detection in urine from bladder cancer patients as non-invasive prognostic tool.** - poster esposto all'American Association for Cancer Research (AACR) Annual Meeting 2015 (Philadelphia)

Fasanelli F, Ricceri F, Giraudo MT, Troncoso Baltar V, Grioni S, Panico S, Masala G, Tumino R, EPIC-InterAct collaborators, Vineis P, Sacerdote C. **Analisi del rapporto tra status socio-economico e diabete utilizzando i modelli di equazioni strutturali: lo studio EPIC-Interact.** - poster esposto alla XXXVIII riunione annuale dell'Associazione Italiana di Epidemiologia (Napoli)

Critelli R, Naccarati A, Assumma M, Polidoro S, Fasanelli F, Russo A, Modica F, Sacerdote C, Zwarthoff Ellen, Matullo G, Vineis P. **Mutation detection in urine from bladder cancer patients as non-invasive prognostic tool.** – poster presentato alla "The European Human Genetics Conference 2014" (Milano)

Presentazioni in atti di convegno

Fasanelli F, Zugna D. **Analisi di mediazione multipla per l'associazione tra depressione materna e sibili e fischi al torace nei primi 18 mesi di vita del bambino.** presentazione orale esposta al XLI Convegno AIE 2017 (Mantova, 25-27 ottobre 2017)

Fasanelli F, Ricceri F, Giraudo MT, Polidoro S et al. **L'analisi delle componenti principali applicata ad uno studio epigenome-wide innestato nella coorte EPIC Italia.** presentazione orale esposta al Convegno di Primavera AIE 2017 (Roma, 5-6 giugno 2017)

Sciannameo V, Carta, D'Errico A, Giraudo MT, Fasanelli F et al**. L'associazione tra tumore della vescica e esposizioni professionali: analisi pooled di due studi casocontrollo italiani.** presentazione orale esposta al XL Congresso AIE 2016 (Torino, 19-21 ottobre 2016)

Fasanelli F, Giraudo MT et al. **WEIGHTING APPROACH PER MEDIATORI MULTIPLI NELL'ANALISI DI SOPRAVVIVENZA.** presentazione orale esposta al XL Congresso AIE 2016 (Torino, 19-21 ottobre 2016)

Giraudo MT, Ricceri F, Fasanelli F et al. **Analisi di mediazione per l'associazione tra livello di istruzione, markers infiammatori e malattie cardiovascolari: risultati dallo studio EPIC-Italia.** presentazione orale esposta al XL Congresso AIE 2016 (Torino, 19-21 ottobre 2016)

Fasanelli F, Ricceri F, Trevisan M et al. **L'effetto protettivo della dieta mediterranea sul tumore del colon e' mediato dai livelli di metilazione nei geni dell'infiammazione? Uno studio caso controllo innestato nella coorte di epic italia.** presentazione orale esposta al XL Congresso AIE 2016 (Torino, 19-21 ottobre 2016)

Zugna D, Fasanelli F, Richiardi L. **Impatto della non-proporzionalità dei rischi in un'analisi di mediazione su dati di sopravvivenza.** presentazione orale esposta al XL Congresso AIE 2016 (Torino, 19-21 ottobre 2016)

Simeon V, Chiodini P, Mattiello A, Krogh V, Pala V, Fasanelli F et al. **Nuovo indicatore antropometrico di obesità centrale e identificazione del rischio di mortalità generale e per cancro nella coorte Italiana di EPIC.** presentazione orale esposta al XL Congresso AIE 2016 (Torino, 19-21 ottobre 2016)

Saieva C, Caini S, Ceroti M, Fasanelli F, Ricceri F, Agnoli C, Grioni S, Mattiello A, Santucci De Magistris M, Tumino R, Martorana C, Masala G. **Consumo di bevande alcoliche e rischio di tumori epiteliali nella coorte EPIC-ITALIA.** presentazione orale esposta al XXXIX Congresso AIE 2015 (Milano, 27-30 ottobre 2015)

Ricceri F, Fasanelli F, Giraudo MT, Sieri S, Pala V, Masala G, Ermini I, Giurdanella MC, Martorana C, Mattiello A, Chiodini P, Vineis P, Sacerdote C.**Abitudini alimentari e disuguaglianze sociali: l'esperienza della collaborazione EPICItalia.** presentazione orale esposta al XXXIX Congresso AIE 2015 (Milano, 27-30 ottobre 2015)

Fasanelli F, Ricceri F, Francia A, Zugna D, Giraudo MT, Krogh V, Grioni S, Mattiello A, Panico S, Masala G, Caini S, Tumino R, Frasca G, Vineis P, Sacerdote C. **DIETA MEDITERRANEA E RISCHIO DI CANCRO DEL COLON-RETTO: UN'ANALISI DI MEDIAZIONE.** presentazione orale esposta al XXXIX Congresso AIE 2015 (Milano, 27-30 ottobre 2015)

Ricceri F, Fasanelli F, Allione A, D'Errico A, Giraudo MT, Matullo G, Rapallo F, Roggero M, Terracini L, Vineis P, Sacerdote C. **USO DELLA STATISTICA ALGEBRICA PER STUDIARE L'INTERAZIONE GENE-GENE.** presentazione orale esposta all'VIII Congresso Nazionale SISMEC 2015 (Torino, 16-19 settembre 2015)

Ricceri F, Sacerdote C, Fasanelli F, EPIC-Interact Collaborators, Vineis P, Gonzales CA, Zamora-Ros R. **ASSOCIAZIONE TRA CONSUMO DI FLAVONOIDI E RIDUZIONE DEL RISCHIO DI TUMORE DELLA VESCICA: RISULTATI DELLO STUDIO EPIC EUROPA**. presentazione orale esposta alla XXXVIII riunione annuale dell'Associazione Italiana di Epidemiologia (Napoli, 5-7 novembre 2014)

Appendix

Marginal time-dependent causal effects in mediation analysis with survival data

Abstract

The main aim of mediation analysis is to study the direct (not mediated) and indirect (mediated) effects of an exposure on an outcome of interest. To date, the literature on mediation analysis with multiple mediators has mainly focused on continuous and dichotomous outcomes. However, development of methods for multiple mediation analysis of survival outcome is still limited. In this article, we show how to extend a method for multiple mediation analysis based on the computation of appropriate weights to survival outcome. The method is illustrated along with an estimation algorithm, assuming a proportional hazards model conditional on exposure, mediators and covariates and allowing for marginal direct and indirect effects to vary over time. The method is applied to an example from a dataset coming from a published study on mortality for prostate cancer where the interest was to understand to what extent the effect of DNA methyltransferase genotype on mortality was explained by DNA methylation and tumor aggressiveness. The approach described is straightforward and can be used to quantify the marginal time-dependent direct and indirect effects carried by multiple indirect pathways.

Keywords

multiple mediation analysis, natural indirect effect, proportional hazards model, pure direct effect, weighting approach

1 Introduction

In medical and epidemiological research it is often of interest to understand the biological or mechanistic pathways that contribute to the effect of an exposure on an outcome. The aim of mediation analysis is to disentangle the total effect of the exposure on the outcome into the indirect effect, i.e. the effect through intermediate variables (mediators), and the direct effect, i.e. the effect through pathways independent of the hypothesized mediators.

A first approach to mediation analysis was proposed by Baron and Kenny in 1986 [1]. The theory was later generalized through a counterfactual approach that gave more general definitions of the direct and indirect effects allowing the presence of nonlinearities and interactions between the exposure and the mediators in the models for the outcome [2, 3, 4, 5, 6].

