From
the INSTITUTE OF EPIDEMIOLOGY II,
Helmholtz Zentrum München -
German Center for Environmental Health (GmbH)
Director: Adjunct. Prof. Dr. rer. hum. biol. Annette Peters

and

the Institute for Medical Informatics, Biometry and Epidemiology,
Ludwig-Maximilians-Universität München
Director: Prof. Dr. rer. nat. Ulrich Mansmann

Impact of Air Pollution Exposure on Genome-Wide DNA Methylation and its Association with Socio-Economic Status

Thesis
Submitted for a Doctoral Degree in Human Biology at the Faculty of Medicine,
Ludwig-Maximilians-Universität München

Tommaso Panni
from Ostra

München 2017

Printed with the approval of the Faculty of Medicine
of the Ludwig-Maximilians-Universität München

Supervisor:	Adjunct. Prof. Dr. rer. hum. biol. Annette Peters
Co-examiner:	Priv. Doz. Dr. rer. nat. Rudolf A. Jörres
Dean:	Prof. Dr. med. dent. Reinhard Hickel
Date of the Oral Examination:	22.09.2017

...we only had one chance to win the [Olympic] Games...

M. Ginobili

I know thy works, that thou art neither cold nor hot;
I would thou wert cold or hot
F. Dostoevskij, Possessed

Contents

List of Figures

List of Tables

List of Abbreviations

$PM_{2.5}$ Particulate Matter with aerodynamic diameter smaller than 2.5 μm

PM_{10} Particulate Matter with aerodynamic diameter smaller than 10 μm

PM_{coarse} Particulate Matter with aerodynamic diameter between 2.5 and 10 μm

NO_2 Nitrogen Dioxide

NO_x Mono-nitrogen oxides

CpG/CG Cytosine-phosphate-Guanine dinucleotide

DNA Deoxyribonucleic Acid

RNA Ribonucleic Acid

CHR Chromosome

CRP C-reactive Protein

NF-kB Nuclear Factor kappa-light-chain-enhancer of activated B cells

SES Socio-economic Status

KORA Kooperative Gesundheitsforschung in der Region Augsburg

NAS Normative Aging Study

VA Veteran Affairs

ESCAPE European Study of Cohorts for Air Pollution Effects

LUR Land-use Regression model

CVD Cardiovascular Disease

SNP Single Nucleotide Polymorphisms

SQN Subset Quantile Normalization

BMIQ Beta-Mixture Quantile Normalization

EM Expectation Maximization

SWAN Subset-quantile Within Array Normalization

WBC White Blood Cells

EWA Epigenome-wide Analysis

BMI Body Mass Index

SD Standard Deviation

FVC Forced vital capacity

FEV$_1$ Forced Expiratory Volume at timed intervals of 1.0 second

FEF$_{25\%-75\%}$ Forced Expiratory Flow 25–75%

PFT Pulmonary Function Test

NA Non Acceptable

N.S. Non Significant

FDR False Discovery Rate

ICC Intra Correlation Coefficient

AML Acute Myeloid Leukemia

GEE Generalized Estimating Equations

DMR Differentially Methylated Regions

MCA Multiple Correspondence Analysis

SEM Structural Equation Modeling

DAG Directed Acyclic Graph

TE Total Effect

DE Direct Effect

IE Indirect Effect

Chapter 1

Introduction

Ambient air pollution exposure has been a public health problem for decades. We can consider particulate air pollution any particle of any matter (solid, liquid and gaseous), that are suspended in the air. In this definition are included both products of combustion, mostly as result of human activity like smoke, fumes and soot or of condensation of vapors and oxidation of gases in the atmosphere. But they may also be natural particles like sea salt, windblown dust or pollen. In conclusion we can say that particulate air pollution is a collection of particles coming from different sources and different materials. Particles have been classified according to their aerodynamics characteristics by measuring the particle diameter in micrometers (μm). Initially, particles with aerodynamic diameter lower than 10 μm were considered as inhalable (the so called PM_{10}), but up to the '90s, evidence suggested that it might have been useful to isolate even smaller particles and a new category was created: $PM_{2.5}$, for particles with aerodynamic diameter lower than 2.5 μm. Nowadays, also particles with diameters lower that 1 μm are under observation. Adverse health effects after increased air pollution exposure include not only the respiratory system (covering asthma, pulmonary effects, lung functions and others), but a range of different outcomes including the cardiovascular system, the autonomic nervous system, endothelial functions and even increasing mortality and hospital admissions. However, differently than other more traditional risk factors (such as smoking), consequences of increased air pollution exposure are smaller. It is more difficult to isolate the effects from the noise of the natural human variability and fluctuation on either hospital admissions or mortality and it may be attributable to other factors. Accurate environmental assessment and sophisticated statistical techniques have been developed in order to correctly address the issue.

We can set the two key years 1936 and 1952 as the birth of air pollution epidemiology with the first studies on Meuse Valley fog and London smog, and 1970 as a landmark year with the first release of the Clean Air Act by the Environment Protection Agency (US EPA) where every state was invited to set and achieve the Natural Ambient Air Quality Standard. Nevertheless, a significant increase of number of publications on air pollution

exposure related with health effect has broken out only in the last two decades where the focus was also directed to short-term effects. In fact, it has been demonstrated by different studies that rate in cardiovascular disease exacerbation, hospital admissions and ischemic heart disease rate may raise in association with outdoor air pollution daily fluctuations.
Once this spectrum of health outcomes have been associated with an increase of ambient air pollution exposure, epidemiological research started to get "smaller". On one side, scientists started to look at the different PM components (metals and other elements), trying to better detail their combination. On another side they joined physicians and biologists in order to better understand the physiological factors that induce the increase of the adverse health effect rate. Most of diseases reflect the effect of several complex elements that combine their effect and the environmental influence has emerged as a key factor that can influence immunological response and lead to a pathological status.

This work touches two important points within the field of environmental epidemiology and specifically within air pollution research, therefore, after introducing the KORA study (source of most of the data analyzed), the thesis will be divided in two parts. The first one (which will take a greater part) will test whether air pollution exposure is related to epigenetic changes, a recently discovered area. Several pollutants will be considered and linked to the measure of DNA methylation in whole blood, with the aim of identifying novel methylation sites that could play a role in the path between ambient exposure and diseases. Our results increase the level of knowledge regarding the association between epigenetic biomarkers and environmental factors. The second one, instead, is considering the issue of confounding in air pollution research focusing mostly in the association between pollutants and Socio-economic status (SES) factors accounting for both area and individual level. Thanks to an alternative approach, the results of this study may add a piece to the discussion regarding the association between SES and air pollution exposure.

In Part I of this work we will focus mostly in the second aspect while in Part II we will deepen how geographical location have an indirect influence on personal air pollution exposure.

1.1 Epigenomics

Let's start with Part I. The key biological processes that have been found as associated with air pollution exposure in the alveoli are inflammatory responses, coagulation and oxidative stress. Several studies have confirmed that particulate matter exposure may potentially influence C-reactive protein (CRP), marker of inflammation, or fibrinogen, marker of coagulation or NF-kB, marker of oxidant mechanisms [Bind et al., 2012, Yang and Omaye, 2009]. CRP is produced by the liver and high values are evidence for an acute inflammation. Elevated CRP is considered as non-specific "marker" for disease, however, several studies suggest that even low but constant levels of internal inflammation can lead to age-related pathologies such as heart disease and neurodegenerative conditions. Fibrinogen instead, plays a key role in the clotting cascade. It is converted in fibrin and stabilizes blood clots after injuries. Fibrinogen promotes atherothrombosis thanks to its procoagulant and proinflammatory characteristics. Instead, NF-kB is a complex protein that controls transcription of DNA of a variety of pro inflammatory cytokines, enzymes responsible for inflammation mediators and immune receptors. These pathophysiologic mechanisms play a key role into the pathways that link air pollution exposure and ambient fine particles to both respiratory and cardiovascular disease, and have also been demonstrated in animal models [Cassee et al., 2013].

In recent years, air pollution exposure has been also linked to cancer development, including lung cancer [Soberanes et al., 2012, Zhao et al., 2013, Raaschou-Nielsen et al., 2013]. Pathophysiological mechanisms like inflammation and oxidative stress have been found to constitute plausible mediators but despite recent conclusion regarding the strength and the consistency of this scientific evidence, the extent to which these systematic effects are elicited by ambient pollution and which biological pathways are stimulated is still undetermined and under debate [Peters, 2012]. Therefore, studies that are helping to enlighten and detail the systemic impact of ambient particles need to take into account and substantiate the multi-organ involvement in response to inhalation of particulate matter. Moreover, deepening the knowledge regarding the genome, it has become more evident that genetics alone is not sufficient to explain the risk of common diseases. There are, in fact, non-genetic and extra-genetic factors that play an important role. Focusing on cardiovascular disease, Baccarelli et al. [Baccarelli et al., 2010] produced a clear conceptual model of how the different worlds influence each other. Epigenetics lies in the middle being influenced by genomics but also by the environment and these three elements combine for subclinical diseases that lead to cardiovascular diseases. Therefore it is no surprise to observe how epigenetics has arisen in the recent year as a key research area in both biomedicine and public health. A first and sharp definition of epigenetics was given by Sir Conrad Waddington in 1942, who defined it as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" [Waddington, 2012]. Now we can define it as the heritable changes in phenotype and gene expression that are occurring without a change in the genomic sequence. In fact, the prefix "Epi" comes from ancient Greek and means "upon", "above", "on", "on top of", "over" and defines something that is happening on genetics, over genetics, referring to non- and extra-genetics mechanisms. The most understood epigenetic markers are DNA methylation, histone modification and microRNA. In this work we will consider DNA methylation.

Within the epigenetic markers, DNA methylation is surely the best studied and understood. It represents a covalent modification that is heritable by somatic cells after cell division. It mostly occurs on CG dinucleotides when a methyl group (CH_3) is added at the 5-carbon of the cytosine ring resulting in 5-methylcytosine (5-mC). They represent approximately 2-5% of all cytosines in mammal genome. Being found in proximity to critically important cis elements within promoters, these methyl groups project into the major channels of DNA and are often found as associated with a repressed chromatin state and inhibit transcription [Orphanides and Reinberg, 2002]. DNA Methylation plays also an important role in maintaining genome integrity by transcriptional silencing of repetitive DNA sequences and endogenous transposons [Bestor, 1998, Hedges and Deininger, 2007]. Many studies have both observed a link between environmental exogenous factors and aberrant changes in DNA methylation at both experimental an epidemiological level at both global and gene-specific level. Moreover, it might mediate some toxicity mechanisms and responses to certain chemicals. Within the list of exposure associated with aberrant changes in DNA methylation we can find metals (cadmium, arsenic, nickel, chromium, methylmercury), Trichloroethylene (TCE), dichloroacetic acid (DCA), and trichloroacetic acid (TCA), air pollution, benzene, Hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) and Endocrine-disrupting Chemicals and Reproductive Toxicants (Diethylstilbestrol, Bisphenol A, Persistent Organic Pollutants, Dioxin) [Baccarelli and Bollati, 2009]. Several studies observed altered DNA methylation in a priori identified areas of the genome or candidate genes. For example, Tarantini et al. in 2009 focused they research on linking air pollution exposure and repetitive elements (such as long interspersed nuclear elements, LINE, and Alu) and iNOS [Tarantini et al., 2009]. Chanda et al. in 2006, instead, focused on promoters of *p53* and *p16* genes in order to ascertain whether perturbation of DNA methylation play a role in arsenicosis and cancer after chronic arsenic exposure [Chanda et al., 2006]. And, a last example is provided by the study carried by Christensen et al. in 2008, where they associated DNA hypermethylation in promoter of *APC*, *CCND2*, *CDKN2A*, *CDKN2B*, *HPPBP1* and *RASSF1* and asbestos exposure [Christensen et al., 2008]. All these studies and many others increased the importance and the focus around the environmental epigenetics but after the advent of genome-wide studies, researchers started to wonder why not to expand it also to epigenetic and give birth to epigenomics. Previous intermediate steps occurred between 2005 and 2010, but probably the greatest technological innovation in this direction landed in spring 2011, when the company Illumina released the Infinium 450k Beadchip, able to estimate DNA methylation in a number of CpG sites 20 times larger than his predecessor, the Illumina 27k. Since then, several groups around the world were not only trying to use this new innovative technology but also to develop methods and techniques in order to answer and address the novel challenges provided by the new pioneering tool. The first innovation of a genome-wide approach is allowing a hypothesis-free assessment of changes in regulation and activation of blood leukocytes, involved in CVD development as part of atherosclerotic plaque formation and inflammation initiation [Madjid et al., 2004], and may offer novel avenues for understanding the role of

environmental stressors. This might not seem completely true since the CG loci had not been selected completely at random, but focusing on hypothesized previous knowledge about the genes and the genetic areas involved. However, the never-before large number of loci involved and the wide range of genomic functional regions included (relation with CpG Islands, proximity to transcription start sites for coding genes and corresponding gene-bodies, 3'-UTRs and intergenic regions) allow to obtain a comprehensive DNA methylome. Even if the molecular mechanisms behind the mentioned relationships are mostly unclear, several studies showed biological pathways that link changes in DNA methylation of candidate genes to oxidative stress, immune deficiency, chronic inflammation, and other carcinogenesis-related biological processes [Brook and Rajagoapalan, 2010] that may alter gene expression [Baccarelli and Bollati, 2009]. Here's why the new hypothesis-free approach increased the excitement around epigenomics: the idea of explore so far untouched areas of the human methylome. Baccarelli et al. in 2010 [Baccarelli et al., 2010], went even beyond asking the question "how many epigenomes?". They observed how "epigenomics markers showed both tissue specificity and correlations across different tissues depending on the loci". Despite the enthusiasm and curiosity raised by that question, this work will only discuss results on DNA methylation measured on whole blood samples.
After this initial excursus into epigenetics, go back to environmental epidemiology. As stated, changes in global methylation [Zhu et al., 2010] as well as in candidate genes [Bind et al., 2014] were observed in individuals with high occupational exposure such as police men or in response to variation in ambient air pollution concentrations [Fustinoni et al., 2012]. What remains still unclear is the time after exposure at which changes in DNA methylation are occurring. Another recent study by the same group demonstrated DNA methylation fluctuations in a short-term period (hours or few days) of elevated particle concentrations [Baccarelli et al., 2009]. Genome-wide methylation assays allow taking advantage of advances in biological technologies in epidemiological studies [Christensen and Marsit, 2012] and studying in particular the role of ambient fine particle concentrations in the days and weeks before biosample collection. Previously unstudied regions of the epigenome are now accessible, opening a new way to discover associations that may link air pollution and DNA methylation. The study objective is to identify and investigate CpG sites which DNA methylation, measured through a genome-wide screening, is associated with short- and mid-$PM_{2.5}$ ambient exposure, linking them to novel biological pathways.

1.2 Socio-economic Status and Air Pollution

Moving now to Part II, the focus is on the association between SES factors and air pollution exposure. As said, epidemiologists unveiled the link between increased air pollution exposure and adverse health effects including respiratory and cardiovascular diseases and also cancer [Brook et al., 2010, Ruckerl et al., 2011]. Deepening origin and causes of these associations, many risk factors have been proven to play a role in the environment-health path [Dai et al., 2014, Dubowsky et al., 2006, McConnell et al., 2015, Ponticiello et al., 2015]. There are two key issues that required researchers' effort in environmental epidemiology: causality and confounding. Literature on both fields is quite large and epidemiologists all over the world contributed to the discussion. Up to date several techniques have been developed to account for both issues, leading to unbiased estimates and to provide a correct interpretation of the results. Even if in this section we will mostly touch the issue of confounding, it is worth to clarify how it also interplays with causality and to clarify the rationale of our work. Regression models are built to test the hypothesis of a plausible association between one or more independent variables with one (or more) dependent variables. Sometimes it is clear from third or previous information that a casual direction can be assumed, for example, it makes sense to assume that is smoking that may lead to adverse health effects, while it doesn't make much sense to assume that people with respiratory disorders are more prone to starting smoking. In this example is the temporal dimension that helps to enlighten the relationship: if a person smokes at time A and health issues may occur at time B, which must follow A, they might be a consequence of the event at time A which stressed the corporal equilibrium. Extending this concept to a population-based study and applying the correct statistical tools, we are induced to confirm the existence of a causal association that goes from smoking to health issues and not the opposite. But this kind of information is not always available. Now, going back to SES and air pollution, which one comes first? What is known is that for sure air pollution doesn't, while we cannot exclude the other direction. That opens to another question, are SES factors influencing air pollution exposure? Or, in other words, are they causing a raise in air pollution exposure? An obvious answer would be clearly negative. How is it possible that the social status, the flow of money in bank accounts and educational studies influence the amount of particles that the lungs are inhaling? Surely not directly, but indirectly? The spatial variability of long-term air pollution exposure in large areas of urbanized districts is related to the uneven distribution of social deprivation of the neighborhoods and, up to date, some studies demonstrated that lower SES levels are associated with increased air pollution exposure that may lead to higher mortality and increased hospitalizations [Gray et al., 2013, Yap et al., 2013]. Hence, specially, in recent years, a particular interest has been directed to the impact of socio-economic status factors with the result of classifying them as "modification effect" [Blanco-Becerra et al., 2014, Forastiere et al., 2007, Ou et al., 2008]. Up to date, SES factors have been considered mostly as confounders and used as adjustment factors in epidemiological models, in order to reduce bias. However, how SES and air pollution exposure are associated has not been fully described. We can hypothesize the existence of a third unmeasured variable, probably house location, that may play two different roles: 1) be linked to both SES factors and air pollution exposure and this case is called confounding; 2) alter the association SES factors and air pollution and it would be moderation or interaction [MacKinnon et al., 2007]. Despite both scenarios look plausible, our preliminary results didn't confirm any of the hypotheses. This led us to think a third alternative: whether the association between SES factors and air pollution exposure might be not only driven but emphasized by accounting for household density. The statistical technique which purpose is to answer that aim is called mediation analysis. Given two variables and their association, it has the power to separate the part of the association that represents how much of it is mediated by a third factor (the mediator), which is in a causal sequence between the two variables. The idea behind it comes from the fact that the location of the house is surely influenced by the individual income, but there are other factors (such as proximity to job place/schools/subway stations, children, personal history) that may influence the choice as well. Therefore, the aim of this study is to identify and quantify factors that are influencing the effect. Simply adjusting by the residential area, the association between SES and pollutants may result as masqueraded, therefore a mediation scheme, including the household density as the mediator, can help to better separate the sources of the variability of the link.