In the counterfactual framework, the methods to estimate the direct and indirect effects differ according to the type of outcome. As far as the survival framework is concerned, a mediation approach involving a single mediator was firstly proposed by Lange et al. in [7] where an additive hazard model was employed to model the time to an event as the outcome of interest. Consequently Vanderweele in [8] discussed several effect measures in survival analysis and extended Lange's approach using both an accelerated failure time model and the Cox proportional hazards model with a rare outcome. These standard approaches in the presence of a single mediator were based on combining parameter estimates from the model for the outcome and for the mediator respectively, but the former was employable with a normal continuous mediator and the latter with rare outcomes. Tchetgen Tchetgen in [9] derived new estimators for mediation analysis for proportional hazards and additive hazards models with appealing robustness properties. Lange in [10] proposed a weighting approach for the proportional hazards model with a non-rare outcome. In a more recent work, Wang and Albert proposed a mediation formula approach for survival outcome with a normally distributed mediator [11].

Several methods have also been proposed to study mediation effects for scenarios where multiple mediators are considered [12, 13, 14, 15, 16, 17, 18, 19, 20] but only in [14, 15, 17, 18, 19] the focus was on survival analysis. The purpose of the present paper is to show how to extend to survival outcome the weighting approach for multiple mediators proposed by Vanderweele et al. in [13] focusing on proportional hazards models. The main advantage of the method is its applicability in frameworks where mediators are dependent on each other. Furthermore it does not require specific models for the mediators thus avoiding the problem of model incompatibility and, similarly to the other weighting approaches, it does not rely on the assumption of rare outcomes.

2 Definitions and assumptions

Let the non-negative random variable T denote the time until the occurrence of the event of interest and let U denote the censoring time. Hence (Y, Δ) are the observed data, where $Y = min(T, U)$, $\Delta = I(T \leq U)$ and $I(.)$ is the indicator function. Let $S_T(t)$ be the survival function, $\lambda_T(t)$ be the hazard function and $f_T(t)$ be the density function at time t. We assume the independence of T and U, so that $S_T(t)$, $\lambda_T(t)$ and $f_T(t)$ can be identified and consistently estimated. Let A be a dichotomous or a categorical exposure, with a and a^* two possible values of A, and let $\mathbf{M} = (M^1, \cdots, M^K)$ be the vector of multiple mediators. We suppose to be interested in evaluating how much of the effect of A on T is mediated through $\mathbf{M}$ jointly and through pathways other than through $\mathbf{M}$. Within the context of mediation in survival analysis, the decomposition of the total effect of an exposure on the outcome in the indirect and direct effects can be expressed in different ways and scales [8]. We will consider here the decomposition on multiplicative scale in terms of hazard functions. By indicating the counterfactual hazard function $\lambda_{T^a}(t)$ as the potential value of the hazard had the exposure A been set at a and the counterfactual $\lambda_{T^{a,\mathbf{m}}}(t)$ as the potential value of the hazard had the exposure A and the mediators $\mathbf{M}$ been set at a and $\mathbf{m}$ respectively, we can give the following formal definitions in terms of hazard functions:

- Total causal effect, $TCE(t) = \lambda_{T^a}(t)/\lambda_{T^{a^*}}(t)$;
- Pure direct effect, $PDE(t) = \lambda_{T^{a,\mathbf{M}^{a^*}}}(t)/\lambda_{T^{a^*,\mathbf{M}^{a^*}}}(t)$;
- Natural indirect effect, $NIE(t) = \lambda_{T^{a,\mathbf{M}^a}}(t)/\lambda_{T^{a,\mathbf{M}^{a^*}}}(t)$.

Briefly, the $TCE(t)$ expresses how much the hazard at time t would change if the exposure were changed from level a^* to level a uniformly in the population. The $PDE(t)$ expresses how much the hazard at time t would change if the exposure were set at $A = a$ versus $A = a^*$ but the mediators were kept at the level they would have taken had the exposure been set at $A = a^*$. Thus the PDE captures which part of the effect of the exposure on the outcome would be maintained if we were to disable the pathways from the exposure to the mediators. Finally, the $NIE(t)$ expresses how much the hazard at time t would change if the exposure were fixed at the level $A = a$ but the mediators were changed from the level they would have taken if $A = a^*$ to the level they would have taken if $A = a$. Thus the NIE captures the effect of the exposure on the outcome that operates through the mediators. Under the composition assumption $T^a = T^{a,\mathbf{M}^a}$, the total effect is given by the product of the natural indirect effect and the pure direct one $(TCE(t) = NIE(t) \cdot PDE(t))$.

In order to estimate the causal direct and indirect effects, several hypotheses need to be satisfied, specifically the absence of unmeasured confounders for the exposure-outcome relationship, exposure-mediators relatioships, mediators-outcome relationships and the absence of an effect of the exposure that itself confounds the mediators-outcome relationship.

The approach we propose in this paper is an extension of the method proposed for continuous and binary outcomes by Vanderweele and Vanstenlandt in [13] to survival outcome. The marginal hazard function can be estimated as the ratio between the marginal density and survival functions, both obtained by means of the mediation formula as follows:

$$\lambda_{T^{a,\mathbf{M}^{a^*}}}(t) = \frac{\mathbb{E}_{[C,\mathbf{M}]^{a^*}}\left[\frac{P(A=a^*)}{P(A=a^*|C)}\, f_T(t \mid A=a, \mathbf{M}, C)\right]}{\mathbb{E}_{[C,\mathbf{M}]^{a^*}}\left[\frac{P(A=a^*)}{P(A=a^*|C)}\, S_T(t \mid A=a, \mathbf{M}, C)\right]}. \tag{1}$$

A proof of (1) is provided in Appendix A. The approach is then based on inverse probability weighting. Its main feature is that it does not require models for the mediators but only for the exposure conditional on covariates and for the outcome conditional on the exposure, the mediators and the covariates. Exposure-mediator and mediators interactions can also be included and the independence between mediators is not necessary. However it allows only for binary or categorical exposures. Since the assumption of proportional hazards model may not hold for both the conditional and the marginal hazard function because of non-collapsibility [22], pure direct and natural indirect effects may vary over time in the presence of non-rare outcome.

3 The estimation procedure

The algorithm for the estimation of causal effects requires the computation at any time $\tilde{t}$ of three weighted averages that we will call $Q_1(\tilde{t})$, $Q_2(\tilde{t})$ and $Q_3(\tilde{t})$. If we suppose that $a = 1$ and $a^* = 0$, these weighted averages correspond to the counterfactual $\lambda_{T^{1,\mathbf{M}^0}}(\tilde{t})$, $\lambda_{T^{0,\mathbf{M}^0}}(\tilde{t})$ and $\lambda_{T^{1,\mathbf{M}^1}}(\tilde{t})$ respectively. The algorithm for the estimation of the effects at a specific time $\tilde{t}$ proceeds as follows:

1. Estimation of $\lambda_{T^{1,\mathbf{M}^0}}(\tilde{t})$:

$$\lambda_{T^{1,\mathbf{M}^0}}(\tilde{t}) = \frac{\mathbb{E}_{[C,\mathbf{M}]^0}\left[\frac{P(A=0)}{P(A=0|C)}\, f_T(\tilde{t} \mid A=1, \mathbf{M}, C)\right]}{\mathbb{E}_{[C,\mathbf{M}]^0}\left[\frac{P(A=0)}{P(A=0|C)}\, S_T(\tilde{t} \mid A=1, \mathbf{M}, C)\right]}$$

- for each subject with $A = 0$ the hazard function is modeled to obtain a predicted estimate of the density and of the survival functions at time $\tilde{t}$ separately if the subject had had $A = 1$ rather than $A = 0$, but using the individual's own values of mediators and covariates;
- two weighted averages of these predicted values are computed for subjects with $A = 0$ (each subject i is given a weight $\frac{P(A=0)}{P(A=0|\mathbf{C}_i)}$ where $\mathbf{C}_i$ denotes the actual covariate values for subject i);
- the ratio of the two weighted averages is computed.