Path analysis has been selected as the statistical technique able to differentiate the origin of the effect. We provide results for nitrogen dioxide (NO_2) and particulate matter smaller than 2.5 and 10 μm ($PM_{2.5}$ and PM_{10}). We present here analyses of the pathways potentially mediating the association between SES factors on area-based as well as individual levels and spatially modelled annual air pollution concentrations. Our analyses compare data collected as part of the ESCAPE Study (European Study of Cohorts for Air Pollution Effects, www.escapeproject.eu) in three European cities, Helsinki, Finland, Augsburg, Germany and Rome, Italy [Stafoggia et al., 2014].

Chapter 2

Objectives of the Dissertation

2.1 Epigenomics

Regarding epigenetics, objective of this dissertation is the increment the information about air pollution influence on DNA methylation aberrant changes. This is not only relevant in identifying CpG sites involved in so far undiscovered biological pathways, but also deepening the knowledge regarding the behavior of untouched areas of the methylome at different time point. The identification of extra- and non-genetic processes may increase the awareness regarding diseases' characteristic and their risk estimation.

2.2 Socio-economic Status and Air Pollution

On the other side, we give importance to the understanding of the function of socio-economic status factors involved in the level of air pollution exposure. Objective of this study is to clarify the role of these factors and increase the awareness of their influence. By differentiating both area based and individual level we also aim to add significant pieces to the discussion regarding their differences in air pollution models.

Chapter 3

The KORA Study

KORA stands for the German acronym: Kooperative Gesundheitsforschung in der Region Augsburg, which means Cooperative Health Research in the Region Augsburg. The platform KORA was initiated in 1996 continuing and expanding the research started in 1984 by the international WHO MONICA project [Holle et al., 2005, Wichmann et al., 2005]. In addition to the three independent cross-sectional surveys conducted between 1984/85 and 1994/95 (called S1, S2 and S3), KORA began a fourth study in 1999/2001, with the name S4. Merging data and samples from existing and new studies with the possibility of long-term follow-ups, the aim of the KORA platform is to provide relevant knowledge to the fields of epidemiology, health economics and public health research. The region of Augsburg is situated in the South of Germany and it counts for around 600,000 inhabitants and 430,000 of them belong to the 24-75 years age range. At every survey, information regarding sociodemographic variables, risk factors such as smoking, alcohol consumption and physical activity, medical and family history of chronical diseases and medication use were collected. Additionally, the subjects underwent a standardized medical examination in order to collect blood samples and other anthropometric measurements. All KORA studies have been approved by the Ethics Commitee of Bavarian Medical Association and the Bavarian commissioner for data protection and privacy.

Main epidemiological area of research of the KORA study is cardiovascular disease. That also expands to associated diseases (diabetes, metabolic syndrome) and potential pathological mechanisms (such as inflammation, stress, endothelial dysfunction, etc. . .). Moreover, new topics that have been included in KORA have been OMICS integration, psycho-social risk factors and environmental factors. In recent year, with the advent of genetics, a great effort was also pushed in that direction with the creation of the KORA-gen resource. Aim of the KORA-gen is to contribute to the field genetic epidemiology. The focus is dual, on one side identification of relevant genes in complex disease and on the other side the study of gene-environment interaction approach. In complex diseases, it is more likely that the combination of the two factors may result in an exacerbation of the genes' effect on disease.

Several sub-studies with specific aims (like the KORA-B "MI family study" in 1996/97 on myocardial infarction and the KORA-A "Diabetes study" on diabetes) were also conducted as well as follow-up studies. In 1987/88 was arranged the follow-up of S1, in 2004/05 F3 re-examined S3 and in 2006/08, F4 re-examined S4. Despite cardiovascular diseases kept to be the main area of research, interest on other topics also increased. A special attention was also given to diabetes, myocardial infarction, allergies, asthma, aging, air pollution and other risk factors.

The Environmental Assessment

This short paragraph will describe the method that had been used to assess the environmental measurements. A paramount feature in Environmental Epidemiology is how can environmental measurements be assessed. Improving the technologies and the details of the outcomes, novel techniques must point to an increase in preciseness that often require deep preliminary analysis and expertise in the field. The area interested by the KORA study is the region of Augsburg in south Germany and an extended program of measurement stations have been undertaken. For the two different type of exposure (as it will be better explained later), two different approaches and data sources have been considered according to the problem that needed to be addressed. Despite these differences, the measurements have been performed with the same tool, the Tapered Element Oscillating Microbalance (TEOM model 1400A device Rupprecht and Patashnick). Here the two approaches.

For the short-term exposure it has only been considered one monitoring station positioned approximately 1km South-east of the city of Augsburg, in the scientific campus of the University of Augsburg. Particle concentration has been measured on an hourly basis and daily averages have been evaluated when at least two measurements were considered valid. The picture for the long-term exposure is, instead, quite more complicated. The main focus is to assess personal yearly averages, concentrating more on the spatial variability inducing the study to locate 20 monitoring stations. This approach has been developed, discussed, and commented within the framework of the ESCAPE project across all the participant research centers (http://www.escapeproject.eu/). The first step is the to geocode the participant addresses and the location of the monitoring stations via WiGeoGis and to consequently calculate the different buffers (125m, 250m, 500m, 1000m, 5000m) for each address position. The next step requires the estimation of a number of variables that need to be considered for every buffer from four different macro-areas: 1) land-use: residential land, industry, forested/green areas; 2) demographic: population and density; 3) geophysical: altitude; 4) traffic: intensity, distance to the nearest road, road length, load, major road. Per each monitor, the annual average concentrations have been calculated (20 values) based on three times 2-weeks measurements (in cold, mild and warm season) and regression models on observed concentrations for each pollutant have been developed. Models have been adjusted based with the routine since the two weeks were not matching across the stations. As further step, Land-use Regression models (LUR) have been applied to cohort addresses and the final model from the 20-observation dataset in order to obtain the final references on which apply participant personal values for all the estimated variables, selected at the previous step and finally seize the personal exposure values. Last operations would be to trim the extreme values of predictors to minimum and maximum observed values (at monitoring level) and back-extrapolate concentrations based on routine measurements. Results of the LUR models were published per each pollutant [Beelen et al., 2013, Cyrys et al., 2012, Eeftens et al., 2012].

Part I

Epigenomics

Chapter 4

What is Epigenetics

The word "epigenetics" is already very fascinating from the etymology. By merging the Greek term "epi" with "genetics", it is intended something that is working over genetics, on top of genetics, above genetics. And genetics, on turn, has already been fascinating for thousands of researchers all over the world from different areas (biologists, physicians, epidemiologists, etc...). One of the most illuminating examples that clarifies this definition is provided by Prof. Andrea Baccarelli. He introduces genetics as the score of an opera play: since its composition, it doesn't change any more. However, the outcome is almost never the same for several ambient conditions including different actors, different orchestras and different theaters. And then there is also the director's interpretation of the opera, who, without changing the score, is adjusting it with some musical rearrangement according to his/her taste or interpretation. Conclusion: keeping the same score, its final exhibition, or, translating it into scientific terms, the "phenotype" of the play, differs. Thanks to this brilliant metaphor, we can translate all these musical rearrangements in science as the epigenetic markers that influence the final phenotype. A bold definition of epigenetics might be: "it studies how we change without changing". While our genome doesn't change, the phenotype does change and epigenetics is the field that studies all the biological and physiological mechanisms which are interplaying and completing genetics in phenotype determination (including aging and disease risk). According to that, epigenetics emerges as a multi-dimensional phenomenon. Laird in 2005, discussing cancer epigenetics, clarified how the different mechanisms are interacting in order to reach phenotypic variations, in the state of chromatin structure, through histone modification, associated protein composition, transcriptional activity and DNA methylation [Laird, 2005]. DNA methylation turned to be a useful marker and the development of new technologies allowed the researchers to obtain accurate measurements in recent years. Not only, already in 2005, Laird (and probably many others with him) was fascinated by the idea of a genome-wide analysis of DNA methylation, and the possibility to uncover so far unveiled areas of the human genome. As said, epigenetics is a multidimensional phenomenon and it is not only touching exogenous processes but is also associated with environmental factors, increasing the complexity (and the interest) on the field. A first epigenetic impact happens already during the prenatal state of life. It has been demonstrated that an increased risk of cardiovascular disease (CVD) at grown-up stage is associated with prenatal exposure to tobacco and this effect may be (at least partially) mediated by epigenetics alterations. Several hypothesis and novel discoveries have already been revealed but knowledge regarding the underlying factors under the observed long latency period that elapses between the in utero tobacco exposure and the CVD development later in life is still not very expanded [Anon., 1994, Breton et al., 2009]. In addition, endogenous environmental factors have also been observed in grown-up stages of life as connected with diseases through epigenetic changes. In this direction, three independent studies observed that high traffic exposure and particle concentrations, well-known environmental elements associated with CVD risk, affected DNA methylation [Baccarelli et al., 2009, Tarantini et al., 2009, Yauk et al., 2008]. Despite DNA methylation is still a mono-dimensional marker for epigenetics, its revealed link with the environment increased the focus on this biological process. This scenario opens to a large pack of new questions. First of all, which environmental elements are influencing DNA methylation? Secondly, how long does it take to a variation in DNA methylation to occur? At this point a step-back is required to clarify what a variation in DNA methylation is. As defined, DNA methylation represents the addition of a methyl group at the 5-carbon of the cytosine ring giving to a specific locus the state of methylated or non-methylated. Methylation is measured on thousands of cells in order to obtain a percentage: for each CpG site, it is estimated the proportion of methylated and non-methylated cells. Loci with low proportions (approximately below 0.30) are called hypomethylated and loci with high percentages (approximately above 0.70) hypermethylated. DNA methylation depends on a group of three enzymes which are in charge of adding the methyl group and they are known as *DNMT1*, *DNMT3a* and *DNMT3b*. DNA methylation variations can also go in the other direction and a reduction of the percentage happens when methyl groups are removed from CpG sites and this process is called DNA demethylation. However, when we refer to DNA methylation level we are always describing the estimated percentage of cells where the methyl group is present at the moment of the medical examination for a specific CG dinucleotide. In environmental epidemiology this concept has to be clearly stated in order to avoid ambiguous and misleading interpretations. For example, observing a positive significant association between DNA methylation in a specific CpG site and air pollution exposure allows us to think that an increase in air pollution exposure is associated with an increase in DNA methylation. But increase from what? Is there any starting point? The association must be interpreted "according to" rather than "independently from" the exposure since a theoretical hypothetical DNA methylation value does not exist. An epidemiological conclusion might then sound like: a systematic higher value of DNA methylation was observed in a group of high exposed subjects vs low exposed.

Literature in the last two decades regarding epigenetics underwent a consistent raise, its impact on modern medicine is huge. The number of publications on epigenetics passed from around 400 in 1995 to almost 9000 in 2009. Of them, according to Pubmed, in 1995, only 12 involved cardiovascular diseases, while in 2009 they were almost 500 [Baccarelli et al., 2010]. Focused on previously hypothesized areas of the genome (mostly promoters, repeated elements and CpG islands), gene-specific results showed, with a certain degree of consistence, evidence of associations between epigenetic markers and increased environmental exposure. Aim of the genome-wide approach is to expand the knowledge already assessed and try to involve also other areas of the human DNA. Novel insights are needed to better enlighten the biological processes that happen behind diseases, behind aging, behind hereditary traits and their further implications.

But new technologies carry not only new questions but also new issues. First, at the planning stage it's important to take into account that dealing with genome-wide data needs an increase in power and sample size. Conduct a study with at least a few hundreds of participants in order not to end up underpowered is necessary. A second clutch point is then to find replication studies. In this work we'll try to solve this problem in two different ways: when at least three studies are involved it is reasonable to run a random-effect meta-analysis while for only two studies it would be enough to replicate the magnitude of the significance level for the CpGs identified in the first study. Despite finding other cohorts with similar data at both biological/medical and environmental level is rather hard, replicate the results in another independent population not only strengthens the evidence and helps to avoid false positives, but also increases generalizability. Further issue regarding the data quality, the normalization and the data preparation will be later discussed.

The Illumina Infinium 450k Beadchip

In May 2010 the company Illumina released what researchers in epigenetics had already been waited for years: the most extended DNA methylation measurement device, the Illumina Infinium 450k Beadchip [Sandoval et al., 2011]. It has the power of measure DNA methylation level in 485,764 loci, of which the 99.3% (482,421) are CpG dinucleotides and 3,343 (0.7%) are CNG targets. It is circa 20-times larger than the previous Illumina Infinium 27k Beadchip that accounted for approximately 27,000 CpG sites. Under the functional genomic standpoint, around 40% of the annotated CpGs (200,339) are located in proximity of promoters of genes, of which 62,625 are within 200 bp and 77,375 within 1500 bp upstream the starting transcription site, 49,525 are in the 5'untranslated region and 10,810 are in Exon 1. A 31% of the CpGs are then annotated in the body of the genes, a 25% in intergenic regions and a last 3% in 3'UTR. Relation to CpG Island is also an important feature. They were firstly described in 1985 and then have been found as highly sensible areas in the methylome, bestowing a special focus to the CpGs belonging to these highly-CG-dense regions [Gardiner-Garden and Frommer, 1987]. 31% of the annotated CpGs (150,254) are then found in CpG Islands, 23% (112,072) in CpG shores (North either South), 31% (47,161) in CpG Shelves (either North or South) and the rest 36% (176,127) are annotated in other areas or the so called "Open Sea". A vast majority is also associated with RNA coding areas (74%, 361,766), around 25% (119,830) are associated with intergenic regions and the remaining 1% (4,168) is annotated into non-coding regions. All the 22 autosomal chromosomes, as well as the two sexual chromosomes are touched by the chip. Chromosome 1 is the most represented with 46,867 CpG sites involved and the autosomal chromosome with the least amount of CpGs is number 21 with 4,246. This was a short overview of the coverage of the chip, let's look at how it is working. We'll refer to the DNA methylation measurements as β-values that represent the mean value for a specific CG dinucleotide and a delta β-values, meaning the difference in DNA methylation between the control and the experimental group. For a generic i$_{th}$ CpG site, this is the formula:

$$\beta = \frac{Max(y_{i,meth}, 0)}{Max(y_{i,meth}, 0) \ + \ Max(y_{i,unmeth}, 0) \ + \ \alpha} \tag{4.1}$$

where y$_{i,meth}$ and y$_{i,unmeth}$ for the CpG site i are the intensities measured by the methylated and un-methylated probes, respectively and α is an offset constant (by default, $\alpha = 100$) added to the denominator to regularize β value in case the intensities of both methylated and un-methylated probes are low. The spectrum of values for the β-value statistic is between 0 and 1, or, in percentage, 0 and 100%.

The greatest difference from the previous Illumina 27k chip lays in the two different types of assay that have been used to measure the DNA methylation level, the so-called Infinium I and Infinium II. Infinium I measures around one third of the CpGs and Infinium II the other two thirds, however, this proportion is uneven in representing other regions of the genome (e.g.: in CpG islands, 50% of loci are measured with Infinium I). While Infinium I uses a classical dual-probe well-established approach, is the Infinium II that brings out a novel technology. Severe repercussions appear as a consequence of this separation on the detection of differentially methylated regions and are well elucidated by [Dedeurwaerder et al., 2011]. Whereas Infinium I leans on two different probes, located on two different bead types with the purpose of separate the measure for the methylated and the unmethylated allele, Infinium II uses only one type of probe on a unique bead type. For the second, is the dye of the channel that indicates whether the signal is methylated (green) or unmethylated (red). Result of this hybrid approach is a difference in the distribution of the measured DNA methylation across the two probe types that are highlighted by three points. First, there is a difference in the range of the distribution and the Infinium II lacks in recognizing highly methylated CpG sites coming up short at the upper bound. Secondly and partially related to the first point, the DNA methylation distribution across the CpG sites is acknowledged to be two-peaked. Empirical results show a shift of the distance between the two peaks comparing Infinium I and II distribution, and, specifically, in Infinium II the summits appear closer than in the Infinium I (also considering the lowest range that this bead type covers, as said in the point one). And, thirdly, the probe-wise variance of DNA methylation resulted to be higher for CpG sites measured with Infinium II.

Several groups of researchers started to look at this problem and in the years following the release of the chip, new methods had been developed and reviewed in order to adjust this shift and obtain the highest possible consistency across the two Infinium bead types.

Chapter 5

Preprocessing of the Genome-wide Methylation Data

Objective of this chapter is to clarify the preprocessing approached that was used in this work after a short review of the available methods.