2. <u>Estimation of $\lambda_{T^{0,\mathbf{M}^0}}(\tilde{t})$:</u>

$$\lambda_{T^{0,\mathbf{M}^0}}(\tilde{t}) = \frac{\mathbb{E}_{[C,\mathbf{M}]^0}\left[\frac{P(A=0)}{P(A=0|C)}\, f_T(\tilde{t} \mid A = 0, \mathbf{M}, C)\right]}{\mathbb{E}_{[C,\mathbf{M}]^0}\left[\frac{P(A=0)}{P(A=0|C)}\, S_T(\tilde{t} \mid A = 0, \mathbf{M}, C)\right]}$$

- for each subject with $A = 0$ the hazard function is modeled to obtain a predicted estimate of the density and of the survival functions at time $\tilde{t}$ separately using the individual's own values of exposure, mediators and covariates;
- two weighted averages of these predicted values are computed for subjects with $A = 0$ (each subject i is given a weight $\frac{P(A=0)}{P(A=0|\mathbf{C}_i)}$ where $\mathbf{C}_i$ denotes the actual covariate values for subject i);
- the ratio of the two weighted averages is computed.

3. <u>Estimation of $\lambda_{T^{1,\mathbf{M}^1}}(\tilde{t})$:</u>

$$\lambda_{T^{1,\mathbf{M}^1}}(\tilde{t}) = \frac{\mathbb{E}_{[C,\mathbf{M}]^1}\left[\frac{P(A=1)}{P(A=1|C)}\, f_T(\tilde{t} \mid A = 1, \mathbf{M}, C)\right]}{\mathbb{E}_{[C,\mathbf{M}]^1}\left[\frac{P(A=1)}{P(A=1|C)}\, S_T(\tilde{t} \mid A = 1, \mathbf{M}, C)\right]}$$

- for each subject with $A = 1$ the hazard function is modeled to obtain a predicted estimate of the density and of the survival functions at time $\tilde{t}$ separately using the individual's own values of exposure, mediators and covariates;
- two weighted averages of these predicted values are computed for subjects with $A = 1$ (each subject i is given a weight $\frac{P(A=1)}{P(A=1|\mathbf{C}_i)}$ where $\mathbf{C}_i$ denotes the actual covariate values for subject i);

- the ratio of the two weighted averages is computed.

The probabilities $P(A = 0|\mathbf{C}_i)$ and $P(A = 1|\mathbf{C}_i)$ in the denominator of the weights are obtained by fitting suitable logistic regressions.

4. Computation of the effects: the pure direct effect, the natural indirect effect and the total causal effect at time $\tilde{t}$ can than be obtained as follows:

$$PDE(\tilde{t}) = \frac{Q_1(\tilde{t})}{Q_2(\tilde{t})} = \frac{\lambda_{T_{1M_0}}(\tilde{t})}{\lambda_{T_{0M_0}}(\tilde{t})},$$

$$NIE(\tilde{t}) = \frac{Q_3(\tilde{t})}{Q_1(\tilde{t})} = \frac{\lambda_{T_{1M_1}}(\tilde{t})}{\lambda_{T_{1M_0}}(\tilde{t})},$$

$$TCE(\tilde{t}) = NIE(\tilde{t}) \cdot PDE(\tilde{t}).$$

5. Computation of the confidence intervals of the effects: using bootstrapping.

The procedure described above can be repeated for a given sequence of times $\tilde{t}$ thus allowing to observe how the causal effects possibly vary over time. The density and survival functions can be estimated using the Royston-Parmar model [23, 24], a flexible parametric Cox model that allows the estimation of the baseline hazards using natural cubic splines.
All analyses were performed using the computing environment R (R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/). We report in Appendix B the R code for the implementation of the estimation algorithms described above.

4 Empirical data example

In this Section we illustrate usefulness of the methodology proposed using data from [21]. In that paper the relationships among DNA methyltransferase genotype (DNMT, polymorphism rs406193), DNA methylation, tumor aggressiveness (measured by means of the Gleason score) and long-term mortality for prostate cancer were studied. In particular, it was hypothesized that DNMT activity affected mortality directly and indirectly via tumor tissue methylation and Gleason score. It is known that DNA methylation is affected by the family of DNA methyltransferase enzymes (DNMTs), among which DNMT3b that is considered in the study. In a previous study [25] an association was found between tumor tissue DNA methylation in three selected genes (GSTP1, APC, RUNX3) and prostate cancer-specific mortality. Details on the study population and on DNMT3b genotyping methodology to target the single-nucleotide polymorphism rs406193 considered in the analysis are given in [21].

Some preliminary analyses were performed to assess the associations between the variables involved assuming that: i) the activity of DNMT3b affects the methylation status of the three genes GSTP1, APC and RUNX3; ii) the methylation status of these genes affects the Gleason score and not viceversa; iii) DNA methylation of these genes affects prostate cancer mortality directly and indirectly through Gleason score (Figure 1). In mediation analysis terms, the exposure was the DNMT3b variant (carriers of at least one T compared to CC carriers), the two mediators were the DNA methylation (coded with three levels corresponding to the number of methylated genes out of ADC, GSTP1 and RUNX3: 0-1, 2 or 3 respectively) and the Gleason score (coded as a dichotomous variable with the two levels corresponding to having or not a score ≥ 8) and the outcome was the time to death for prostate cancer. It is important to underline that the relationship between DNMT3b and the Gleason score could be mediated also by the unmeasured DNA methylation of further genes. Therefore in our analysis the direct effect of the exposure on the outcome included also the path DNMT3b→DNA methylation (APC, GSTP1, RUNX3 excluded)→prostate cancer mortality. The assumption that both the mediators followed temporally the exposure was obviously reasonable. The age at diagnosis, the source used for tumor tissue typing and the period of diagnosis were considered as potential confounders. We also assumed the absence of other unmeasured confounders of exposure-outcome, exposure-mediators, mediators-outcome associations and the absence of unmeasured/unknown mediators-outcome association affected by the exposure.