5.1 Quality Control

The first step of the preprocessing of the data is the Quality Control. During microarray experiments, technical issues may occur, resulting in a poor performance of the chip; therefore it is paramount to detect all possible sources of dirtiness in the data, find a way to purify them and to discard data that are not needed or are not necessary. The first step is to retain DNA methylation regarding only autosomal chromosomes and exclude rs-probes (65 loci), markers that are not CpG sites but rather SNPs (Single Nucleotide Polymorphisms). The second step was to remove background noise and it was accomplished using the R package *minfi*. Afterwards it was necessary a bead filtering. Briefly, within the same array is contained a random number of technical replicates, the so-called "beads", for every probe. Their scope is to open unique opportunity of quality control of the data. According to the manufacturer's recommendation, to be valid, at least three functional beads on the array have to summarize either the methylated or unmethylated signals that are associated to a probe. Rather than eliminate the probes that do not fit with this criteria, a detection p-value of 1 is associated to the β-value for these probes. Detection p-values represent the confidence that a given transcript is expressed above the background defined by negative control probes and the score whether a transcript on the array has been detected. A non-acceptable (NA) value is also assigned to probes with a detection p-value higher than 0.01. At this point, most of the deceptive information should have been detected, and the second last measure that needs to be accounted is excluding the CpG sites with more than 5% of NAs value and samples with more than 20% of NAs values. When a probe showed a high degree of non-valid information then it is likely that the all data regarding that probe are not reliable. Finally, a dye bias correction was applied according to the R package *lumi*. Considering the Quality Control completed, we can move on to one of the most controversial topic: the normalization process.

5.2 The Normalization Process

For the first two/three years after the release of the chip, the problem of normalization opened a huge discussion across scientists all over the world such that a gold standard method has not yet been found. Despite several reviews and revision articles, which clarified strengths and weaknesses of the different proposed pipelines and ranked their performance including also mixture of different proposed methods, the final decision is not univocal. Here we provide a short history of the pipelines that received a larger echo. An historical overview is provided in Figure 5.1. The first method was proposed by Dedeurwaeder et al. together with the description of the two different types of bead probe [Dedeurwaerder et al., 2011]. They proposed a simple normalization process (logarithmical rescaling of DNA methylation proportions: $M\text{-}values = \log_2(\beta\text{-}values/(1-\beta\text{-}values))$) and a peak-based correction that works as follows. Through kernel density estimation, the methylated and unmethylated peaks for both Infinium I and II have been determined.

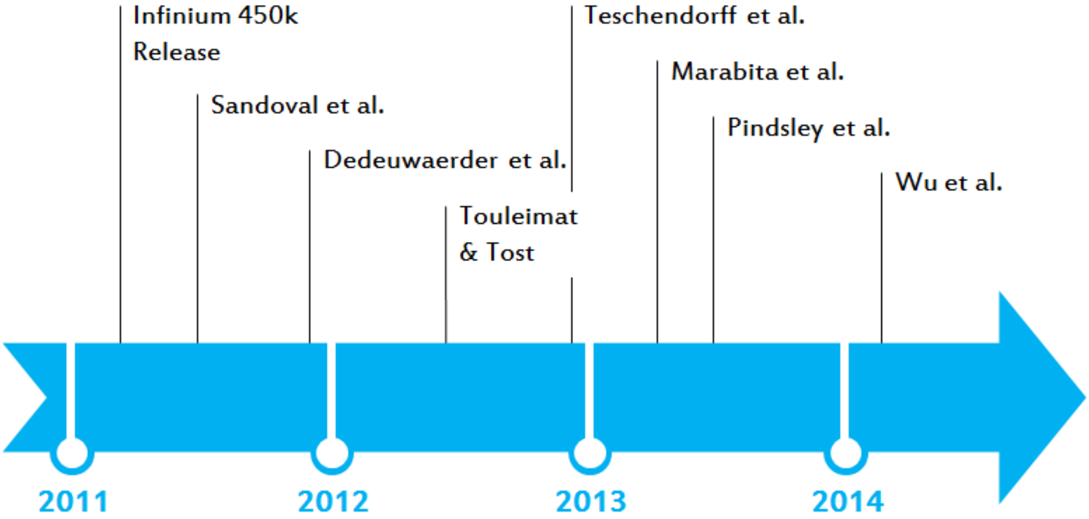

Figure 5.1: A few milestone points in Illumina 450k data evolution. From the release, in spring 2011, several pipelines have been developed, from the Peak-based by Dedeuwaerder (late 2011), to the SQN from Touleimat and Tost (mid-2012), to BMIQ from Teschendorff (late 2012) before other studies were undertaken in order to establish the most reliable method.

Prior transformation to M-values allows to separate the distribution in order to have the unmethylated summit in the negative side and the methylated in the positive side of the new [-Inf; +Inf] range. The function *Argmax(density M-value)* has been used on both sides to determine univocally the two summits S_U and S_M for the unmethylated and the methylated side, respectively. The next step would then be to rescale independently the negative and positive M-values based on the distance between the summits and zero. The corrected M-values follow these equations for the negative side: *corrected M-values = M-values/*σ_u where $\sigma_u = 0$ - S_U and the positive side: *corrected M-values = M-values/*σ_m with $\sigma_m =$ S_M - 0. Lastly, the M-values were rescaled to match Infinium I distribution range and re-converted back into β-values.
It was then succeeded by a pipeline developed by the researchers Touleimat and Tost in 2012 [Touleimat and Tost, 2012]. They addressed all the issues there had been observed before and provided a panel of plausible solutions that have also been tested through simulations. Main novelty of the Touleimat-Tost approach is the Subset Quantile Normalization (SQN). They are aware of the problems and the differences due to the hybrid approach used from Illumina and they proposed a method that uses, after quality control, the more reliable estimation from Infinium I to normalize and correct the DNA methylation levels measured with Infinium II. This approach, that applies the concept of "anchor probes", is slightly different than previously described approaches since it doesn't modify the distribution based on value equivalence but on rank equivalence. They further applied this method taking into account the unbalanced distribution of Infinium I and II in the different regions of the genome. So, using the information provided by Illumina regarding "relation to CpG Island" and "relation to gene sequence", they computed SQN stratifying by the category of the two mentioned variables. Then, they compared different methods including no preprocessing, the Dedeurwaerder peak-based approach, a global SQN, a CpG Island stratified SQN and a gene sequence stratified SQN. The results were finally compared with pyrosequencing data, as best experimental expression of methylation measurements, and underlined the SQN approach used in relation with CpG Island annotation as the best solution.
However, a new revolution was about to come. During late 2012, Teschendorff et al. published a new method called Beta-Mixture Quantile Normalization, that we'll refer as BMIQ [Teschendorff et al., 2013]. The basic idea of this method is not very complicated but touches a rather problematic point of previous pipelines: global quantile normalization might force a specific DNA methylation value into a quantile too far away from its original position without any possibility of control. This intuition led to the idea of stratify the distribution in three sections and parametrize them. These are the stages. After the Quality Control step, this method considers the fitting of a three-state beta mixture model in order to separate the three biological statuses of the methylation distribution: unmethylated, hemimethylated and fully methylated. This step will be performed on separated Infinium

I and II probes. Two parameters of a beta distribution are evaluated for the Infinium I and II probes for each of the three identified set of CpGs (un-, hemi- or fully methylated). What happens next is recreating a distribution of the un- and fully methylated status using an Expectation Maximization (EM) algorithm with a function that denotes the probability of belonging to a specific state on both left and right sides of the mean. This separation is important since the EM algorithm estimates are two-tailed. Turning, then, these probabilities into the quantiles of the beta-distribution using the Type I parameters, the normalized β-values for the probe I are set. Finally, the hemimethylated values are remaining and an empirical approach can be applied. Firstly setting the robust bounds different than 0 and 1 but maximum of unmethylated or minimum of hemimethylated and maximum of hemimethylated or minimum of fully methylated and then applying the shift and dilatation factors would let the normalized values to be computed. There is no uniform re-scaling but instead a probe-specific transformation, allowing the exclusion of holes in the distribution. Even if in the same paper they tested the validity of their approach compared with the other methods previously described, a more detailed and extended evaluation was provided by Marabita et al. a few months later, in spring 2013 [Marabita et al., 2013]. Using two independent datasets that ensured high level of both biological and technical replications, following a strict design, they focused on four issues to validate the methods. The four criteria were:

1. how is technical variability reduced during normalization;
2. how much probe design bias is removed;
3. how is batch effect reduced;
4. how does it help to identify differentially methylated regions.

The conclusion is that in general BMIQ was the optimal method to provide a good assessment and reliable methylation values. A valid alternative is also represented by the combination of quantile normalization and BMIQ.
Following Teschendorff's criteria we tested all these methods in our data and decided to choose BMIQ without QN. To be thorough, other methods were also developed and considered by Marabita et al. like SWAN (Subset-quantile Within Array Normalization) of the GenomeStudio Quantile Normalization. But these methods were not taken into account in this study. Figure 5.2 shows a summary of the steps applied to our data.

5.3 The Role of White Blood Cells

Hematopoiesis is the name of the biological process that gives birth to all the range of blood cells subclasses from pluripotent hematopoietic stem cells. Leukocytes, commonly

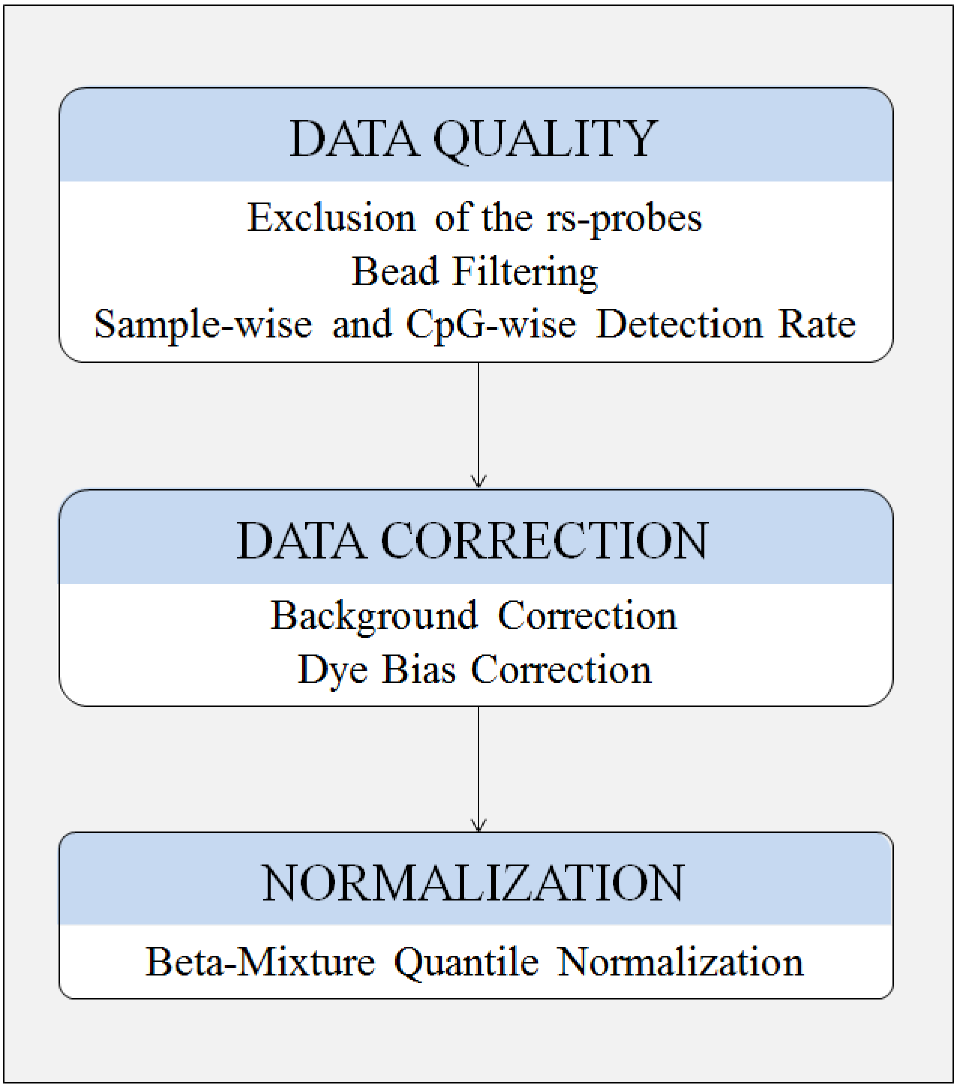
DATA QUALITY
Exclusion of the rs-probes
Bead Filtering
Sample-wise and CpG-wise Detection Rate
DATA CORRECTION
Background Correction
Dye Bias Correction
NORMALIZATION
Beta-Mixture Quantile Normalization

Figure 5.2: Performed steps, by thematic area, on KORA data.

called white blood cells (WBC), play an important role in response to pathogens and foreign antigens. Therefore, the composition of the different leukocytes type has been found to reflect occurrences of disease states or exposure to toxicants leading to alterations in whole types of cells in or out tissues [Wieczorek et al., 2009]. Moreover, white blood cells composition strongly differs in DNA methylation levels of their respective CpG sites. As a consequence, thanks to this dual implication of leukocytes, they arise as strong confounding factor in environmental epigenetics. The composition of white blood cells is related to phenotype or disease and DNA methylation, therefore biased estimates and spurious correlations may be observed when excluded from the analysis. This is the reason why in 2012 a group headed by Houseman E.A. proposed a new method, using the concept of regression calibration, with the aim of estimate at the same time from DNA methylation data the personal proportion of six different white blood cell types: B cells, Granulocytes, Monocytes, Natural Killer cells, CD4+ and CD8+ T cells [Houseman et al., 2012]. The group of researchers identified a number of differentially methylated regions that can be used as reliable and steady biomarkers for individual leukocyte types and thanks this property, proposed a set of analytical tools able to reach reliable estimates of white blood cell proportions from whole blood samples. R code and methylation "purified" cell samples (representing the external validation dataset or gold-standard data) have been made freely accessible to all the users for easy implementation. Alternative gold-standard data have been proposed by Reinius et al. [Reinius et al., 2012]. An advantage of this alternative dataset is that most cell type specific CpG sites were selected from the Infinium 450k, whereas the previous Infinium 27k was used by Houseman et al. Furthermore, Reinius et al. also differentiated between Eosinophils and Neutrophils that in Houseman are both ranked as Granulocytes and the purified cells were obtained from the same six subjects differently than Houseman. However, data used by Reinius consider only male samples, which implications in a mixed population are difficult to assess. This drawback let us decide for the implementation of the Houseman data, being in possess of a mixed sex population.

Chapter 6

Statistical Analyses

After describing the data and their preprocessing, it is time now to explain how an epigenetic genome-wide analysis is performed. First of all it takes the name EWA, from epigenome-wide analysis. As discussed above, data are obtained from more than 400,000 CG dinucleotides for all the study participants. For each subject we then collected its corresponding exposure values or outcome of interest and all the other useful information we would like to use in order to purify as much as we can our analysis. The way to proceed then is very straightforward, by running a model for each CpG site, reaching the remarkable amount of results from more than 400,000 models. Useful information are tracked for each of them (usually β coefficient, Standard Error and p-value of the exposure or of the CpG site whether it is not the outcome) and results are then evaluated and interpreted. This is the basic rationale behind every EWA: study the association of each CpG site with the exposure or outcome of interest. Methylation can be either the dependent or the independent variable, conditional to the design of the study, if it is considered to suffer effects from other variables or if it supposed to cause changes to other factors. Aim of the EWA: identify novel CpG sites associated with the variable of interest and look for plausible biological pathways that might have elicited physiological reactions leading to diseases.

6.1 Identification of the Model

EWAs require heavy computational calculations. For this reason an accurate preparation is necessary, not only regarding the organization of the data processing but also to considering previous findings. We selected our model based on prior knowledge and we then tested our results in order to evaluate the robustness of the model and to identify plausible problematic sources. Starting now with the a priori model, the sections below will describe some applications and how they had been tested. We obtained data following a cross-sectional scheme with single individual measurements (as explained in Chapter 3 – The KORA Study), thus a linear regression model would be fitted in order to study the linear association between the exposure and DNA methylation. Repeated over the time measurements would require a random-effect model. In environmental epidemiology, the identification of the model has to be accurately planned and the first element to take into account is the type of exposure we are studying. All analyses have been performed with the statistical software R, version 2.14. Some applications are described in the following paragraphs.

6.1.1 Short- and Mid-term Model

The first exposure that will be discussed is the short- and mid-term $PM_{2.5}$ and following is described the way the model was build. A crude analysis of any methylation data always require age and sex, mandatorily. Next, an important confounding factor that have been observed to modify the methylation values are smoking status as categorical (stratified as current, former and never smoker) and Body Mass Index (BMI) and alcohol consumption (number of drinks per week) as continuous. Smoking-related CpG sites are particularly susceptible to changes in methylation and some preliminary analysis already demonstrated that unadjusted models lead to spurious false positive results. In line, we had to take into account the social and environmental factors that epidemiologist found as possible effect-modifications in air pollution studies: socio-economic status (personal income), day of the week, season (according to the astronomical definition) and temperature (daily averages). These variables are particularly important in short- and mid-term exposure where our focus is two-sided: 1) to study short term variation in DNA methylation related to increased air pollution exposure in a limited time window and 2) to focus on the temporal side of the variability rather than the spatial one. Complete coverage regarding the variables in the involved studies is provided in Table 6.1. Control for temperature and season allow us to purify from possible effects due to the influence of the climate and the weather conditions. Finally, as stated in the previous chapter, white blood cell proportions, estimated via the method developed by Houseman et al., have been included in the model. Here is the equation:

$$Y_i = \beta_0 + \beta_1 PM_{2.5_i} + \beta_2 Temperature_i + \beta_3 X_{3,i} + ... + \beta_p X_{p,i} + \varepsilon_i \tag{6.1}$$

Where Y_i is the methylation measurement for i_{th} subject, β_0 is the intercept, β_1 and β_2 are the coefficients for the effect of the trailing average values for exposure and temperature at a specific time window on methylation, $X_{3,i}$ to $X_{p,i}$ are the $p-2$ covariates with the corresponding $\beta_3, \dots, \beta_p$ coefficient and ε_i is the error.