Firstly we report the results obtained by a standard regression approach and then we perform the mediation analysis through the weighting approach to estimate the magnitude of marginal direct and indirect effects. To estimate the effects of the DNMT3b variant on the number of methylated genes and on the level of Gleason score, an ordinal logistic regression model and a logistic regression model were used respectively. While there was no evidence of association between carriers of the rs406193 T allele and the number of methylated genes (adjusted odds ratio of each increase in the number of methylated genes = 0.84, 95% confidence interval (CI): $0.57 - 1.23$), an association was found with the levels of Gleason score (adjusted odds ratio of having a score of 8 or more = 0.57, 95% CI: $0.39, 0.85$). Moreover, there was also an association between the two candidate mediators (adjusted odds ratio of having a Gleason score of 8 or more = 1.45, 95% CI: $1.08, 1.94$, the DNMT3b

variant was considered among the covariates).
A Royston-Parmar regression model was fitted to estimate the mutually adjusted associations between the two mediators and the outcome. This model was also adjusted for the exposure, the age at diagnosis, the source of tumor tissue and the period of diagnosis. It was found that subjects with 2 or 3 methylated genes had an increased risk of mortality compared to those with 0-1 methylated genes (adjusted hazards ratio: 1.50, 95% CI: $1.02, 2.24$ for 2 versus 0-1 and 1.98, 95% CI: $1.26, 3.12$ for 3 versus 0-1). For subjects with a Gleason score higher than 8 the adjusted hazards ratio was 2.49, 95% CI: $1.79, 3.46$. We applied the weighting approach adapted to survival outcomes to test how much of the protective effect of the variant on prostate cancer mortality could be mediated by a decrease in the number of methylated genes and in the level of Gleason score, on their turn associated with mortality. We estimated direct and indirect effects over about 100 equidistant time values between the minimum value and the maximum value of observed survival times.

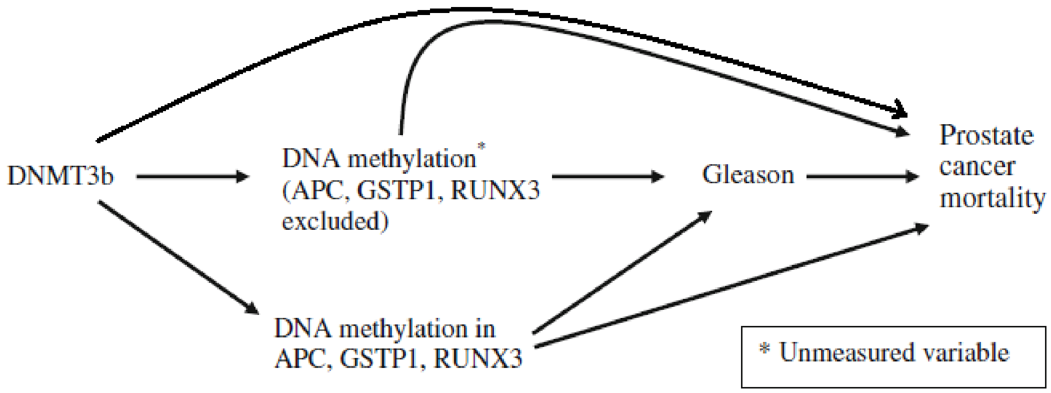


Figure 1: The assumed causal relationships. DNMT3b genotype is the exposure variable evaluated in association with prostate cancer mortality. DNA methylation in APC, GSTP1, and RUNX3 genes is considered as an intermediate variable. Gleason score is the further intermediate variable. The relationship between DNMT3b and the Gleason score could be mediated on its turn by unmeasured DNA methylation of further genes. The direct effect of the exposure on the outcome is represented by the pathways that do not cross the two measured mediators, therefore it also includes the path DNMT3b→DNA methylation (APC, GSTP1, RUNX3 excluded)→prostate cancer mortality.

Figure (2) shows the results of mediation analysis through the plots of the causal effects as functions of time. The pure direct effect appeared to be always close to

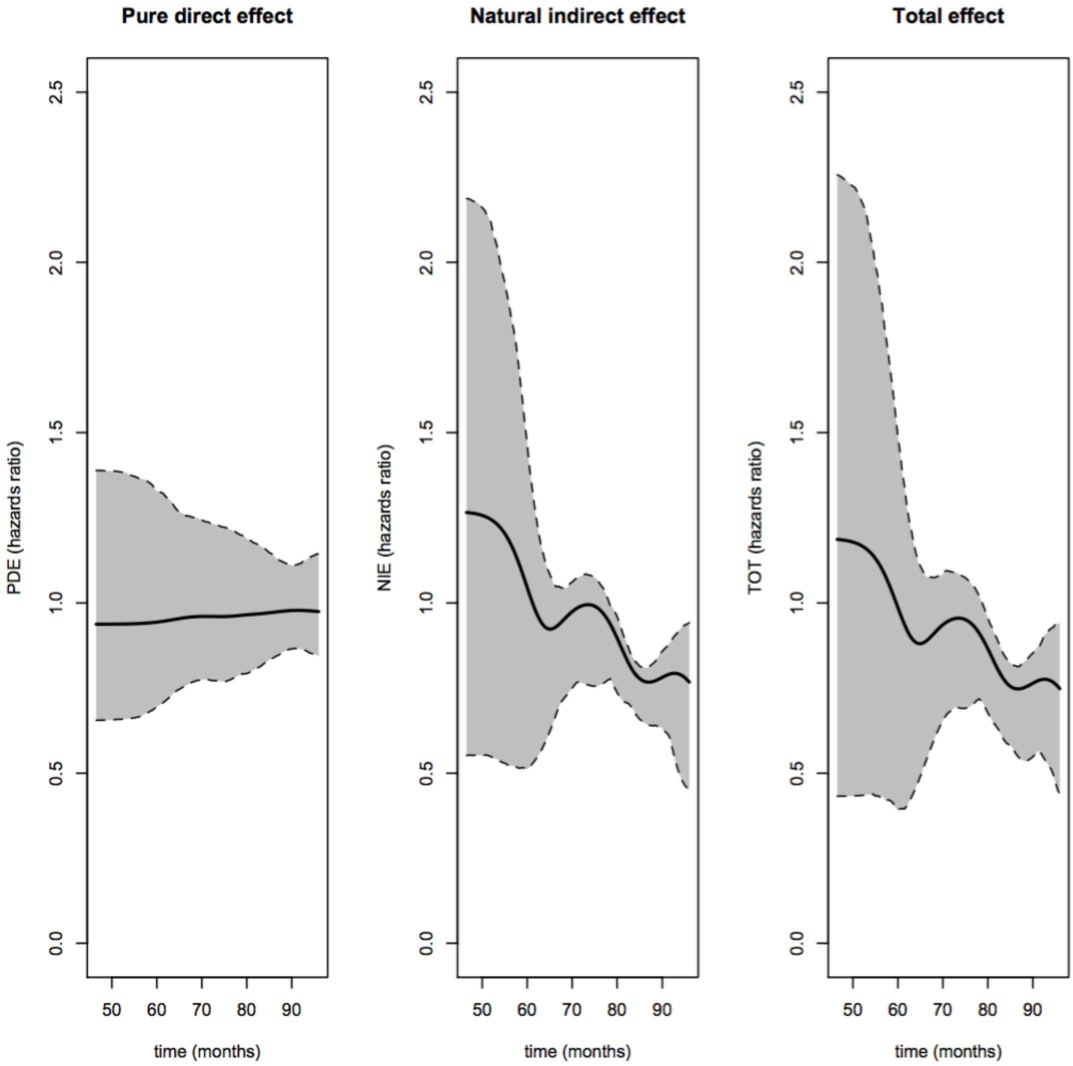


Figure 2: Results of mediation analysis considering DNA methylation and Gleason score as mediators.

	DNMT3b rs406193		
	CC	CT+TT	95% CI
PDE($\tilde{t} = 78$)	1	0.96	0.78, 1.20
NIE($\tilde{t} = 78$) (through DNA methylation and Gleason score)	1	0.95	0.77, 1.00
TCE($\tilde{t} = 78$)	1	0.91	0.72, 1.02
PDE($\tilde{t} = 90$)	1	0.97	0.86, 1.11
NIE($\tilde{t} = 90$) (through DNA methylation and Gleason score)	1	0.78	0.63, 0.86
TCE($\tilde{t} = 90$)	1	0.76	0.55, 0.85

Table 1: Causal effects at times $\tilde{t} = 78$ and $\tilde{t} = 90$ months (CI= confidence interval; PDE=pure direct effect; NIE=natural indirect effect; TCE= total causal effect).

the unit value over time and hence the TCE seemed to be explained mostly by the NIE.

Table 1 shows in details the causal effects estimated at the two values of time $\tilde{t} = 78$ and $\tilde{t} = 90$ months (corresponding to the median and to the 95th percentile of the observed survival times respectively). At 78 months from diagnosis, the TCE on mortality risk for prostate cancer was 0.91 (95% CI: $0.72, 1.02$), the NIE was 0.95 (95% CI: $0.77, 1.00$) and the PDE was 0.96 (95% CI: $0.78, 1.20$). The direct and indirect effects were therefore similar to each other. At 90 months, the TCE on mortality risk for prostate cancer for patients carrying the variant was stronger (0.76, 95% CI: $0.55, 0.85$), the NIE was 0.78 (95% CI: $0.63, 0.86$) and the PDE was 0.97 (95% CI: $0.86, 1.11$). The analysis then suggests that at 90 months from diagnosis the total effect of the variant on the cause-specific mortality is mostly attributable to the indirect effect through tumor tissue methylation and Gleason score. However the estimates obtained of the PDE and the NIE and, hence, of the TCE, could be biased by the presence of some unmeasured mediator-outcome confounders such as a number of possible non-epigenetic molecular signatures pointing toward Gleason score and prostate cancer mortality.

To explore the role of single mediators, we conducted an additional analysis by including only DNA methylation as a mediator. The models with and without Gleason score may be not directly comparable because of non-collapsibility of hazards ratio, however if this phenomenon is assumed not to affect greatly the estimates as well as the models aptness, this analysis may suggest the extent at which the addition of Gleason score as a second mediator modifies the direct and indirect effect estimates. Appendix C contains the results of this analysis. In this case the PDE resulted to be protective, but its strength was decreasing over time. For times greater than 85 months the results were similar to those obtained in the analysis with both mediators. In this case the indirect effect incorporates all the pathways through DNA methylation in APC, GSTP1 and RUNX3 including the path exposure→first mediator→second mediator→outcome. Therefore, the comparison between the results obtained by the two mediation analyses (with and without Gleason score) suggests that Gleason score has a relevant role in explaining the protective effect of DNMT3b on prostate cancer mortality, independently from DNA methylation in APC, GSTP1 and RUNX3 genes, for shorter times.

5 Discussion

In this article we have introduced a procedure to estimate pure direct and natural indirect effects through multiple mediators in survival settings by showing how to extend the weighting approach proposed by Vanderweele et al. in [13] to survival outcomes. The applications to real data highlight the practical utility of the method proposed.

Few methods have been introduced in literature for multiple mediation analysis with survival data. A simple approach that can be used with any generalized linear model including survival ones was developed by Tchetgen Tchetgen et al. in [14]. The method estimates conditional causal effects using inverse odds ratio weighting. It allows to include multiple mediators of a categorical, discrete or continuous nature and binary or continuous exposures. The approach has the advantage of overcoming the need to specify possible interactions between the exposure and the mediators and it can be implemented with standard softwares [26]. Its main limitation is that difficulties may arise in detecting small indirect effects. In the same year, Lange et al. proposed a weighting approach for multiple mediation that can be used for most types of outcomes, including survival outcomes [15]. The method is applicable to all types of mediators and exposures. It requires distinct causal pathways for the mediators and shows a worse performance in the case of continuous mediators. Despite being a weighting approach, the estimation procedure requires besides a model for the exposure also a model for each mediator in the construction of the weights. Huang et al. proposed another multi-mediator model devised specifically only for survival data [17]. This is a regression-based approach where the survival distribution is modeled through a flexible semiparametric probit model and the mediators are modeled through linear regressions. The approach requires continuous mediators and a continuous or binary exposure. Its main advantage is that it allows the examination of path-specific effects of each mediator. However it carries the limitation to be employable only in a low-dimensional setting (one or two mediators).

More recently, Huang and Yang [18] have proposed methods for multi-mediator analyses using Aalen additive hazards models, Cox proportional hazards models with rare outcomes and semiparametric probit models. They have provided closed-form expressions for path-specific effects requiring models for the mediators with normal errors. Lin et al. have proposed an approach to estimate interventional analogues of direct and indirect effects through a survival mediational g-formula [19]. The approach has been inspired by the one proposed previously in [12] and can be used with time-varying exposures, mediators, and confounders. However the outcome only focuses on survival probability at the end of follow-up and the extension to different survival models such as the proportional hazards model is proposed as a future perspective.

The method described in the present paper allows the estimation of marginal causal effects assuming the validity of the proportional hazards hypothesis for observed hazard function conditional on exposure, mediators and covariates. To obtain the causal mediation effects on the hazard function scale, the method requires the choice of a grid of times at which the effects have to be estimated. The result is a time-dependent estimation of the causal effects that allows to investigate how the mediation effects change as a function of time. For this aspect, the method is similar to that proposed by Wang and Albert in [11], but has the advantage of being able to be used in the presence of multiple mediators of any nature, also not normally distributed. It bears also the advantages of allowing the presence of exposure-mediators and mediators interactions and of requiring neither models for the mediators nor their independence. Moreover, the method can be used also with non-rare outcomes, while in the presence of rare outcomes we expect the estimated effects to be constant over time. The estimation performance of the method is highly dependent on the validity of the assumptions listed in Section 2. The hypothesis of the correct specification of the model for the outcome is crucial and the bias due to misspecification of this model will be the subject of a future work. No constraints are imposed for extensions of the approach to other survival models.

A limitation of the present method is its inability to characterize the path-specific effects of each mediator [27]. Several procedures have been proposed in literature under various settings [16, 28, 29, 30, 31] but none explicitly for survival analysis except for the ones in [17, 18]. Since proportional hazards models are commonly used in biomedical research, the development of methodologies for mediation analysis that enable to incorporate multiple mediators and to characterize the path-specific effects may be an important direction for future research. Finally the procedure is based on the computation of weights, which could become unstable if large weights are given to very few subjects included in the dataset, and it can be used only with binary or categorical exposures. In fact, although the paper primarily focuses on binary exposures, the approach equally applies for categorical exposures considering a fixed category as the reference and estimating the causal effects for each of the others with respect to that one.

The main contribution of this paper is to give a useful tool in mediation analysis in the presence of multiple mediators and survival outcomes. The proposed approach involves probability weights that relate the exposure, the mediators and the confounders and therefore can be implemented in most standard regression softwares, provided that a weight is assigned to each observation.

A Proof of expression (1)

In the following we extend the weighting approach for multiple mediators proposed in [13] to the case where the time-to-event outcome is described by means of hazard functions. We show in detail the case of continuous mediators, but the binary or categorical cases can be treated in the same way substituting integrals with sums.