As sensitivity analysis we then tested two different partially adjusted models. The first one is a crude model, adjusting only for age and sex, while the second one added to the crude model also the white blood cell proportions. A special focus on the white blood cells is important in order to keep their effect under control. The question is: which of

	Mean ± SD / N(%)		
	KORA F3 **(n=500, 2004-05)**	**KORA F4** **(n=1799, 2006-08)**	**NAS baseline[a]** **(n=657)**
Participants Characteristics			
Males	260 (52.0)	887 (49.3)	657 (100)
Age, years	53.12 ± 9.6	60.92 ± 8.9	72.44 ±m 6.9
BMI,[b] kg/cm^2	27.15 ± 4.5	28.15 ± 4.8	28.07 ± 4.1
Monthly Income, euro	1,104.8 ± 583.9	1,159.84 ± 556.6	*
Education, years	11.7 ± 2.8	11.5 ± 2.5	15.07 ± 2.9
Drinkers[c]	296 (59.2)	1038 (57.7)	130 (19.7)
Alcohol Consumption, g/day	16.11 ± 19.6	15.49 ± 20.4	*
Smoking			
Never Smokers	226 (45.2)	226 (12.6)	188 (28.6)
Former Smokers	11 (2.2)	782 (43.5)	446 (67.9)
Current Smokers	232 (46.0)	753 (41.9)	23 (3.5)
Passive Smokers (either Former or Never)	11 (2.2)	36 (2.0)	*
Missing	20 (4.4)	2 (0.0)	0 (0.0)
Environmental Exposure (mean of the daily average of the day before the visit)			
PM$_{2.5}$,[d] μg/m^3 **Percentiles (25th, 50th, 75th)**	20.0 ± 11.6 14.0, 17.7, 25.9	14.2 ± 10.2 6.7, 12.2, 18.8	10.6 ± 7.1 6.3, 9.0, 13.2
Temperature, °C **Percentiles (25th, 50th, 75th)**	7.1 ± 7.5 0.9, 7.9, 13.2	8.7 ± 6.6 3.9, 7.5, 13.1	12.5 ± 8.5 6.4, 12.7, 19.8

[a] First time blood sample was collected (time window: 1999-2007)
[b] Body Mass Index
[c] Participants with at least 2 drinks per week
[d] Particulate Matter smaller than 2.5 μm
* Data not available

Table 6.1: Descriptive statistics of the study participants in the KORA F3, KORA F4 and US Veteran Affairs (VA) Normative Aging Study

these confounders are leading to statistical significance? A comparison of the results across the three models (crude, crude plus white blood cells and full) allows the confirmation of the reasonability of the original model, including the necessary adjustments, and to even exclude spurious implications by other covariates. Furthermore, the issue of long-term exposure also needed consideration. In order to reduce the risk of running into false positive associations between short- and mid-term $PM_{2.5}$ exposure and DNA methylation we also run a sensitivity analysis by including yearly averages. Usual air pollution exposure is known to influence DNA methylation values by taking part of the variability and biasing the results and being in possess of long-term $PM_{2.5}$ data allowed us to check variability due to this specific exposure and strengthen the evidence of the results.

Another important question touches the length of the time window exposure we considered. Based on previous knowledge [Bind et al., 2014, Rückerl et al., 2007, Schwartz, 2000], we looked at three different trailing averages prior the visit day. 2-, 7- and 28-day were selected in order to represent variations within days and weeks (short- and mid-term exposure) from measurement that took place at the same measurement station (in KORA located at the scientific campus of the University). Looking for temporal rather than spatial variability, daily averages from the city center can sufficiently simulate peaks of air pollution that may trigger adverse health outcomes. According to the findings by Bind et al., variations of DNA methylation due to endogenous exposures are possible after an increase of exposure in a 4-week time window. Secondary aim of the three lags is also to catch plausible DNA methylation fluctuations over time. A univocal pattern for most of the CpG sites seems to be an unrealistic conclusion, while is more likely that different loci may be associated at different temporal intervals. Genome-wide results looking at answers to this open question do not yet exist. The results from single day windows led us also to verify cumulative exposure periods (2nd and 3rd weeks before the day visit) defined and evaluated in regression models as secondary analysis.

We applied the same approach also to particle number concentration (the number of particles present in any given volume of air). We averaged the exposure at the three same time-windows as $PM_{2.5}$ and meta-analyzed the results of both exposures among KORA F3, KORA F4 and NAS (Normative Aging Study, that will be later introduced) since the strategy to measure the exposure was consistent across the studies.

6.1.2 Long-term Model

Possible implications of long-term air pollution exposure were also studied. Long-term exposure values have been defined according to LUR models within the framework of the ESCAPE study and have already been explained in Chapter 3. Models for long-term exposure are slightly less complex than models for short-term exposure (partly because many factors have already taken into account in the estimation of the long-term variables) and as covariates account only sex, age, BMI, smoking status, alcohol consumption and educational years. Sensitivity analysis has also been performed with crude model (only age and sex) and an alternatively adjusted model (with income instead of educational years). Pollutant considered were $PM_{2.5}$, PM_{10}, NO_2 as well as traffic exposure. Analyses have been performed by pooling KORA F3 and F4 in order to increase power and were not followed by replication sets. Finally, both BMIQ- and SQN-preprocessing pipelines have been applied to the data and results were evaluated for both.

6.1.3 Lung Function Analysis

Another pioneer project that involved KORA DNA methylation data aims to link smoking, epigenetic changes and lung functions. The conceptual framework Figure 6.1. is the prior identification of the CpG sites annotated within genes associated with smoking and at a later stage the test of their possible implications with lung functions. The rationale is: changes in DNA methylation associated with (and eventually caused by) smoking exposure are then considered as possible enhancer of lung function variations. Lung function decline over the time might be mediated by epigenetics, lying in between the physiological changes and the environmental influence. Genome-wide smoking effect on DNA methylation and their plausible recovery after quitting smoking have already been innovatively described by Zeilinger et al. [Zeilinger et al., 2013]. By including spirometric evaluations of FVC, FEV_1 and $FEF_{25\%-75\%}$ lung functions we have the chance of adding a piece to the puzzle. Again in collaboration with the Harvard School of Public Health and the laboratory of Prof. Andrea Baccarelli, we discussed a shared strategy that best

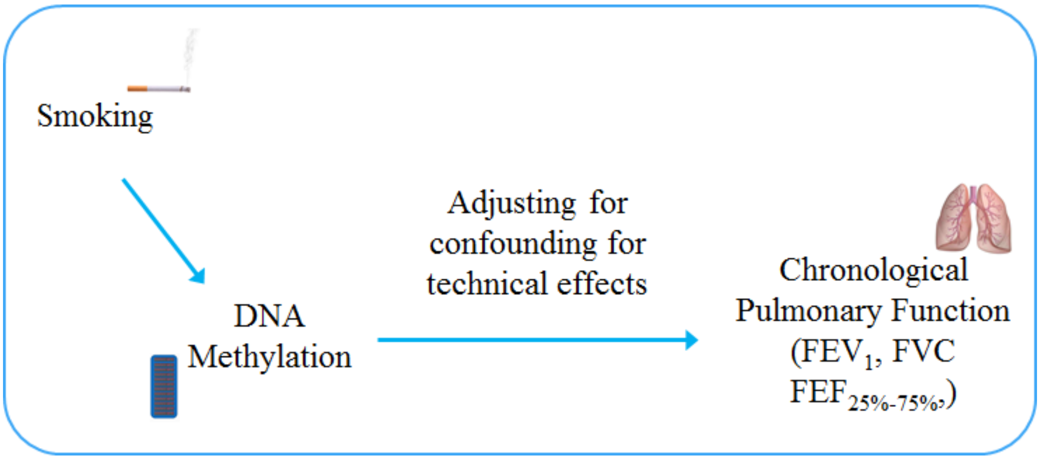
Smoking
DNA Methylation
Adjusting for confounding for technical effects
Chronological Pulmonary Function (FEV_1, FVC $FEF_{25\%-75\%,}$)

Figure 6.1: Conceptual Framework of the lung function project. Confounders: log(height baseline), weight, $weight^2$, pack-years, follow-up time, max education, indicator of taking medicine, Houseman cell proportions, indicator for season, day of week, vitamin C intake and plate.

fits the data and replicate the results in two independent cohorts (in KORA spirometry data are available only for F4). A pilot EWA on NAS data using smoking as exposure variable had the scope of compile a Master List of CpGs for replication in KORA. This list includes all the identified probes, any other CpGs found within the gene body, in addition to any CpGs +/-5kb from the identified target and for the CpGs identified in NAS but not mapped to any gene, all the local CpGs at +/-5kb were also added to the list. The definitive Master List includes 2003 unique CpG sites all over the genome. In 2008 a subset of the KORA F4 called F4L gave birth to a study with the aim of controlling lung functions. Responsible for this project is prof. Holger Schulz who, together with dr. Stephan Karrasch took part of the joined decision group. Lung functions have been consistently measured at different time points (four times in the NAS study and two times in KORA) and have been considered as the outcome variables. We run a linear model for each of the three measurements for each of the selected methylation site like the one that follows:

$$\begin{aligned} Y_{i,j} &= \beta_{0,j} + \beta_{1,j} Age_{baseline,j} + \beta_{2,j} CpG_{i,j,time1} + \beta_{3,j}(Age_{k,j} - Age_{baseline,j}) + \\ &+ \beta_{4,j}(Age_{k,j} - Age_{baseline,j}) \star CpG_{i,j,time1} + \beta_5 X_{5,i} + ... + \beta_p X_{p,i} + b_{i,j} + \varepsilon_{i,t} \end{aligned} \tag{6.2}$$

Where Y$_{i,j}$ represent the lung functions, $\beta_{0,j}$, $\beta_{1,j}$, $\beta_{2,j}$, $\beta_{3,j}$, $\beta_{4,j}$ are intercept and coefficient for age (at baseline), the CpG site, the difference between age at the follow-up points and the baseline, and the interaction between the difference between age at the follow-up points and the baseline and the methylation measurement. $\beta_{5,j}$, ... ,$\beta_{p,j}$ represent the $p - 5$ coefficient for the list of confounders: log(height baseline), weight, squared weight, pack-years, follow-up time, maximum education, indicator of taking medicine, Houseman cell proportions, indicator for season, day of week, vitamin C intake and technical effect adjusting. $b_{i,j}$ represent the random intercept for the individuals. The models have been applied to all folks, and replicated in a set of KORA subsets with only men, only women, and men older than 55 years in order to increase the comparability with NAS data (an only old male cohort). Results are reported for two parameters: β_4, which measures the effect of DNA methylation on the rate of decline in lung function, i.e., the interaction between DNA methylation and follow-up time, and β_2, which represents the cross-sectional effect of DNA methylation. As final sensitivity analysis, trying to confirm that the effect was not driven by the current smokers, the analysis was also run only on former and current. Replication of the results with a subset implies a confirmation of the discovered signals while a non-replication wouldn't directly mean the denial of the observed signals. The reason is simple, by reducing the sample size we reduce the power and it is not possible anymore to define if the signal disappeared due to a lack of power or because it was previously driven by an extra non-necessary adjustment.

6.2 Outliers

Outliers might constitute a serious problem. Preliminary results showed that without excluding them, the risk of increasing the false positive rate is higher. Here's how we faced this issue. Scatterplots of residuals have been plotted and plausible outliers have been manually checked and discarded. Our belief is that unusual values are still plausible and remove them from the analysis would reduce power. The basic idea is to follow the concept of hypo-, hemi- and hypermethylation. Unusual or extreme values for a CpG wouldn't be discarded if they keep being within one of the three mentioned areas (identified with these ranges [0 – 0.35], [0.35 – 0.65] and [0.65 – 1]), so for example if a CpG has an average value of 0.85, displaying hypermethylation, all the records above 0.65 would not be removed. This rule of thumb method gives the possibility of keeping the most reasonable data. Since preliminary results confirmed that is more likely that outliers highlight false positive associations instead of hiding true signals, it makes sense to check for outliers after the analysis and not before. Other techniques based on distance from the median have also been developed, but often they result in being too conservative, by simply drawing a threshold that risk to discard extreme values without a particular biological reason.

6.3 Replication of Findings

As mentioned earlier, an important point in DNA methylation studies is to replicate the results in different cohorts. Thanks to our collaboration with the Environmental Health department at the Harvard School of Public Health, and especially thanks to Professors Joel Schwartz and Andrea Baccarelli, we were able to perform the same analysis on short- and mid-term $PM_{2.5}$ within the framework of the Normative Aging Study (NAS) and merge the results in a meta-analysis. This was also possible thanks to the high consistency of the methods that were used to get the data in the two research unit, the Helmholtz Zentrum München in Munich, Germany and the Harvard School of Public Health in Boston (MA), United States. Before moving to the meta-analysis and the multiple-comparison, a short introduction on the NAS and the methods applied is given below.

In 1963 at the Veteran Affairs (VA) Outpatient Clinic in Boston (MA) was initiated the Normative Aging Study, a comprehensive interdisciplinary longitudinal study [Bell et al., 1972]. Aim of NAS is to study the biomedical, physiological, psychosocial and disease-related changes and effects associated with aging [Bossé et al., 1984]. Thanks to its statutory responsibility for the medical care data of 25 million war veterans (mostly from World War II and the Korean War), the VA could compose the first sample with 2,280 men. They were enrolled as research individuals for their lifetime, be subject to recurrent medical examinations (at 3-5 year interval) on an outpatient basis and supplemented with periodic mail surveys, interviews and examinations. In order to center the attention on non-pathological aging, the subjects were carefully screened prior the visit to meet rigid health requirements regardless of age. Exceptionality of this study is its large sample size and a vast socioeconomic diversity of its population. Focus of the analysis is dual on both the clinical and the social side collecting biological, anthropometrical and medical data as well as socio-behavioral. The design of the study aims to sharpen the association between the natural course of aging and the regular history of chronic diseases. Important element is also given by the environmental implications, which have been included throughout the longitudinal scheme accounting for the endogenous and exogenous character of the aging process.

Data provided for this project belong to the batch of samples collected between 1999 and 2007. Blood samples were provided for most of the 657 participants at two different time points (1,119 total samples). DNA methylation has been measured with the Illumina Infinium 450k Beadchip and particulate matter concentration with the same device as in Augsburg and it was located at the Boston Logan International airport measurement station. Preprocessing strategy was discussed together according to the data and the oncoming literature. The result was a highly comparable pipeline with only minor differences. Only white individuals from the NAS were included in the analysis. Lastly, the model that was used in KORA cannot be perfectly applied in NAS due to the inclusion of replicate measurements from the same subjects. Keeping the same set of covariates, a mixed effect model was required in order to take into account the intra-person variability:

$$Y_{i,t} = \beta_0 + \beta_1 Exposure_{i,t} + \beta_2 X_{2,i,t} + ... + \beta_p X_{p,i,t} + u_{i,j}\varepsilon_{i,t} \qquad (6.3)$$

This model follows the one applied in KORA and differs only in the u$_{i,j}$ element that represents the random participant effect.

6.4 Meta-analysis

Results provided in this work for the short- and mid-term analysis come from a meta-analysis between three independent studies: two surveys F3 and F4 from the KORA study (Augsburg, Germany) and the NAS (Boston, US). KORA F3 counts for 500 subjects enrolled between 2004 and 2005, KORA F4 counts 1799 study participants who underwent the visit in the years 2006-2008 while the NAS includes 657 subjects examined between 1999 and 2007, of which, most underwent the examination twice. We selected random effect meta-analysis in order to control and take into account the heterogeneity. Being yet impossible to extensively expand the analysis to a large number of studies, homogeneity of the estimates slightly increases the reliability of the results. Heterogeneity has been assessed through the I-squared test (I^2) on fixed-effect estimates and CpGs with p-values $>$ 0.05 and I^2 $<$ 0.5 were indicated as homogenous. Methylation changes might be very little but consistency across the studies helps to identify more plausible signals. Being in possess of at least three studies allows the adoption of the meta-analysis in order to summarize the results. But when we only possess two studies, a confirmation of the significant findings is a valuable replication measure, and this is the case of the lung function study. For this project we matched the list of the significant CpG sites at both studies across the three different lung functions.

6.5 Multiple Comparison

As said before, EWAs are so designed in order to test the association between each CpG sites and endogenous exposure as well as other physiological index. But facing thousands of tests implies to face the problem of multiple comparisons: increasing the number of estimated p-values, some of them will fall below the significance threshold just by chance, signaling false positive associations. In order to avoid this, several techniques have been developed and within the most common there are the Bonferroni and the False Discovery Rate (FDR). The Bonferroni method aims to correct the familywise error rate by reducing the threshold by the number of performed tests, for example fixing the significance threshold at 0.05 and running 100 tests, the new adjusted threshold for p-values would be 0.05/100=0.0005. This method has been acknowledged as a very conservative method even leaving the chance to observe false negatives. The FDR instead aims to control for the false discovery rate which represents the proportion of discoveries that classified as false positive. Developed in details by Benjamini & Hochberg in 1995 [Benjamini and Hochberg, 1995] it is less conservative than Bonferroni and creates a new rescaled p-values called q-values starting by ordering the raw p-values by descending. They are then compared with a standard threshold to select the significant hits. The method choice depends on the situations, usually if the risk of accounting for some false positives is not too expensive, FDR let to enlarge the number of positive associations. Otherwise if the researcher is very worried in catching false positives even at the cost of discarding plausible good signals, then Bonferroni might be preferable. However, a standard threshold for EWAs (and more general genome-wide studies) has not yet been defined, not even within the frame of a specific method (eg: Bonferroni). A solid, but now slightly less common, alternative is represented by the Holm sequential technique which, like Bonferroni, corrects the familywise error rate [Holm, 1979]. It is more powerful than Bonferroni keeping the error rate under control but in recent years FDR became a more common solution.

For the meta-analysis of the short- and mid-term exposure at the three different trailing averages, we based our choice following a paper from Dudbridge F. and Gusnanto A. [Dudbridge and Gusnanto, 2008] where they do not address a specific edge but a reasonable range. We were slightly more severe than how a pure Bonferroni threshold would have been with our data ($0.05/430.000 \sim 1.2E-07$), setting it at 7.5E-08. It has become common for genome-wide studies to use FDR as correction method arguing that Bonferroni assumes independence of the tests, which is surely not our case or any EWA case. However, we judged the risk of noise leading to false positives, in genome-wide DNA methylation studies, very high and decided to keep as main multiple comparison approach the Bonferroni method. FDR was further used as less conservative technique in order to look at CpGs that displayed a lower degree of significance but annotated with genes annotated with loci identified through Bonferroni. This double-step approach might lead to the discovery of regions within the genome that might show sensibility to endogenous exposures instead of single loci. This leads the discussion into a new arena of which yet very little is known: the study of correlation and interaction across CpG sites throughout the genome. Based on published literature about this theme, a short discussion will take place later in this work.

A different approach was used in the lung function project. As mentioned before, meta-analysis wasn't applied on these but a simpler replication strategy. A probe was considered genome-wide significant at the first step (from NAS results) if it had a Holm p-value equal or lower than 0.05. Moreover, always in NAS, CpG sites that were significant at the FDR level of 0.10 were also investigated for pathway analysis as well as for replication. At the second stage, for replication on KORA results, despite analysis were run for the whole package of selected CpGs (∼2000), FDR was used accounting only for the loci that showed significance in NAS, independently per each considered lung function. Finally, Intra Correlation Coefficient (ICC) was estimated in order to check stability across the studies (otherwise known as metastability).

6.6 Functional Analysis

After highlighting the CpGs from the genome-wide analysis, several tools have been used in order to determine plausible functional associations. Through Pubmed and GeneCards (http://www.genecards.org/), the involved genes were defined and classified. Then, to look at possible links across them and with any other gene, a web-interface, GeneMania, was extensively consulted [Warde-Farley et al., 2010]. GeneMania uses protein databases in order to map and link genes. In this way we extended our list of genes with others that look very associated or show interplay with the ones identified in our analysis. The approach was dual: to look for previous publications that found the genes in our list involved with physiological processes or even directly to diseases and to find if they had already been linked with air pollution or similar exposures.

Chapter 7

Results

This chapter will present results obtained in the different aforementioned projects. It will start with results on short- and mid-term exposure (that will take the larger part), then discuss long-term and finally lung functions.

7.1 Short- and Mid-term Results

We obtained data from three independent studies: KORA F3 and KORA F4 from the region Augsburg (Germany) and NAS (Boston Area, USA) (Table 6.1). While both the KORA studies accounted for around an equal sex distribution (52% of males in F3 and 49.3% in F4), the NAS is a totally male cohort. Differences can also be found in average age with 53 and 60 years old for the German cohorts and 72 for the American, in mean educational years with around 11 in both F3 and F4 and 15 in NAS and in drinker proportion where F3 registered 59.2% of drinkers, F4 57.7% and NAS 19.7%. On the other side, BMI resulted rather consistent with average values around 27 and 28. Substantial differences were also observed for smoking. F3 participants were mostly split between never (45.2%) and current smokers (46%), F4 in former (43.5%) and current (41.9%) and NAS registered mostly former smokers (67.9%), a smaller percentage of never smokers (28.6%) and a few current (3.5%). Regarding the particle concentration the day before the visit we observed a higher average in F3 with a value of 20.0 μg/m^3, while it was assessed at 14.2 μg/m^3 in F4 and 10.6 μg/m^3 in NAS. Instead temperature was higher in NAS with 12.5 °C against 7.1 °C and 8.7 °C in F3 and F4 respectively. Finally, DNA methylation showed consistency across the three studies with a relatively small standard deviation.
Results of the meta-analysis showed the identification of significant loci at all the three trailing averages considered, from 2-day up to 4-week exposure (Table 7.1, Table 7.2, Table 7.3, Table 7.4, Figure 7.1).

Coefficients are expressed as per increase of 10 μg/m^3 in PM$_{2.5}$. One CpG site was observed as Bonferroni genome-wide significant at 2-day trailing average, (cg25575464 within *NEURL4*, chromosome 17), displaying a positive association implying an increase in

Name	Chr.[a]	Reference Gene Name	Relation to CpG Island	Methylation level Illumina Beta, Mean ± SD		
				F3	F4	NAS
Trailing 2-day average $PM_{2.5}$						
cg25575464	17	*NEURL4*	Island	.03 ± .01	.02 ± .01	.01 ± .01
Trailing 7-day average $PM_{2.5}$						
cg04078416	3	*CCDC12*	Island	.05 ± .01	.05 ± .01	.02 ± .01
cg15996282	5	*LMBRD2; SKP2*	Island	.04 ± .01	.04 ± .03	.02 ± .01
cg00402617	8	*YWHAZ*	Island	.07 ± .01	.06 ± .02	.03 ± .01
cg19963313	8	*NSMAF*	Island	.04 ± .01	.03 ± .01	.02 ± .01
cg15883382	10		Island	.04 ± .01	.05 ± .01	.02 ± .01
cg09225537	15	*MAG*	N. Shore	.03 ± .01	.02 ± .01	.01 ± .01
cg08757611	17		Island	.03 ± .01	.03 ± .01	.02 ± .01
cg25575464	17	*NEURL4*	Island	.03 ± .01	.02 ± .01	.01 ± .01
cg02608596	19	*MPND*	Island	.04 ± .01	.03 ± .02	.02 ± .01

[a] CHR: chromosome

Table 7.1: Characteristics of the CpG sites from meta-analyses of 2- and 7-day trailing averages, significant with Bonferroni and FDR methods.

Name	Methylation level Illumina Beta, Mean ± SD[a]	Regression Coefficient[b]	P-value[c]	FDR[d]
Trailing 2-day average $PM_{2.5}$				
cg25575464	.02 ± .01	0.00082	4.69E-08	0.005
Trailing 7-day average $PM_{2.5}$				
cg04078416	.04 ± .01	0.0001	4.19E-07	0.027
cg15996282	.04 ± .02	0.0017	7.69E-08	0.010
cg00402617	.06 ± .02	0.0002	1.29E-07	0.018
cg19963313[e]	.03 ± .01	0.0018	2.49E-08	0.016
cg15883382	.04 ± .01	0.0001	8.43E-07	0.040
cg09225537	.02 ± .01	0.0001	4.44E-07	0.027
cg08757611	.03 ± .01	9.70E-05	2.15E-07	0.018
cg25575464	.02 ± .01	0.0001	1.76E-07	0.018
cg02608596[e]	.03 ± .02	0.0017	7.69E-08	0.010

[a] Calculated across KORA F3, F4 and NAS

[b] Methylation change for an increase of $PM_{2.5}$ of 10 $\mu g/m^3$ adjusted for sex, age, income (education years for NAS, in which information on income was not available), smoking status, alcohol intake, BMI, temperature (moving average always matching with the PM exposure window), day of the week, season and the proportion of five estimated white blood cell types: Monocytes, B Cells, CD8 T Cells, CD4 T Cells, NK

[c] Bonferroni significance level at 7.5E-08

[d] FDR: False Discovery Rate with Benjamini-Hochberg method, significance level at 0.05

[e] Shown in Figure 7.2

Table 7.2: Mean and meta-analysis results of the CpG sites of 2- and 7-day trailing averages, significant with Bonferroni and FDR methods.

Name	**CHR**[a]	**Reference Gene Name**	**Relation to CpG Island**	**Methylation level Illumina Beta, Mean ± SD**		
				F3	F4	NAS
cg16308101	1	*SERBP1*	Island	.45 ± .03	.46 ± .03	.44 ± .03
cg16856342	1	*SERBP1*	Island	.46 ± .02	.46 ± .02	.38 ± .02
cg23276912	1	*C1orf212*	S. Shore	.87 ± .03	.89 ± .03	.86 ± .04
cg03455255	2	*TSPYL6; ACYP2*	Island	.90 ± .02	.92 ± .01	.93 ± .02
cg11046593	2	*MSGN1*	-	.80 ± .05	.83 ± .09	.86 ± .07
cg04423572	3	*LOC100128640*	N. Shelf	.70 ± .04	.74 ± .04	.74 ± .03
cg19963313	8	*NSMAF*	Island	.04 ± .01	.03 ± .01	.02 ± .01
cg13169286	10	-	S. Shore	.55 ± .03	.59 ± .07	.51 ± .06
cg02795981	10	*ZMIZ1*	-	.78 ± .05	.78 ± .06	.79 ± .08
cg19215199	10	*ZMIZ1*	S. Shore	.82 ± .04	.83 ± .04	.82 ± .06
cg13527922	11	*F2*	-	.86 ± .02	.87 ± .02	.87 ± .02
cg24101979	17	*NXN*	N. Shore	.81 ± .03	.77 ± .04	.80 ± .05
cg26003785	17	*NXN*	N. Shore	.94 ± .01	.96 ± .01	.97 ± .02
cg26283240	17	*NXN*	S. Shore	.87 ± .03	.86 ± .03	.88 ± .04
cg06004017	22	*MN1*	N. Shore	.86 ± .02	.90 ± .02	.87 ± .03
cg20680669	22	*MN1*	N. Shelf	.96 ± .02	.96 ± .02	.99 ± .01

[a] CHR: chromosome

Table 7.3: Characteristics of the CpG sites from meta-analysis of 28-day trailing average, significant with Bonferroni method, or FDR significant and located in a gene with another CpG that meets genome-wide significance, or FDR significant and Bonferroni significant at shorter time-window.

Name	Reference Gene Name	Methylation level Illumina Beta, Mean ± SD[a]	Regression Coefficient[b]	P-value[c]	FDR[d]
cg16308101	*SERBP1*	.45 ± .03	-0.0076	2.86E-08	0.002
cg16856342	*SERBP1*	.44 ± .02	-0.0061	1.74E-07	0.003
cg23276912[e]	*C1orf212*	.90 ± .03	0.0073	4.56E-08	0.002
cg03455255	*TSPYL6; ACYP2*	.92 ± .02	0.0047	1.86E-08	0.001
cg11046593[e]	*MSGN1*	.83 ± .08	0.016	1.12E-08	0.001
cg04423572	*LOC100128640*	.73 ± .04	0.013	7.26E-09	0.001
cg19963313	*NSMAF*	.03 ± .01	0.0024	4.12E-07	0.005
cg13169286	-	.57 ± .06	-0.013	6.21E-08	0.003
cg02795981	*ZMIZ1*	.78 ± .06	0.0093	3.94E-05	0.029
cg19215199	*ZMIZ1*	.83 ± .04	0.0093	3.66E-08	0.002
cg13527922	*F2*	.87 ± .02	0.0051	1.54E-08	0.001
cg24101979	*NXN*	.78 ± .04	0.0072	8.95E-05	0.001
cg26003785[e]	*NXN*	.96 ± .01	0.0038	9.53E-09	0.001
cg26283240	*NXN*	.87 ± .03	0.0065	2.03E-05	0.024
cg06004017	*MN1*	.89 ± .02	0.0046	0.00019	0.048
cg20680669	*MN1*	.97 ± .02	-0.0049	2.09E-08	0.001

[a] Calculated across KORA F3, F4 and NAS

[b] Methylation change for an increase of $PM_{2.5}$ of 10 $\mu g/m^3$ adjusted for sex, age, income (education years for NAS, in which information on income was not available), smoking status, alcohol intake, BMI, temperature (moving average always matching with the PM exposure window), day of the week, season and the proportion of five estimated white blood cell types: Monocytes, B Cells, CD8 T Cells, CD4 T Cells, NK

[c] Bonferroni significance level at 7.5E-08

[d] FDR: False Discovery Rate with Benjamini-Hochberg method, significance level at 0.05

[e] Shown in Figure 7.3

Table 7.4: Mean and meta-analysis results of the CpG sites of 28-day trailing averages, significant with Bonferroni method, or FDR significant and located in a gene with another CpG that meets genome-wide significance, or FDR significant and Bonferroni significant at shorter time-window.

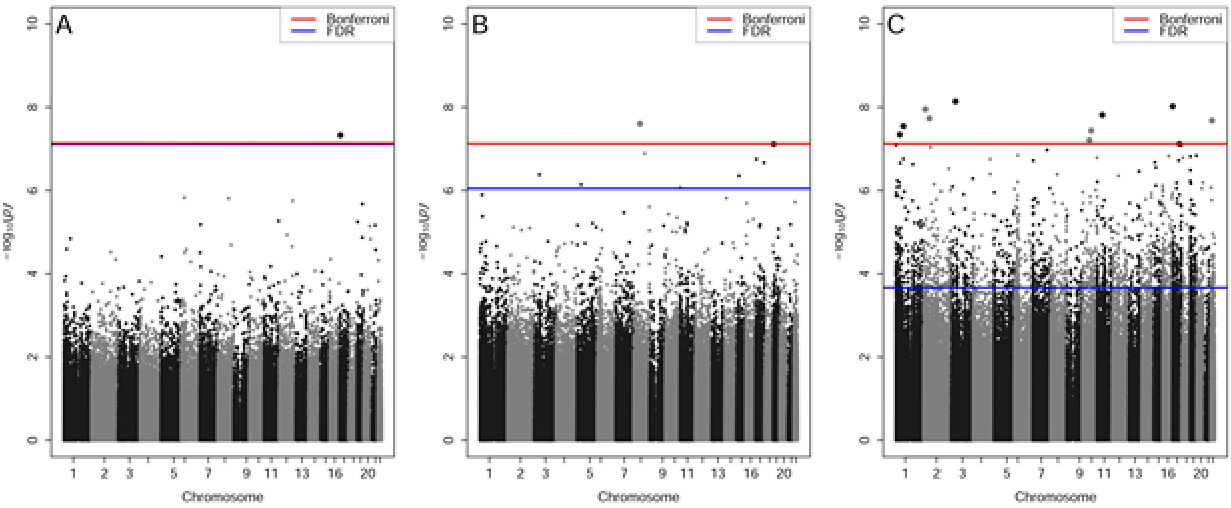

Figure 7.1: Manhattan plots showing p-values from the meta-analysis of KORA F3, KORA F4 and NAS longitudinal cohort studies across the human genome. Each dot corresponds to a CpG methylation site. Panel A: 2-day $PM_{2.5}$ exposure; Panel B: 7-day $PM_{2.5}$ exposure; Panel C: 28-day $PM_{2.5}$ exposure.