Theorem A.1. *Consider a binary exposure A, a vector of continuous mediators $\boldsymbol{M}$, a set of covariates C and a time-to-event outcome T. Suppose that:*

- *conditional on C, there is no unmeasured exposure-outcome confounding; it follows that:*
$$T^{a,\boldsymbol{m}} \amalg A \mid C \quad \forall a, \boldsymbol{m}; \tag{2}$$
- *conditional on A and C, there is no unmeasured mediators-outcome confounding; it follows that:*
$$T^{a,\boldsymbol{m}} \amalg \boldsymbol{M} \mid \{A, C\} \quad \forall a, \boldsymbol{m}; \tag{3}$$
- *conditional on C, there is no unmeasured exposure-mediators confounding; it follows that:*
$$\boldsymbol{M}^a \amalg A \mid C \quad \forall a; \tag{4}$$
- *there is no effect of the exposure that itself confounds the mediators-outcome relationship; it follows that:*
$$T^{a,\boldsymbol{m}} \amalg \boldsymbol{M}^{a^*} \mid C \quad \forall a, a^*, \boldsymbol{m}. \tag{5}$$

*Under the additional assumptions (**consistency**):*

$$\textit{if } A = a \textit{ then } T^a = T \textit{ and } \boldsymbol{M}^a = \boldsymbol{M} \tag{6}$$

$$\textit{if } A = a \textit{ and } \boldsymbol{M} = \boldsymbol{m} \textit{ then } T^{a,\boldsymbol{m}} = T \tag{7}$$

the counterfactual hazard $\lambda_{T^{a,\boldsymbol{M}^{a^*}}}(t)$ can be written as:

$$\lambda_{T^{a,\boldsymbol{M}^{a^*}}}(t) = \frac{\mathbb{E}_{[C,\boldsymbol{M}]^{a^*}}\left[\frac{P(A=a^*)}{P(A=a^*|C)}\, f_T(t \mid A = a, \mathbf{M}, C)\right]}{\mathbb{E}_{[C,\boldsymbol{M}]^{a^*}}\left[\frac{P(A=a^*)}{P(A=a^*|C)}\, S_T(t \mid A = a, \mathbf{M}, C)\right]} \tag{8}$$

where $\mathbb{E}_{[C,\boldsymbol{M}]^{a^*}}$ indicates the expectation with respect to the joint density of $\mathbf{M}$ and C conditional on $A = a^*$.

Proof.

Consider first the conditional density function $f_{T^{a,\mathbf{M}^{a^*}}}(t \mid c)$ for $T^{a,\mathbf{M}^{a^*}}$.
Using the law of total probability and assumptions (5), (2) and(4) in sequence:

$$\begin{aligned} f_{T^{a,\mathbf{M}^{a^*}}}(t \mid c) &= \int_{\mathbf{m}} f_{T^{a,\mathbf{m}}}(t \mid \mathbf{M}^{a^*} = \mathbf{m}, c)\, f_{\mathbf{M}^{a^*}|C}(\mathbf{m} \mid c)\, \mathbf{dm} = \\ &= \int_{\mathbf{m}} f_{T^{a,\mathbf{m}}}(t \mid c)\, f_{\mathbf{M}^{a^*}|C}(\mathbf{m} \mid c)\, \mathbf{dm} = \\ &= \int_{\mathbf{m}} f_{T^{a,\mathbf{m}}}(t \mid A = a, c)\, f_{\mathbf{M}^{a^*}|C}(\mathbf{m} \mid c)\, \mathbf{dm} = \\ &= \int_{\mathbf{m}} f_{T^{a,\mathbf{m}}}(t \mid A = a, c)\, f_{\mathbf{M}^{a^*}|C,A}(\mathbf{m} \mid c, A = a^*)\, \mathbf{dm}. \end{aligned}$$

Then, applying (3) and (6) it follows:

$$f_{T^{a,\mathbf{M}^{a^*}}}(t \mid c) = \int_{\mathbf{m}} f_{T^{a,\mathbf{m}}}(t \mid A = a, \mathbf{M} = \mathbf{m}, c)\, f_{\mathbf{M}|C,A}(\mathbf{m} \mid c, A = a^*)\, \mathbf{dm}.$$

Using (7), we have:

$$f_{T^{a,\mathbf{M}^{a^*}}}(t \mid c) = \int_{\mathbf{m}} f_{T}(t \mid A = a, \mathbf{M} = \mathbf{m}, c)\, f_{\mathbf{M}|C,A}(\mathbf{m} \mid c, A = a^*)\, \mathbf{dm}. \tag{9}$$

Consider now the counterfactual density function $f_{T^{a,\mathbf{M}^{a^*}}}(t)$. Using the law of total probability and then the result obtained in the previous point, we have:

$$\begin{aligned} f_{T^{a,\mathbf{M}^{a^*}}}(t) &= \int_{c} f_{T^{a,\mathbf{M}^{a^*}}}(t \mid c) f(c)\, dc = \\ &= \int_{c} \left[\int_{\mathbf{m}} f_{T}(t \mid A = a, \mathbf{M} = \mathbf{m}, c)\, f_{\mathbf{M}|C,A}(\mathbf{m} \mid c, A = a^*)\, \mathbf{dm} \right] f(c)\, dc. \end{aligned}$$

Using now the equality:

$$f_{\mathbf{M},C,A}(\mathbf{m}, c, A = a^*) = f_{\mathbf{M}|C,A}(\mathbf{m} \mid c, A = a^*) \cdot P(A = a^* \mid C = c) \cdot f(c),$$

we have:

$$\begin{aligned} f_{T^{a,\mathbf{M}^{a^*}}}(t) &= \int_{c}\int_{\mathbf{m}} f_{T}(t \mid A = a, \mathbf{m}, c)\, \frac{f_{\mathbf{M},C,A}(\mathbf{m}, c, A = a^*)}{P(A = a^* \mid C = c)\, f(c)}\, f(c)\, d\mathbf{m}\, dc = \\ &= \int_{c}\int_{\mathbf{m}} f_{T}(t \mid A = a, \mathbf{m}, c)\, \frac{f_{\mathbf{M},C|A}(\mathbf{m}, c \mid A = a^*) P(A = a^*)}{P(A = a^* \mid C = c)}\, d\mathbf{m}\, dc = \\ &= \mathbb{E}_{[C,\mathbf{M}]^{a^*}} \Big[\frac{P(A = a^*)}{P(A = a^* \mid C)}\, f_{T}(t \mid A = a, \mathbf{M}, C) \Big]. \end{aligned}$$

Similarly, considering the survival function $S_{T^{a,\mathbf{M}^{a^*}}}(t)$ for $T^{a,\mathbf{M}^{a^*}}$, we have:

$$S_{T^{a,\mathbf{M}^{a^*}}}(t) = \mathbb{E}_{[C,\mathbf{M}]^{a^*}}\Big[\frac{P(A=a^*)}{P(A=a^* \mid C)}\ S_T(t \mid A=a,\mathbf{M},C)\Big].$$

Finally, we have:

$$\lambda_{T^{a,\mathbf{M}^{a^*}}}(t) = \frac{f_{T^{a,\mathbf{M}^{a^*}}}(t)}{S_{T^{a,\mathbf{M}^{a^*}}}(t)} = \frac{\mathbb{E}_{[C,\mathbf{M}]^{a^*}}\Big[\frac{P(A=a^*)}{P(A=a^*|C)}\ f_T(t \mid A=a,\mathbf{M},C)\Big]}{\mathbb{E}_{[C,\mathbf{M}]^{a^*}}\Big[\frac{P(A=a^*)}{P(A=a^*|C)}\ S_T(t \mid A=a,\mathbf{M},C)\Big]}. \quad \square$$

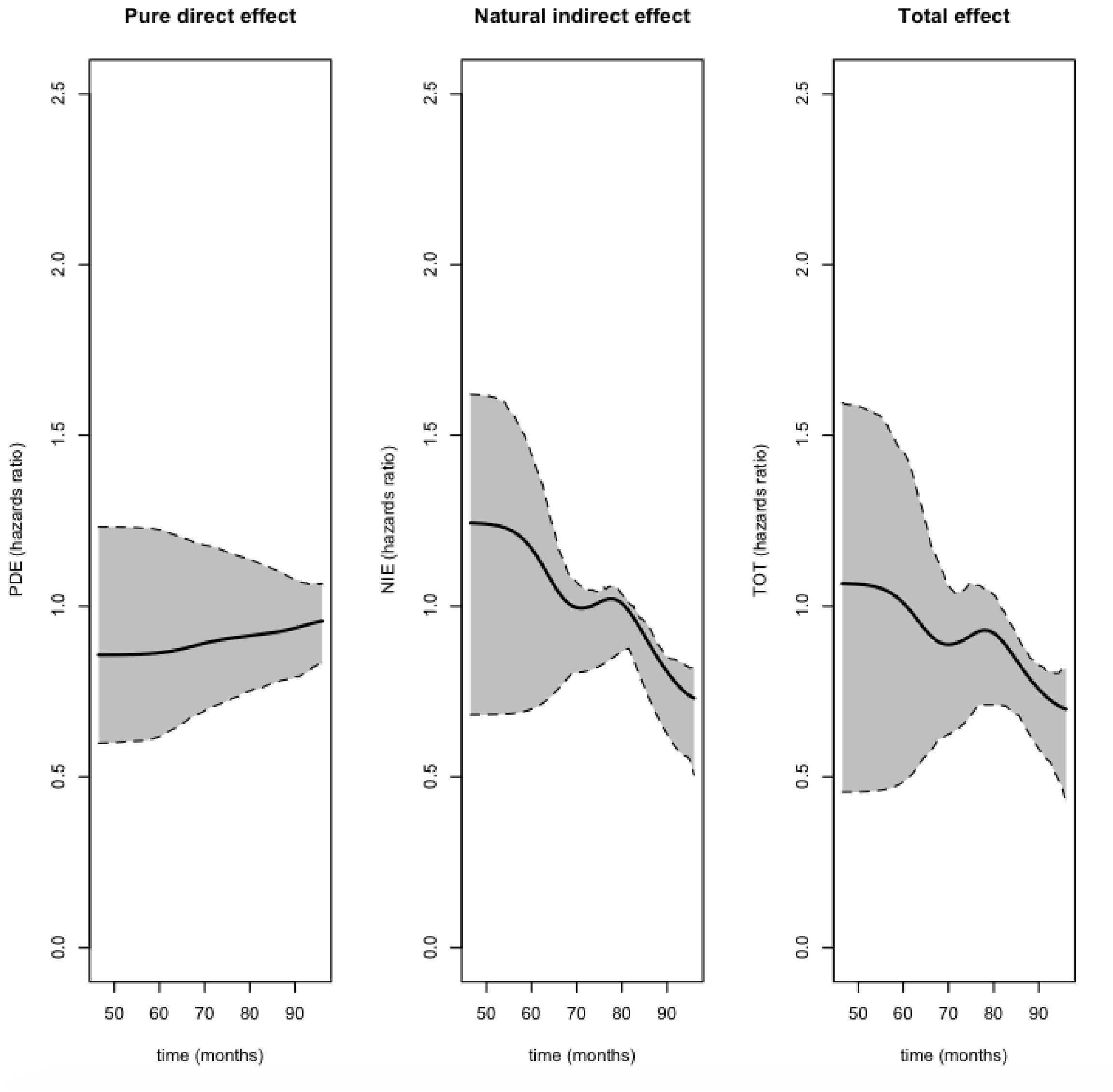
B R Code for direct, indirect and total effect estimates

```
##########################################################
# mydata is a dataframe with the following columns:
         # a is the binary exposure
         # m1 is the first mediator
         # m2 is the second mediator
         # c1 and c2 are two potential confounders
         # time is the follow up time
         # event is the status indicator (0=no, 1=yes)
##########################################################

# choose the correct library in R
library(rstpm2)
library(boot)

pde2=vector()
nie2=vector()
tot2=vector()

#definition of a grid of times at which the effects have to be estimated
step=0.5
times=seq(min(mydata$time),max(mydata$time),step)

###################
# estimation procedure
###################

theta= function(mydata,indices) {
d=mydata[indices,]
# compute the weights
logit=glm(formula = a ~ c1+c2, family = "binomial", data = d)
prob=data.frame(predict(logit, d, type="response"))
colnames(prob)[1]="p"
prob$pdm=prob$p
prob$pdm[d$a==0]=1-prob$p[d$a==0]
prob$w=1/prob$pdm
# model the hazard function
fit=stpm2(Surv(time,event) ~ a+m1+m2+c1+c2, data=d, df=5)
nstep=length(tempi)-1
# for each time t....
for (i in 1:nstep)
  {
# predict the density and the survival forcing A=0
pred.frame0=d
pred.frame0$a=0
pred.frame0$time=tempi[i+1]
surv0=data.frame(predict(fit,se.fit=TRUE,newdata=pred.frame0,type="surv"))
dens0=data.frame(predict(fit,se.fit=TRUE,newdata=pred.frame0,type="density"))
# predict the density and the survival forcing A=1
pred.frame1=d
pred.frame1$a=1
pred.frame1$time=tempi[i+1]
surv1=predict(fit,se.fit=TRUE,newdata=pred.frame1,type="surv")
dens1=predict(fit,se.fit=TRUE,newdata=pred.frame1,type="density")
# lambda[T0M0]
dens00=weighted.mean(dens0$Estimate[d$a==0],prob$w[d$a==0])
surv00=weighted.mean(surv0$Estimate[d$a==0],prob$w[d$a==0])
coef00=dens00/surv00
# lambda[T1M1]
dens11=weighted.mean(dens1$Estimate[d$a==1],prob$w[d$a==1])
surv11=weighted.mean(surv1$Estimate[d$a==1],prob$w[d$a==1])
coef11=dens11/surv11
# lambda[T1M0]
dens10=weighted.mean(dens1$Estimate[d$a==0],prob$w[d$a==0])
surv10=weighted.mean(surv1$Estimate[d$a==0],prob$w[d$a==0])
coef10=dens10/surv10
# compute the effects estimate
pde2[i]=coef10/coef00
nie2[i]=coef11/coef10
tot2[i]=pde2[i]*nie2[i]
}
return(c(pde2,nie2,tot2))
}
```

```
###################
# bootstrap of the estimation procedure
###################

replicates=2000
results <- boot(data=mydata, statistic=theta, R=replicates)
pde=matrix(,nrow=length(times)-1,ncol=3)
nie=matrix(,nrow=length(times)-1,ncol=3)
tot=matrix(,nrow=length(times)-1,ncol=3)

for (i in 1:length(times)-1)
{
interval_conf=boot.ci(results, type="bca", index=i)
pde[i,1]=results$t0[i]
pde[i,2]=interval_conf$bca[4]
pde[i,3]=interval_conf$bca[5]

interval_conf=boot.ci(results, type="bca", index=i+length(times)-1)
nie[i,1]=results$t0[i+length(times)-1]
nie[i,2]=interval_conf$bca[4]
nie[i,3]=interval_conf$bca[5]

interval_conf=boot.ci(results, type="bca", index=i+2*(length(times)-1))
tot[i,1]=results$t0[i+2*(length(times)-1)]
tot[i,2]=interval_conf$bca[4]
tot[i,3]=interval_conf$bca[5]
}

###################
# plot of the effects
###################

x=times[2:length(times)]
# pure direct effect
F=pde[1:length(times)-1,1]
L=pde[1:length(times)-1,2]
U=pde[1:length(times)-1,3]
```

```
par(mfrow=c(1,3))
plot(x,F,ylim=c(0,2.5),type="l",main="Pure direct effect",xlab="time",
                                               ylab="PDE")
polygon(c(x,rev(x)),c(L,rev(U)),col = "grey75", border = FALSE)
lines(x, F, lwd = 2)
lines(x, U, col="black",lty=2)
lines(x, L, col="black",lty=2)
# natural indirect effect
F1=nie[1:length(times)-1,1]
L1=nie[1:length(times)-1,2]
U1=nie[1:length(times)-1,3]
plot(x,F1,ylim=c(0,2.5),type="l", main="Natural indirect effect",xlab="time",
                                                  ylab="NIE")
polygon(c(x,rev(x)),c(L1,rev(U1)),col = "grey75", border = FALSE)
lines(x, F1, lwd = 2)
lines(x, U1, col="black",lty=2)
lines(x, L1, col="black",lty=2)
# total effect
F2=tot[1:length(times)-1,1]
L2=tot[1:length(times)-1,2]
U2=tot[1:length(times)-1,3]
plot(x,F2,ylim=c(0,2.5),type="l", main="Total effect",xlab="time",
                                                 ylab="TOT")
polygon(c(x,rev(x)),c(L2,rev(U2)),col = "grey75", border = FALSE)
lines(x, F2, lwd = 2)
lines(x, U2, col="black",lty=2)
lines(x, L2, col="black",lty=2)
```