DNA methylation at elevated particle exposure. Despite the three study-specific estimates were all positives, heterogeneity was significant among the studies. Applying FDR as less conservative method, no other CpGs site appeared as significant. Moving to the 7-day trailing average, we observed another CpG site that reaches the genome-wide significance level (cg19963313 on *NSMAF*, chr. 8). Despite the positive association, p-value for cg02608596 (on *MPND*, chr. 9) resulted slightly above the threshold. Both showed a positive association and fulfilled the criteria for homogeneity. Conversely than the 2-day average, FDR correction added to the two mentioned CpGs other 7 loci with p-value < 0.05. Within them, it also appears cg25575464, Bonferroni significant at 2-day $PM_{2.5}$. Homogeneity among the studies was observed in four cases and heterogeneity in three. Finally, a 28-day average in $PM_{2.5}$ concentration was found as associated with ten CpG sites: cg16308101 (on *SERBP1*, chr. 1), cg23276912 (*C1orf212*, 1), cg03455255 (*TSPYL6*, *ACYP2*, 2), cg11046593 (*MSGN1*, 2), cg04423572 (*LOC100128640*, 3), cg13169286 (no annotated gene, 10), cg19215199 (*ZMIZ1*, 10), cg13527922 (*F2*, 11), cg26003785 (*NXN*, 17), cg20680669 (*MN1*, 22). Three of them showed decreased DNA methylation while the other seven showed increased DNA methylation at similar effect size. Homogeneity across the study sites was also reached in three out of ten of these CpGs (cg23276912, cg11046593, and cg26003785). Specific Manhattan plots are provided in Figure 7.1 and detailed results on significant CpG sites in Table 7.2 and Table 7.4. Considering FDR, a number of 1,819 additional CpGs displayed a p-value < 0.05 suggesting association with 28-day $PM_{2.5}$ exposure including five sites annotated in genes with at least one Bonferroni significant CpG: cg16856342 (*SERBP1*, chr 1), cg02795981 (*ZMIZ1*, chr 10), cg24101979 and cg26283240 (*NXN*, chr 17) and cg06004017 (*MN1*, chr 22). Additionally, the new list of 1,819 loci FDR associated with 28-day $PM_{2.5}$, included also cg19963313, Bonferroni significant at 7-day exposure. Within this longer list, several CpGs were found to be associated within one single gene: the top-ranked was *MAD1L1* (on chromosome 7) which appeared 11 times and *PRDM16* (on chromosome 1) 8 times. We unfortunately acknowledged that several loci are rather isolated signals as shown by the regional plots in Figure 7.2 and Figure 7.3 (right sides). For this reason we also checked how our results would look like adjusting the multiple comparisons with FDR, a less conservative method. A further plot that may simplify the temporal variation for the top CpGs is showed in Figure 7.4 and it has been recalled "Spider Plot". The top ten significant CpGs per each

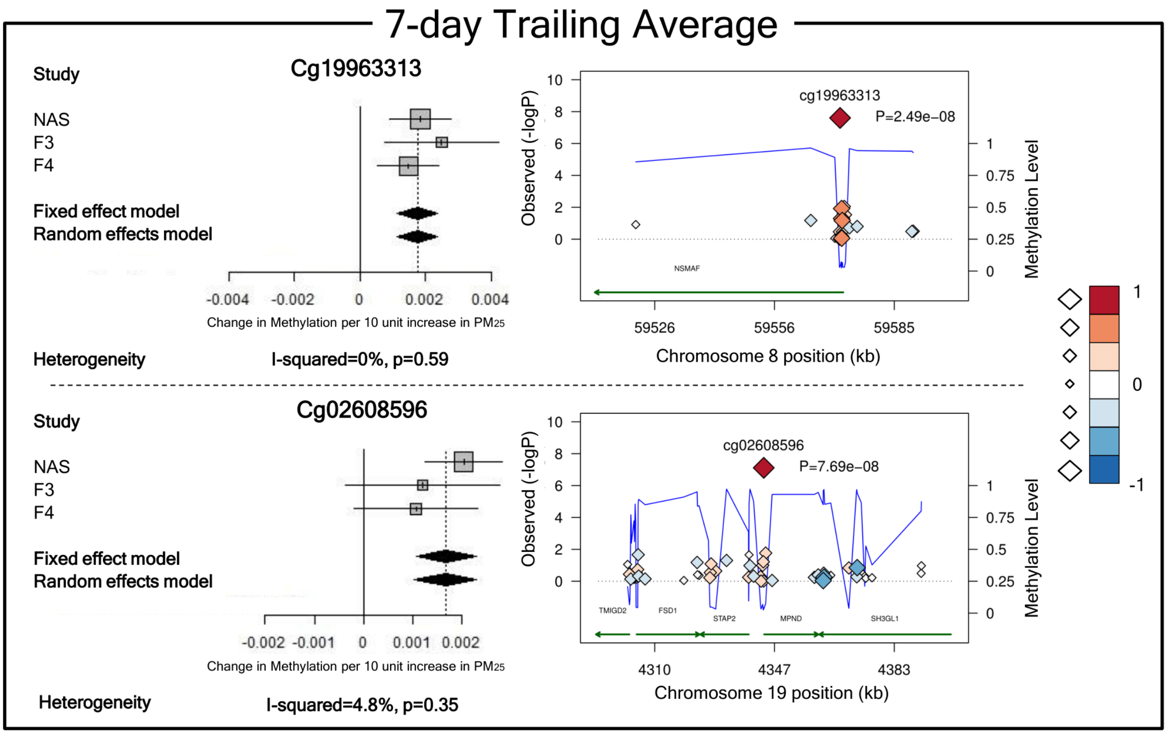

Figure 7.2: Forest plots (left side) and Regional plots regarding the CpG sites that achieved genome-wide significance level and homogeneity at 7-day average. Forest plots show KORA F3, KORA F4 and NAS longitudinal cohort estimates and pooled meta-analysis results. Regional plots show the p-values of each annotated CpG sites (diamonds) in a 200k bp length genome segment around the top CpG. The color and the size of the diamonds represent the intensity of the correlation with the top CpG target (in the center). The blue broken line connects the average methylation value of adjacent CpG sites; the right axis displays the 0-1 methylation scale. Correlations and averages values are calculated as mean of the three studies. Yellow outlined diamonds highlight FDR significant CpG sites.

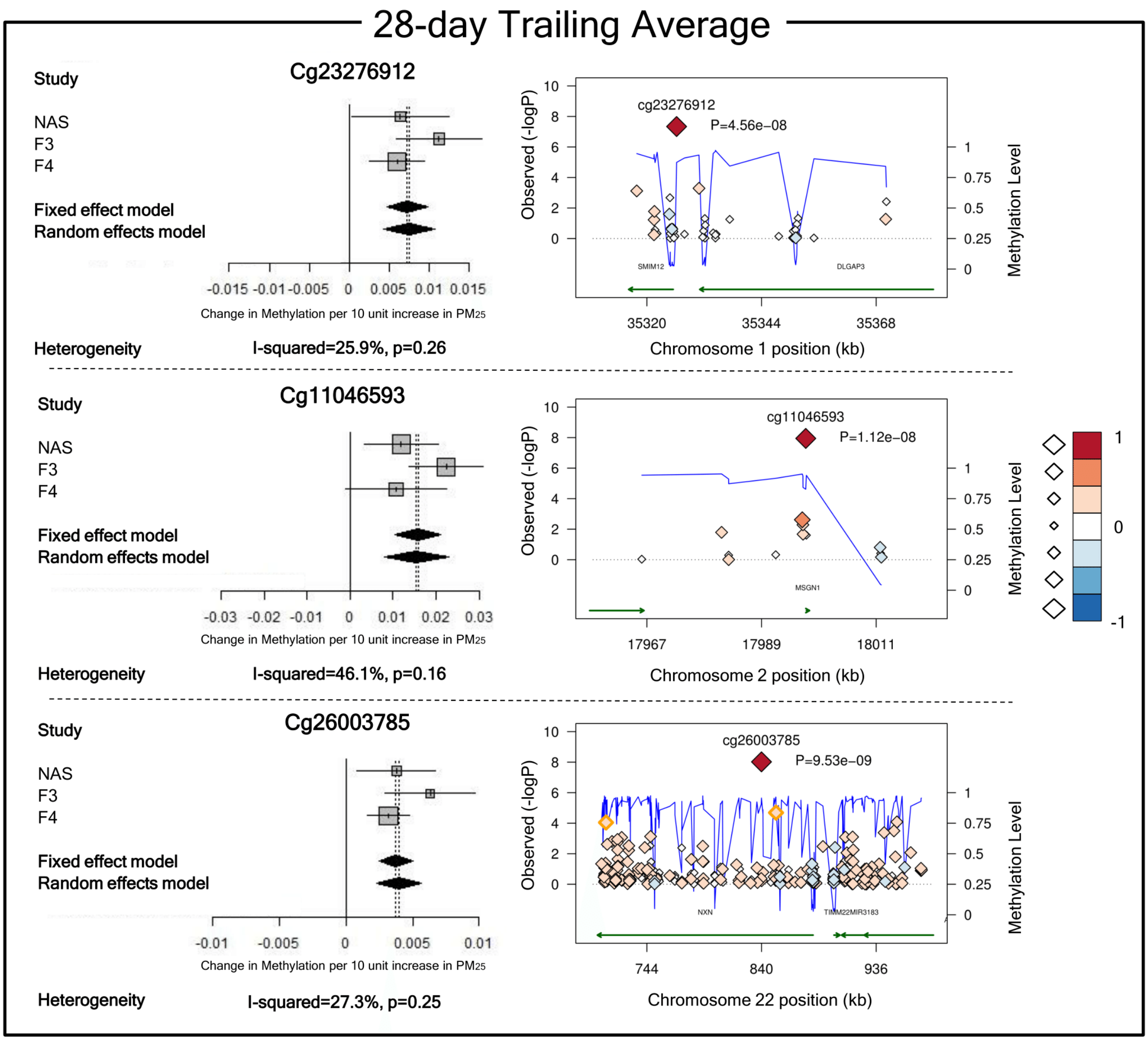

Figure 7.3: Same as Figure 7.2, on 28-day average homogeneous results.

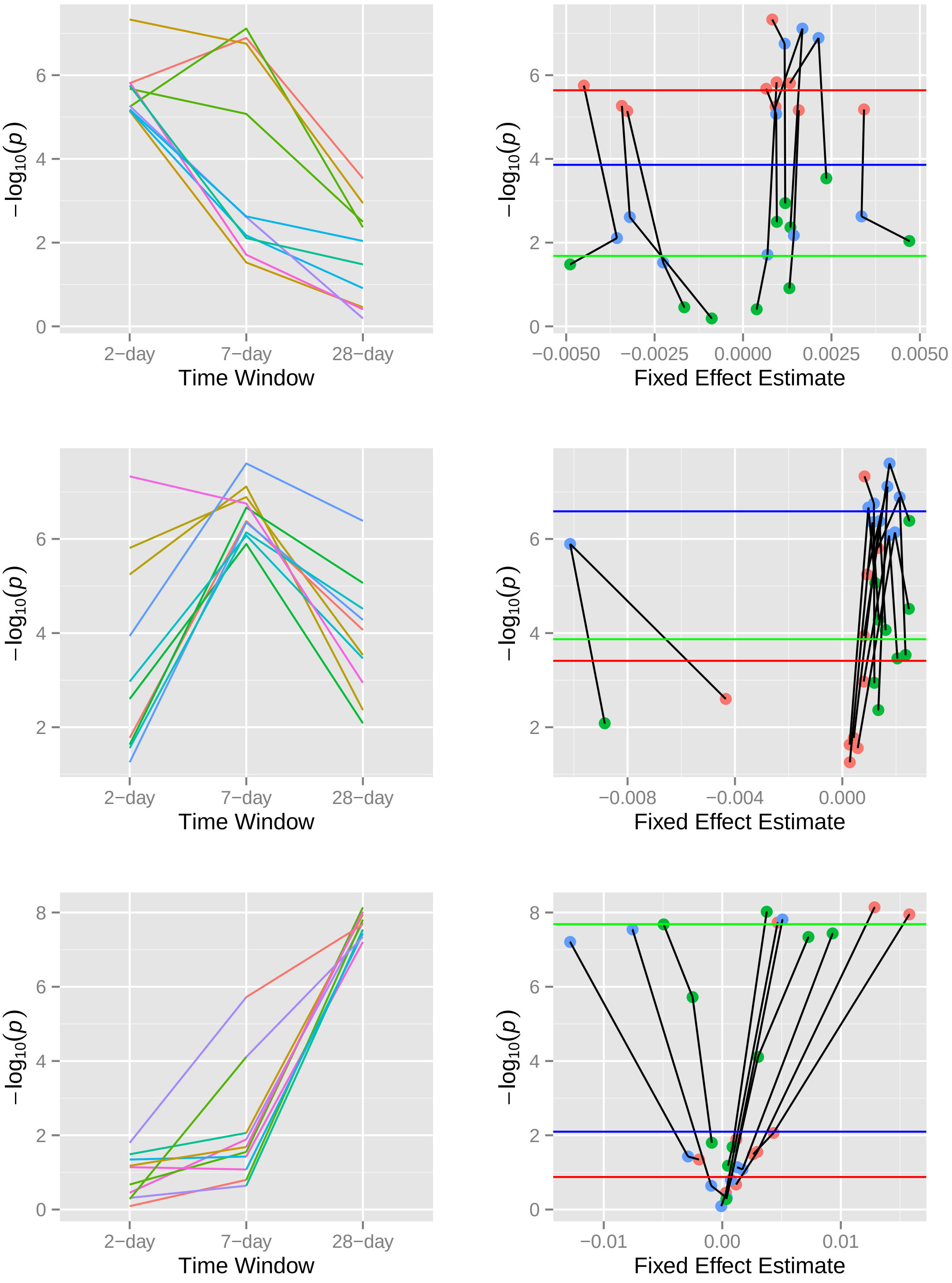

Figure 7.4: Behavior of the top 10 CpG sites identified at 2-day (panel A), 7-day (B) and 28-day (C) exposure over time. On the left panel it is shown how p-values vary within the different averaging period, one broken line for each CpG. On the right panel we show estimates and p-values on horizontal and vertical axis, one broken line for each CpG, the colors represent each temporal average: red for 2-day, blue 7-day, green 28-day and the corresponding colored horizontal lines lie on the average level of the p-values.

trailing average have been considered: on the left side we can observe how the p-values vary from one trailing average to the other, while on the right side it is shown how the effect estimates vary in relationship with the time window (colors) and the p-value (vertical axis). Two more things are also important to highlight. The first is that a general decrease of p-values has been observed extending the length of the time window. The second is the unbalanced proportion of CpGs that displayed increased and decreased methylation, positive coefficients have been observed approximately twice more often than negatives. Results for sensitivity analysis confirmed the reasonability regarding the selected priori model and the fact that it is necessary in order to best separate the sources of the variability and purify the effect of the short-term concentration. The amount of significant hits at the two partially adjusted models is reported in Table 7.5.

	Model 1		**Model 2**		**Model 3**	
	FDR	**Bonferroni**	**FDR**	**Bonferroni**	**FDR**	**Bonferroni**
2-day Trailing Average	29	1	34	1	1	1
7-day Trailing Average	696	18	988	14	9	2
28-day Trailing Average	37377	431	52800	563	1829	10

Table 7.5: Sensitivity analysis, comparison of amount of FDR and Bonferroni significant CpGs for three models: 1- adjusted only with age and sex; 2- model 1 plus white blood cell proportions; 3- full adjustment.

Considering the minimal adjustment with only sex and age, the number of significant CpG sites increases dramatically at all the three time windows. It slightly decreases form the crude model, as expected, by adding the white blood cell proportions. In contrast, after adjusting for the long term exposure, our top 10 CpGs at 28-day average show high consistency with the previous results in both estimate and p-value, except for cg20680669 and cg26003785 which estimates respectively moved from a β = -0.0049 with p = 2.09E-08 (without long-term) to β = -0.0020 with p = 2.36E-03 and from β = 0.0038 with p = 9.53E-09 to β = 0.0033 with p = 1.10E-06. Table 7.6 shows the differences in coefficient and p-value for the 28-day significant CpGs without and with adjustment for the long-term exposure. Furthermore a special focus has also been driven to the problem of outliers. Residuals of FDR significant CpGs at 2- and 7-day trailing averages and Bonferroni significant at 28-day were plotted and manually checked in order to detect unusual values. Cg11046593 have been identified as problematic locus (according to the approach explained in the Chapter 6, Statistical Analyses), therefore we excluded 22 values for F4, 1 for F3 and 12 for NAS. Results of the new values on the reduced dataset didn't delineate a relevant

	With long-term		**Without long-term**	
	β^*	**P-value**	β^*	**P-value**
cg16308101	-0.0070	4.64E-07	-0.0076	2.86E-08
cg23276912	0.0079	8.85E-09	0.0073	4.56E-08
cg03455255	0.0050	5.37E-09	0.0047	1.86E-08
cg11046593	0.016	1.28E-07	0.016	1.12E-08
cg04423572	0.013	1.23E-08	0.013	7.26E-09
cg13169286	-0.014	5.97E-08	-0.013	6.21E-08
cg19215199	0.0095	4.32E-08	0.0093	3.66E-08
cg13527922	0.0056	2.09E-09	0.0051	1.54E-08
cg26003785	0.0033	1.10E-06	0.0038	9.53E-09
cg20680669	-0.0020	2.36E-03	-0.0049	2.09E-08

* β express change in % 5mC according to a 10 unit increase in $PM_{2.5}$ exposure

Table 7.6: Sensitivity analysis, comparison of regression coefficients (β) and p-values in 28-day significant hits with and without yearly $PM_{2.5}$ exposure adjustment.

change and the association remained significant: the estimate changed from 0.016 to 0.012 and the p-value from 1.12E-08 to 5.48E-08. A final comparison between air pollution and temperature estimates have also been conducted and didn't show remarkable differences, resulting similar in magnitude as shown in Table 7.7 and Table 7.8.

Particle number concentration results provided instead a more confusing scenario: after meta-analysis hundreds of CpG sites were found as associated at each trailing average. After several checks such as outliers, confounding and very carefully the code, we were not able to find a rationale behind it. It seems unlikely that all these signals are real and a possibility is that one or more factors, that may confound the real effects, are still missing.

7.2 Long-term Results

LUR long-term exposure estimates have also been used as exposure. Three different air pollution indicators were accounted: the gaseous NO_2 and the particles $PM_{2.5}$, PM_{10}. In order to increase power, KORA F3 and KORA F4 were pooled together accounting for almost 2,300 participants. Despite the increased sample size, it wasn't observed any strong consistent association. NO_2 and $PM_{2.5}$ displayed a few CpG sites as associated with the exposure but only using the Tost-preprocessed data and this severe lack of consistency (even comparing probes with higher p-values, in the order of E-05 or E-06) left behind some doubts regarding the credibility of these results. These loci didn't replicate using the BMIQ pipeline and, additionally, the more the literature updated, the more the skepticism regarding the Tost pipeline increased, until even hypothesizing that can add noise to the data, leading to false positives. In conclusion we couldn't strongly associate any signal to long-term air pollution exposure but these results confirmed the importance of an accurate pipeline: analysis on same data but after different preprocessing produced rather different results. The fact that the scarce evidence that we observed on long-term data, when results have so little consistency just by changing the preprocessing, probably means that they were mostly influenced by noise.