C Results of mediation analysis considering only DNA methylation as mediator



	DNMT3b rs406193		
	CC	CT+TT	95% CI
PDE($\tilde{t} = 78$)	1	0.91	0.74, 1.15
NIE($\tilde{t} = 78$)	1	1.02	0.85, 1.06
(through only DNA methylation)			
TCE($\tilde{t} = 78$)	1	0.93	0.71, 1.05
PDE($\tilde{t} = 90$)	1	0.94	0.79, 1.08
NIE($\tilde{t} = 90$)	1	0.80	0.62, 0.85
(through only DNA methylation)			
TCE($\tilde{t} = 90$)	1	0.76	0.58, 0.83

References

1. Baron RM, Kenny D. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *J Personal Disord.* 1986; 51(6): 1173-1182.

2. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology.* 1992; 3(2): 143-155.

3. Pearl J. Direct and indirect effects. In: Breese J, Koller D, eds. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence.* San Francisco, CA: Morgan Kaufmann; 2001:411-420.

4. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface.* 2009; 2: 457-468.

5. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *Am J Epidemiol.* 2010; 172(12): 1339-1348.

6. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychological Methods.* 2010; 15(4): 309-334.

7. Lange T, Hansen JV. Direct and Indirect Effects in a Survival Context. *Epidemiology.* 2011; 22(4): 575-581.

8. Vanderweele TJ. Causal Mediation Analysis With Survival Data. *Epidemiology.* 2011; 22(4): 582-585.

9. Tchetgen Tchetgen EJ. On causal mediation analysis with a survival outcome. *Int J Biostat.* 2011; 7(1): 33.

10. Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. *Am J Epidemiol.* 2012; 176(3): 190-195.

11. Wang W, Albert JM. Causal Mediation Analysis for the Cox Proportional Hazards Model with a Smooth Baseline Hazard Estimator. *J R Stat Soc Ser C Appl Stat.* 2017; 66(4): 741-757.

12. Zheng W, van der Laan MJ. Causal mediation in a survival setting with time-dependent mediators. 2012.

13. VanderWeele TJ, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiologic Methods.* 2013; 2(1): 95-115.

14. Tchetgen Tchetgen EJ. Inverse odds ratio-weighted estimation for causal mediation analysis. *Stat Med.* 2013; 32(26): 4567-4580.

15. Lange T, Rasmussen M, Thygesen LC. Assessing Natural Direct and Indirect Effects Through Multiple Pathways. *Am J Epidemiol.* 2013; 179(4): 513-518.

16. Daniel RM, De Stavola BM, Cousens SN, et al. Causal mediation analysis with multiple mediators. *Biometrics.* 2014; 71(1): 1-14.

17. Huang YT, Cai T. Mediation Analysis for Survival Data Using Semiparametric Probit Models. *Biometrics.* 2016; 72(2): 563-574.

18. Huang YT, Yang HI. Causal Mediation Analysis of Survival Outcome with Multiple Mediators. *Epidemiology* 2017; 28(3): 370-378.

19. Lin SH, Young JG, Logan R, et al. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat Med.* 2017; 36(26): 4153-4166.

20. Steen J, Loeys T, Moerkerke B, et al. Flexible Mediation Analysis With Multiple Mediators. *Am J Epidemiol.* 2017; 186(2): 184-193.

21. Gillio Tos A, Fiano V, Zugna D, et al. DNA methyltransferase 3b (DNMT3b), tumor tissue DNA methylation, Gleason score, and prostate cancer mortality: investigating causal relationships. *Cancer Causes Control.* 2012; 23(9): 1549-1555.

22. Greenland S, Pearl J, Robins JM. Confounding and Collapsibility in Causal Inference. *Statist. Sci.* 1999; 14(1): 29-46.

23. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med.* 2002; 21(15): 2175-2197.

24. Royston P. Flexible parametric alternatives to the Cox model: update. *The Stata Journal.* 2004; 4(1): 98-101.

25. Richiardi L, Fiano V, Vizzini L, et al. Promoter methylation in APC, RUNX3 and GSTP1 and mortality in prostate cancer patients *J Clin Oncol.* 2009; 27(19): 3161-3168.

26. Nguyen QC, Osypuk TL, Schmidt NM, et al. Practical Guidance for Conducting Mediation Analysis With Multiple Mediators Using Inverse Odds Ratio Weighting. *Am J Epidemiol.* 2015; 181(5): 349-356.

27. Avin C, Shpitser I, Pearl J. Identifiability of Path-Specific Effects. In: Kaelbling LP, Saffioti A, eds. *Proceedings of the International Joint Conferences on Artificial Intelligence.* Denver, CO: Professional Book Center; 2005: 357-363.

28. Albert JM, Nelson S. Generalized Causal Mediation Analysis. *Biometrics.* 2011; 67(3): 1028-1038.

29. VanderWeele TJ, Vansteelandt S, Robins JM. Methods for effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology.* 2014; 25(2): 300-306.

30. Taguri M, Featherstone J, Cheng J. Causal mediation analysis with multiple causally non-ordered mediators. [published online ahead of print November 23, 2015]. *Stat Methods Med Res.* (doi: 10.1177/0962280215615899).

31. Vansteelandt S, Daniel RM. Interventional Effects for Mediation Analysis with Multiple Mediators. *Epidemiology.* 2017; 28(2): 258-265.