7.3 Lung Functions

Differences between KORA F4 and NAS participants have already been described but not regarding lung functions. We observed a high degree of consistency for $FEF_{25\%-75\%}$, but a small difference for the other two categories. In KORA F4, values for FVC and FEV_1 resulted slightly larger than in NAS with values around 4.20 in the German study vs 3.60 in the American for FVC and 3.30 vs 2.70 for $FEF_{25\%-75\%}$. This is probably due to the effect of age, NAS participants are averagely 10 years older than KORAs.
Results of the preliminary EWAs on NAS highlight nine genome-wide significant hits (at Holm correction level) as shown in Table 7.9, Table 7.10.

CpG	**NAS**		**F3**		**F4**	
	β^* **PM**	β^* **Temp.**	β^* **PM**	β^* **Temp.**	β^* **PM**	β^* **Temp.**
Trailing 2-day average PM$_{2.5}$						
cg16308101	0.00086	0.00050	-0.00017	-0.00138	0.00031	0.00016
cg23276912	0.00013	0.00042	0.00119	0.00139	0.00016	0.00127
cg03455255	0.00028	0.00132	0.00094	-0.00052	0.00019	0.00110
cg11046593	0.00468	-0.00162	0.00771	-0.00114	-0.00592	0.00220
cg04423572	0.00189	0.00341	0.00731	0.01175	-0.00100	-0.00373
cg19963313	0.00082	-0.00018	0.00132	0.00414	0.00067	0.00053
cg13169286	0.00016	0.00356	-0.00174	-0.01280	-0.00301	-0.00451
cg19215199	0.00146	-0.00334	0.00584	0.00493	0.00017	0.00162
cg13527922	0.00062	-0.00005	0.00185	0.00230	-0.00068	-0.00035
cg25575464	0.00168	-0.00128	0.00054	0.00177	0.00015	0.00014
cg26003785	0.00100	-0.00026	0.000200	-0.00028	0.00016	-0.00024
cg20680669	-0.00175	0.00061	-0.00171	0.00050	-0.00020	0.00048
Trailing 7-day average PM$_{2.5}$						
cg16308101	0.00340	-0.00401	-0.00440	-0.00481	0.00022	-0.00026
cg23276912	-0.00121	0.00427	0.00449	0.00427	0.00256	0.00245
cg03455255	-0.00089	0.00169	0.00207	-0.00095	0.00092	0.00108
cg11046593	0.00450	0.00394	0.00983	-0.00578	-0.00419	0.01015
cg04423572	0.00350	0.00471	0.01304	0.02375	-0.00076	-0.00529
cg19963313	-0.00035	0.00183	0.00252	0.00465	0.00126	0.00084
cg13169286	0.00042	-0.00106	-0.00533	-0.01860	-0.00196	-0.00128
cg19215199	0.00637	-0.00086	0.00961	0.00336	-0.00029	0.00227
cg13527922	-0.00148	0.00259	0.00348	0.00037	-0.00020	-0.00036
cg25575464	-0.00050	0.00246	0.00047	0.00199	0.00049	0.00057
cg26003785	-0.00293	0.00215	0.00317	-0.00026	0.00029	-0.00037
cg20680669	0.00096	-0.00566	-0.00474	-0.00185	-0.00066	-0.00053

[*] β express change in % 5mC according to a 10 unit increase in PM$_{2.5}$ exposure

Table 7.7: Study specific regression coefficients of PM and Temperature for all Bonferroni significant CpGs at 2- and 7-day trailing averages.

CpG	NAS		F3		F4	
	β^* **PM**	β^* **Temp.**	β^* **PM**	β^* **Temp.**	β^* **PM**	β^* **Temp.**
Trailing 28-day average $PM_{2.5}$						
cg16308101	-0.00934	0.01292	-0.02129	-0.01348	-0.00334	-0.00071
cg23276912	0.00695	-0.00317	0.01048	0.00847	0.00582	0.00235
cg03455255	0.00288	0.00038	0.00863	0.00441	0.00408	0.00086
cg11046593	0.00989	0.00424	0.02228	-0.00868	0.01036	0.00167
cg04423572	0.00634	-0.00117	0.04365	0.05288	0.00021	-0.00776
cg19963313	0.00262	-0.00011	0.00172	0.00088	0.00187	0.00300
cg13169286	-0.00331	-0.00086	-0.02180	-0.03602	-0.00775	-0.00090
cg19215199	0.00248	0.01015	0.02568	0.01168	0.00398	0.00246
cg13527922	0.00361	0.00129	0.01043	0.00580	0.00350	-0.00103
cg25575464	0.00229	-0.00096	-0.00092	-0.00043	0.00043	0.00089
cg26003785	0.00368	-0.00835	0.00655	0.00120	0.00313	-0.00034
cg20680669	-0.00609	-0.00039	-0.00692	0.00054	-0.00253	-0.00164

* β express change in % 5mC according to a 10 unit increase in $PM_{2.5}$ exposure

Table 7.8: Study specific regression coefficients of PM and Temperature for all Bonferroni significant CpGs at 28-day trailing average.

CpG (Gene)	PFT[a]	Unadjusted for smoking		Adjusted for smoking		ICC[b]
		Effect Estimate	P-value	Effect Estimate	P-value	
β_4[c]						
cg02721176 (C10ORF96)	FEV_1	0.03	2.29E-08	0.03	N.S.	0.81
cg09995068	FEV_1	-0.41	2.09E-09	-0.34	N.S.	0.99
cg14292220	FEV_1	0.23	3.89E-10	0.19	N.S.	0.95
cg01249054	FEV_1	0.62	9.43E-08	0.52	N.S.	0.78
cg18476993 (MIR671; CHPF2)	FEV_1	0.00	N.S.	0.14	6.06E-08	0.64
cg05644990 (GPBP1)	FVC	0.84	1.25E-08	0.86	8.42E-09	2.80E-14
cg12565126 (MFSD2B)	FVC	0.09	2.33E-08	0.09	2.45E-08	0.88
cg13532885 (SYN1)	FVC	0.06	4.69E-08	0.06	4.41E-08	0.93
cg26468478 (CELSR3)	FVC	0.90	1.50E-08	0.90	2.17E-08	0.032
cg03867607 (MYL6)	FVC	0.00	N.S.	0.39	9.09E-08	0.76
cg05191655	FVC	0.00	N.S.	0.15	7.34E-08	0.73

[a] The genomic inflation β-values, with no adjustment for smoking for the longitudinal effect (β_4), were 1.36, 1.14, and 1.03 for the FEV_1, FVC and $FEF_{25\%-75\%}$ models, respectively; after adjusting for smoking, the β-values were 1.20, 1.48, and 1.08 for FEV_1, FVC and $FEF_{25\%-75\%}$, respectively.

[b] ICC: Interclass Correlation Coefficient

[c] β_4 measures the effect of DNAm on the rate of decline in lung function, i.e., the interaction between DNA methylation and follow-up time.

[N.S.] Non Significant

Table 7.9: Total of 11 unique DNA methylation sites associated with a given Pulmonary Function Test (PFT) model organized by smoking analysis, unadjusted (left side) and adjusted for smoking (right side), n = 657 participants in the NAS across all three smoking classes, never (190 participants), former (441), and current smokers (26).

CpG (Gene)	PFT[a]	Unadjusted for smoking		Adjusted for smoking		ICC[b]
		Effect Estimate	P-value	Effect Estimate	P-value	
β_2[b]						
cg03636183 (F2RL3)	$FEF_{25\%-75\%}$	1.01	1.30E-08	0.72	N.S.	0.83
cg05575921 (AHRR)	FEV_1	1.79	1.21E-16	1.56	4.74E-11	0.91
	FVC	1.40	3.33E-09	1.25	N.S.	
	$FEF_{25\%-75\%}$	1.26	9.52E-19	1.15	1.30E-12	
cg05951221 (ALPPL2)	$FEF_{25\%-75\%}$	0.96	5.35E-08	0.66	N.S.	0.89
cg06126421 (IER3)	FEV_1	1.44	1.14E-11	1.18	7.46E-08	0.93
	FVC	1.28	1.55E-08	1.14	N.S.	
	$FEF_{25\%-75\%}$	0.85	1.61E-09	0.64	N.S.	
cg21566642	FEV_1	1.29	4.67E-09	1.00	N.S.	0.90
	$FEF_{25\%-75\%}$	0.89	1.15E-09	0.69	N.S.	
cg15342087 (IER3)	FEV_1	3.06	1.20E-08	2.57	N.S.	0.67

[a] The genomic inflation β-values, with no adjustment for smoking for the cross sectional effect (β_2), were 0.71, 0.93, and 0.80 for the FEV_1, FVC and $FEF_{25\%-75\%}$ models, respectively; after adjusting for smoking, the β-values were 0.94, 0.71, and 0.83 for FEV_1, FVC and $FEF_{25\%-75\%}$, respectively.

[b] ICC: Interclass Correlation Coefficient

[c] β_2 represents the cross-sectional effect of DNAm.

[N.S.] Non Significant

Table 7.10: Total of 6 unique DNA methylation sites associated with a given Pulmonary Function Test (PFT) model organized by smoking analysis, unadjusted (left side) and adjusted for smoking (right side), n = 657 participants in the NAS across all three smoking classes, never (190 participants), former (441), and current smokers (26).

Of them, seven referred to the so-called effect of DNA methylation on the rate of decline in lung function (technically, the interaction term between DNA methylation and follow-up length): cg18476993 (on *MIR671* and *CHPF2*, chr. 7), cg05644990 (*GPBP1*, 5), cg12565126 (*MFSD2B*, 2), cg13532885 (*SYN1*, X), cg26468478 (*CELSR3*, 3), cg03867607 (*MYL6*, 12) and cg05191655 (no annotated gene, 4). The two remaining significant CpGs for the cross-sectional effects are cg05575921 (*AHRR*, 5) and cg06126421 (*IER3*, 6). It is worth to also highlight that cg05575921, annotated in the gene *AHRR*, was associated with two lung functions FEV_1 and $\text{FEF}_{25\%-75\%}$. Research on smoking exposure already highlighted more times this probe (and consequently the gene *AHRR*) and is a further discover to see it associated with lung function decline (after adjustment for smoking). Excluding smoking from the covariates provided a few more hits for both the parameters and, for those who kept being Holm significant, effect magnitudes remained regular. After sensitivity analysis on current and former smokers only, the probe on *AHRR* kept being significant as well as two probes associated with the rate of decline (cg05644990 and cg26468478). Results of replication showed that three probes reproduced the NAS effect also in KORA (meeting the two criteria of FDR adjustment and effect in the same direction) as shown in Table 7.11 and Table 7.12.

Two of these were probes linked with the cross-sectional effect (cg05575921 and cg06126421) while the third one (cg01086847) with the lung function rate of decline. A closer examination brought us to discard cg01086847 being very likely simply influenced by SNP near the target. The two CpG sites that replicated the effect, also showed high ICC, meaning high meta-stability. Moreover, they have also been found to be replicated by Shah et al. despite a not very high heritability in cg05575921 [Shah et al., 2014].

PFT	Variable	FDR significant hits in NAS[a]	Number of FDR -significant results in NAS also FDR -significant in KORA[b]	Number of FDR -significant KORA results with effect in same direction as NAS
KORA all participants				
$FEF_{25\%-75\%}$	β_2	1	1	1
	β_4	0	-	-
FVC	β_2	2	2	2
	β_4	270	0	0
FEV_1	β_2	0	-	-
	β_4	1212	0	0
KORA men				
$FEF_{25\%-75\%}$	β_2	1	1	1
	β_4	0	-	-
FVC	β_2	2	2	2
	β_4	270	0	0
FEV_1	β_2	0	-	-
	β_4	1212	15	1

[a] FDR defined as having a Benjamini Hochberg p-value less than 0.1.

[b] FDR defined as having a Benjamini Hochberg p-value less than 0.1, based on the number of FDR significant in NAS

Table 7.11: Comparison of NAS top findings within KORA when adjusted for cigarette use (smoking status and pack-years): all of the KORA participants (males and females) vs. only men.

PFT	Variable	FDR significant hits in NAS[a]	Number of FDR -significant results in NAS also FDR -significant in KORA[b]	Number of FDR -significant KORA results with effect in same direction as NAS
KORA oldest men				
$FEF_{25\%-75\%}$	β_2	1	1	1
	β_4	0	-	-
FVC	β_2	2	2	2
	β_4	270	0	0
FEV_1	β_2	0	-	-
	β_4	1212	0	0
KORA women				
$FEF_{25\%-75\%}$	β_2	1	1	1
	β_4	0	-	-
FVC	β_2	2	1	1
	β_4	270	0	0
FEV_1	β_2	0	-	-
	β_4	1212	0	0

[a] FDR defined as having a Benjamini Hochberg p-value less than 0.1.

[b] FDR defined as having a Benjamini Hochberg p-value less than 0.1, based on the number of FDR significant in NAS

Table 7.12: Comparison of NAS top findings within KORA when adjusted for cigarette use (smoking status and pack-years): oldest KORA men (55 years and up) vs. all women.

Chapter 8

Discussion

This chapter will discuss possible implications of our discoveries. A greater part will be taken by the CpG sites identified with short- and mid-term $PM_{2.5}$ exposure since the analysis led to a larger amount or results.

8.1 Short- and Mid-term Results

This meta-analysis conducted on three independent studies identified twelve CG dinucleotides genome-wide significantly associated with ambient fine particle matter concentration at a threshold of 7.5E-08. Based on previous findings on the literature three trailing averages were defined at 2-, 7- and 28-day prior the visit day. The twelve CpG sites were so temporally distributed: one at 2-day average, one at 7-day and ten at 28-day. Nine out of twelve sites displayed increased methylation and four of them were also found to be homogenous across the three studies. All of them displayed regularly a little overall variation (observed 15% average coefficient of variation) within the study populations. A less conservative multiple comparison approach unveiled a larger number of CpG sites, especially at 7-day (8 additional targets) and at 28-day (1,819 additional probes).

8.1.1 Homogeneous CpGs

The genome-wide significant CpG site (cg19963313) at 7-day average shows homogeneity among the studies. Cg19963313 is located in the gene *NSMAF* that has been previously linked with the 55kD tumor necrosis factor receptor, encoding a WD-repeat protein that binds a cytoplasmic sphingomyelinase activation domain [Montfort et al., 2010]. In addition it participates in the same reaction within a pathway as *SMPD2* [Wu et al., 2010], that, thanks to prior discoveries in primary cells, have been found to be linked to oxidative stress [Byon et al., 2008, Jana and Pahan, 2007]. Furthermore, it has also been identified as a plausible therapeutic target [Liao et al., 2013] and a determinant in cellular response to hypersmolar stress [Robciuc et al., 2012]. It is well known that hypersmolarity imposes a conspicuous stress on membranes, and it is even acuter on those in direct contact with the environment [Hallows et al., 1996], however, very little is known regarding its association with air pollution.

MPND, the gene that includes cg02608596 (which p-value resulted close above the Bonferroni genome-wide significance threshold), has interaction with ubiquitin specific peptidase. Ubiquitination is a well-known process involved post-transcriptional histone modifications (another very well characterized and investigated epigenetic marker related to chromatin modeling) and it induces us to hypothesize that even short temporal peaks of air pollution exposure may widely affect chromatic remodeling by influencing key players in chromatin function. Finally, a further implication of ubiquitination have been found in association with DNA repair and recovering from DNA damage [Al-Hakim et al., 2010, Ulrich and Walden, 2010]. Our work may present cellular response to a DNA damage caused or better provoked by particulate matter.

Three additional CpGs were found significant and fulfilled the criteria for homogeneity at the 28-day trailing average exposure to fine particle: cg26003785, cg23276912 and cg11046593, respectively annotated to *NXN*, *MSGN1* and *C1orf212*. All three genes belong to the categorization "protein-coding gene".

Specifically, *NXN* has been observed in partnership with phosphofructokinase (PFK) 1, a glycolytic enzyme reported to contribute to systemic metabolic conditions and cancerous processes [Mor et al., 2011, Yi et al., 2012]. In addition, *NXN* interacts with *CIR1*, a compressor interacting with *RBPJ*. It plays an important role in the Notch Signaling Pathway, which is involved in many functions regarding inter-cell communication and also gene regulation of various cell differentiation processes at both embryonic and adult stage of life.

DNA methylation was observed as increased at cg23276912, which is located in the promoter of *MSGN1*, which, whether methylated, has been shown to conduct to transcriptional repression [Jones and Takai, 2001]. It resulted reported as clutch element during maturation of mesoderm stage in embryonic development [Fior et al., 2012, Wittler et al., 2007] and, additionally, domain datasets reported shared protein domain between *MSGN1* and *AHR* and *ARNT* (Aryl Hydrocarbon Receptor and Aryl Hydrocarbon Receptor Nuclear Translocator, respectively). There are evidences in literature about the involvement of these genes in several biological processes including regulation of inflammatory and other endogenous processes among the ones known to have been associated with multi-factorial diseases like pulmonary disorders [Scrivo et al., 2011, Ukena et al., 2010]. Ovrevik et al. in 2014 found that these genes play a key role in regulating chemokine-responses mostly relating *AHR* and *ARNT* to NF-kB, the nuclear factor-kB family, where the regulation of the inflammatory responses is usually characterized by the p65/p50 dimer [Ovrevik et al., 2014]. The link between the gene *AHR* and PM exposure is not a novelty since was already highlighted through nongenotoxic events and Th17 polarization

[Andrysik et al., 2011, van Voorhis et al., 2013] but this work provides the first evidence of a possible epigenetic mediation between the ambient effect and the gene. In fact, results from Di Meglio et al. [Di Meglio et al., 2014] demonstrated an absence of *AHR* triggers dysregulation of skin cellular response to inflammatory stimuli, this work provided evidence for the increase of DNA methylation in a CpG site located in the promoter of a gene (*MSGN1*) sharing protein domain with *AHR* associated with an increase in PM exposure. Despite the indirect link, the identification of *MSGN1* may add useful information regarding the complex relationship that occurs between endogenous (often environmental) factors and the consequent responses of the immunological system. On the other side, no other study has previously displayed the gene *ARNT* in an air pollution study, so our finding constitutes a real novelty. Future studies are required with the aim of verifying and clarifying the role of *ARNT*.
Unfortunately we had no access to any functional information about *C1orf212*.

8.1.2 Heterogenous CpGs

Below is briefly described the role of the seven genes annotated with the other genome-wide significant CpGs but not sufficiently homogeneous across the studies.
Cg25575464 (*NEURL4*). After *MPND*, mentioned in the manuscript, this is another gene related with Ubiquitination.
Cg16308101 (*SERBP1*). Koensgen et al. [Koensgen et al., 2007] discovered overexpressed *SERBP1* in ovarian tumor epithelial cells and we observed decreased methylation in a CpG site annotated within that gene. Moreover, Serce and coworkers [Serce et al., 2012] presented a significant association of *SERBP1* expression in human breast carcinoma with favorable prognosis, hypothesizing it as a potential prognostic marker of tumor. This is, to our knowledge, the first study that links the *SERBP1* behavior with an environmental factor.
Cg19215199 (*ZIMZ1*). It was found by Lee et al. [Lee et al., 2007] as regulatory gene for the p53-mediated transcription, a tumor suppressor and it is consistent with our result, showing increased methylation at higher particle matter exposure, hence, eventually, increased gene silencing. Other publications helped to clarify how and where *ZIMZ1* operates [Rakowski et al., 2013, Soler et al., 2008] but this is the first study that links it with an environmental exposure.
Cg13527922 (*F2*, Thrombin). A well-known gene already linked to tumors. It was found by Hu and colleagues [Hu et al., 2004] playing an important role in the three steps of implantation, seeding and spontaneous metastasis. We observed increased methylation at ambient concentration of fine particle.
Cg20680669 (*MN1*). Already known as related to Acute Myeloid Leukemia (AML) and meningioma, in a Chinese cohort, Xiang and coauthors [Xiang et al., 2013] and Aref et

al. [Aref et al., 2013] confirmed the work made by Liu and coworkers [Liu et al., 2010], observing that, in human, overexpressed *MN1* is associated with AML consistently with the decreased methylation we observed.
TSPYL6 and *ACYP2*, annotated for the CpG site cg03455255 and *LOC100128640*, annotated for the CpG site cg04423572, were not found to be related with any specific disease or other outcomes.

8.1.3 Temporal Variation within Days and Weeks before the Visit

As said, one CpG have been found significant at 2-day average, one at 7-day and ten at 28-day, and they don't match, constituting twelve unique loci. This result opens the big question regarding the temporal variation of DNA methylation. But proceed with order. A systematic decrease of p-values has been observed consequently to the extension of the trailing average Figure 7.1. From the Manhattan plot can be clearly seen how the cloud of dots is extending its area towards the more statistically significant side of the graph and this is reflexed by the higher number of significant loci. This trend may suggest that increased levels of fine particle exposure need to persist over longer period of time than just a couple of days before epigenetic mechanisms alter the regulation in leukocytes exceptionally beyond their usual exposure, as presumably pinpointed by the number of differentially methylated CpG sites throughout the genome. P-values of CpG sites significant at 28-day average were decreasing moving from the 2-day to the 7-day and to the 28-day, where finally turned significant (as also shown in Figure 7.4 [Panel C]), meanwhile, p-values of probes significant at 7-day average didn't result significant at 28-day. In order to better understand the behavior of the selected CpG sites, further analysis are required. The hypothesis after this work is that eventual DNA methylation fluctuations according to a short- or mid-term exposure are probe-specific. The cases of the two CpG sites Bonferroni significant at a time-window but only FDR significant (less conservative method) at the next longer trailing average are examples that point to the fact that maybe there are CpGs which DNA methylation level is more sensible at shorter variations. However is very difficult to drawn any conclusion since a reference value doesn't exist. This is a crucial concern for this research and an open question that emerged from our results, having the possibility to choose several temporal windows and perform a rather extensive package of analysis.
A second pattern that we come out looking closely at our results is the proportion of positive and negative coefficients, reporting increased or decreased DNA methylation consequently an increase in $PM_{2.5}$. We predominantly observed positive DNA methylation coefficients, 9 out of 12 only within the significant ones. While a general trend is difficult to interpret, it is still possible to make sense of the results CpG by CpG. It is known from

several studies that epigenetic changes might increase gene silencing and it becomes an issue when tumor suppressors are silenced [Laird, 2005]. In our results, an example of this mechanism might be constituted by the identification of a CpG within the gene *ZIMZ1* which was already connected to skin tumor cells in mouse models [Rogers et al., 2013]. Little is known regarding the role of epigenetics in tumor development mediated by air pollution but this result might provide useful information. A previous study observed, instead, a higher prevalence of decreased DNA methylation associated with short-term PM exposure in tandem repeats [Guo et al., 2014]. However, the basic approach is substantially different: our results are the first that link PM exposure and DNA methylation on a genome-wide basis, while the excellent work done by Guo et al. still reports a gene-specific approach. The completely hypothesis-free approach makes results with previous study not much comparable, where the hypothesis was yet to look at priori selected genes. Since the original dogma of a genome-wide analysis is to identify so far undiscovered areas of the genome (and hopefully, in the future, cross-link them), novel results are very welcome having the power to open so far unasked questions, even if the interpretation might not be suddenly evident.

With respect to that, the regional plots (Figure 7.2 and Figure 7.3, right side) showed that the highly significant hits are mostly isolated probes instead of pointing to a specific region. This opens another interesting subject and it is the relationship across CpGs within the same area. We know that CpGs that are close have an influence each other and it is a common hypothesis that these specific variations related to differentially methylated areas are stronger than the ones related to lonely differentially methylated CpG sites. We will later discuss this topic attaching also a short literature review of developed methods that address the intra-CpG correlation.

Despite we identified several plausible links involving the genes we identified, we couldn't provide any specific causal function that may unveil how is happening that air pollution exposure influences adverse physical responses. This is only an observational study and further mechanistic in-depth analyses are required to better enlighten this possibility. However, another interesting implication of our results is the higher prevalence of positive associations that seems going in the opposite direction of the larger percentage of negative associations found by Zeilinger et al. in relationship with smoking status [Zeilinger et al., 2013]. Deepening the results that we found, we noticed that cg23276912 shares protein domain with gene *AHR*, as explained above, and its coefficient displayed increased methylation, hence, plausibly, gene silencing. The most striking CpG observed by Zeilinger et al. is located within *AHRR*, which represses *AHR* consistently with the hypothesis suggested by our results but possible implications need to be verified by future studies. Being at the dawn of the Epigenomics era, this is another example that shows how important it is to focus on each identified locus and discover its real functions.

The Regional Plots of the CpG sites identified in this study exhibit another pattern, despite we are not fully aware of the rationale behind. Selecting a restricted genetic window (200k bp) around a target (a highly significant probe), all the probes positively/negatively correlated with the target show a similar methylation value, that might be either hyper- or hypo-methylation and independently of the correlation direction, but consistent within them. For example, looking at the CpG sites with blue shade (meaning negative correlation) around cg02608596 in Figure 7.2, the height of the broken blue line (representing their average methylation value) is homogenous (showing hyper-methylation, in this specific example).

8.1.4 Strengths and Limitations

A great strength of the results presented here is that none of the studies that we considered is underpowered since all included at least 500 participants. We strongly believe that it was a necessity in order to have the sufficient power to identify and detect differentially methylated probes with such a little variability. Moreover going into the details of the model, we could take advantage of several positive elements that increased the research value. First of all, both methylation and $PM_{2.5}$ had been measured with the same methods and tools across the studies. Then, we had the potential to adjust for an extensive list of relevant confounders including temperature at the same time window of the $PM_{2.5}$, weather conditions and others that from sensitivity analysis and previous knowledge seem to be pre-requisite for the least possible biased estimation of associations with such ubiquitous exposures like, particulate matters (and more in general air pollution). Finally, performing a number of sensitivity analysis allowed the consideration of the a priori model as the most conservative for the estimates. Our results were not dependent on long-term exposure at residences and not influenced by potential outliers. Some limitations are also acknowledged. Air pollution had been measured from a single monitoring station since personal exposure data were not available. Considering that no coal power plant was in operation in proximity of the subjects and just a little percentage of them is addressed to live near a major road, let us focus on the temporal fluctuations of DNA methylation but not the spatial variability, forced to rely on ambient air pollution estimates. On support of the reliability of this method, we report an exposure-validation study enrolled in Boston, Massachusetts. Home indoor and outdoor concentrations on $PM_{2.5}$ were compared on 25 participants with ambient concentrations from a central monitoring station. They observed that ambient $PM_{2.5}$ was highly associated with both indoor and outdoor concentrations. Primarily Berkson-type measurement error [Zeger et al., 2000] will surely result in measurements error since a single site has been used in this study that implies a bias in the standard error but not in the effect size. Therefore, we considered only white participants limiting our results from plausible generalizations to other races. One more limitation comes from the type of analysis that we enrolled, focusing on the exposure on the days before the medical examination, no longitudinal/long-term effect could be deepened. Lastly, even if the Illumina 450k is the widest used platform with respect of number of successful probes analyzed, it doesn't yet cover the whole epigenome.

8.1.5 Summary

In conclusion we could observe and highlight a number of novel CpG sites associated with increased days and week of ambient fine particle exposure at different trailing averages. Sensitivity analysis also provides the evidence that these results were not influenced and cannot be attributed to long-term effects. We also observed, after meta-analysis, that the significance level tends to decrease extending the time window of exposure. Novel biological pathways have been identified according to the findings of this work and might be link air pollution exposure to several pathophysiological processes and health outcomes such as tumor development, inflammation stimuli, pulmonary disorders, glucose metabolism and also chromatin remodeling and gene regulation. Mechanistic analyses will be required in the future in order to establish the evidence of the effect of these epigenetic changes on the human biological system, including eventually lung cancer. Novel questions have been raised that solicit further research on environmental epigenomics.

8.2 Long-term Analysis

Results of long-term analysis on KORA F3 and F4 combined were not as clear as results on short- and mid-term exposure. Nevertheless, some suggestions and take-home messages can still be discussed. First of all, a long-term study focuses more on the spatial variability of the exposure and is determined by factors that are not much volatile during the time, such as traffic, green areas, major road or presence of massive sources of pollution. Relating to that, it is plausible to imagine that in a medium-sized city like Augsburg, the differences of the elements in play might not re-create the variability needed to display a change in DNA methylation related with conditions over extended time windows. On another side, the cross-sectional design of the analysis doesn't allow us to consider longitudinal variations of DNA methylation. This constitutes an important point since little is known regarding the behavior of DNA methylation, its plausible reversibility and most interestingly its interaction with aging. In western countries, air pollution effects on health are also limited compared to other exposures like smoking and produces just little effects through the time that constantly damage men's health but not massively. For this reason, genome-wide studies and specifically addressing DNA methylation might be very useful and contribute to enlighten the deep biological processes of diseases elicited by endogenous exposures. A genome-wide screening has the power to touch vast areas of the human genetic background, hence to try to merge a large amount of very small variations that, combined, might unlock a physiological outcome and many factors were already identified to play a major role into this system. Analytical tools keep evolving quickly. Considering the relative novelty represented by the Illumina 450k Beadchip and the yet elevated costs needed to enroll projects involving hundreds of sample, eventually longitudinally, and the necessity of a replication base, we consider our results a positive impulse for future analysis. Our DNA methylation measurements were only performed on whole blood samples and even the consideration of other types of tissues may provide a more specific picture of the effects of common environmental exposure on human health. Finally, epigenetic findings might also turn into therapeutic targets expanding the medical frontiers of treatments.

8.3 Lung Functions

As result of EWAs conducted on the American population-based study NAS, a large number of CpG sites have been detected as being Holm-significantly associated with smoking. Aim of the research is to study the possible association of some of the identified CpG sites with variations in lung functions, especially the results of a spirometry test FVC, FEV_1 and $FEF_{25\%-75\%}$. Across the different parameters and adjustment for smoking activity, nine CpG sites revealed an association with the respiratory functions; of them, seven with the rate of decline and the other two with the cross-sectional effect. Results of the selected 1339 CpG sites have been replicated in an independent German study: KORA. Three of the nine identified probes have also been found FDR significant in the replication study. These results confirmed the role of DNA methylation as possible mediator between smoking exposure and variation in lung functions parameters. One of the CpG sites that also replicate in KORA belongs to gene *AHRR*, a well-known gene in the field of smoking. If on one side it is surprising to observe it associated with decline in lung functions, on the other side, by adding a piece to the puzzle, it confirms that it is a key path between smoking and health consequences.

Another major question raised by these results concerns the CpG sites that were not found associated with lung function decline at either rate or cross-sectional level. Throughout our genome there are probes which DNA methylation is directly influenced by smoking but our empirical results haven't yet identified pathophysiologic factors that change in association with this epigenetic marker. Despite the brilliant idea of checking for lung functions parameters, smoking effects mediated by DNA methylation seem to be even more broadly extended.

Chapter 9

Inter-CpG Correlation

Sometimes, from analysis or studies, very simple issues related to the nature of the data and the information that we are handling, emerge from the reality and require a solution or, at least, an investigation. The issue discussed here is the correlation across CpG sites. Are CpG sites close each other following a similar behavior? Are they influencing each other? Are they correlated or do they follow a more complicated and intriguing pattern? Those are just a few exciting questions that arise looking at the epigenome and at the first results in the field of epigenomics and environmental epigenomics. Two published methods will be here discussed and at the end a naïve solution will be proposed.

9.1 A-Clustering

During late 2013 Sofer et al. proposed a computational method with the aim of detect areas of the genome that "exhibit common behavior" but might not be restricted to pre-specified or known methylation domains [Sofer et al., 2013]. The brilliant idea was to use correlation coefficient in order to scroll adjacent CpGs and cluster them according to certain parameters of distance and minimum correlation. Once a cluster have been identified, all the probes are tested together vs a specific exposure using a GEE model. Generalized Estimating Equations (commonly known as GEEs) is a special model, alternative to Mixed Models, developed by Liang and Zeger in 1986 in order to analyze correlated data like longitudinal data [Liang and Zeger, 1986]. The algorithm developed by Sofer et al., gives the possibility to choose the type of correlation that is preferred (Spearman or Pearson) and the way the distance between two clusters is estimated, hence how to join or not them. The analyst can also arbitrarily decide the minimum dimension (in number of probes) of the cluster. Even if the authors provided results from simulations conducted by variating the parameters, it is surely a strength the possibility given to the users to choose based on its own data and beliefs. This method represents a distinct novelty in the field of epigenomics. The dogma of A-clustering lies in the dual-step approach, first, the CpG sites are clustered together independently from the exposure but based on their "behavior", then associated with other factors. Once clusters are calculated, they constitute a kind of completely new dimension of the epigenome that can then be directly considered for further analysis and associated to plausible exposures. It looks like a new hierarchical level, at the base there are the CpGs, above there are the clusters. An interesting study that applied A-clustering has been recently published by Sen et al. [Sen et al., 2015], where they demonstrated that Pb-induced changes in DNA methylation can be modeled as co-regulated cluster. Simulations provided evidence of its increase effectiveness compared with Bump-Hunting, which will be discussed in the next paragraph.

A-clustering has also drawbacks. First of all an extension is needed in order to use all the power of an eventual longitudinal study. The algorithm includes currently only the steps needed for a cross-sectional study and a new strategy for repeated measurements is not yet proposed. However it is fair to remind that the whole code is written in R and is open access, so every user can simply modify and add the features he/she prefers. A second question that is raised is how can a user decide the best distance for his data? Are there biological elements that are needed to be taken into account? A proposed solution has been written in the paper but it is based on computational results on pilot data. A further question comes regarding replications. Clusters are strongly study-specific and the only way to replicate a result would be to ask another study to consider exactly the same grouping structure. This closes the way to a full replication since one study determines the first step for all the others, which would be used only from the second step. Finally, a limitation concerns the use of correlation. Figure 9.1 may help to better understand the problem. The histograms represent the distribution of the correlation between CpGs annotated next to each other. Then the distance was changed, to the second next CpGs, then the fifth, the tenth, the 100th and the 1000th. The distribution never changes across the different distances (as shown by the histograms, which density is then reported in the bottom graph, Figure 9.1) and this might suggest that correlation might not exactly address CpGs that exhibit the same behavior but rather randomness of the data. A second test was also performed. Three CpGs have been randomly selected and their correlation with any other CpG from the same chromosome has been estimated. The results showed a homogeneous histogram across all the estimated values, but more interestingly the same pattern across all the other CpGs, independently from the distance, shown in the horizontal axis of Figure 9.2. Results from Figure 9.1 and Figure 9.2 come from chromosome 5 as example but the pattern didn't change in other chromosomes (results not shown).

9.2 Bump-Hunting

Looking for differentially methylated regions (alternatively called DMR) can solicit several questions and one of them is: "do groups of CpG sites identified in a certain area,

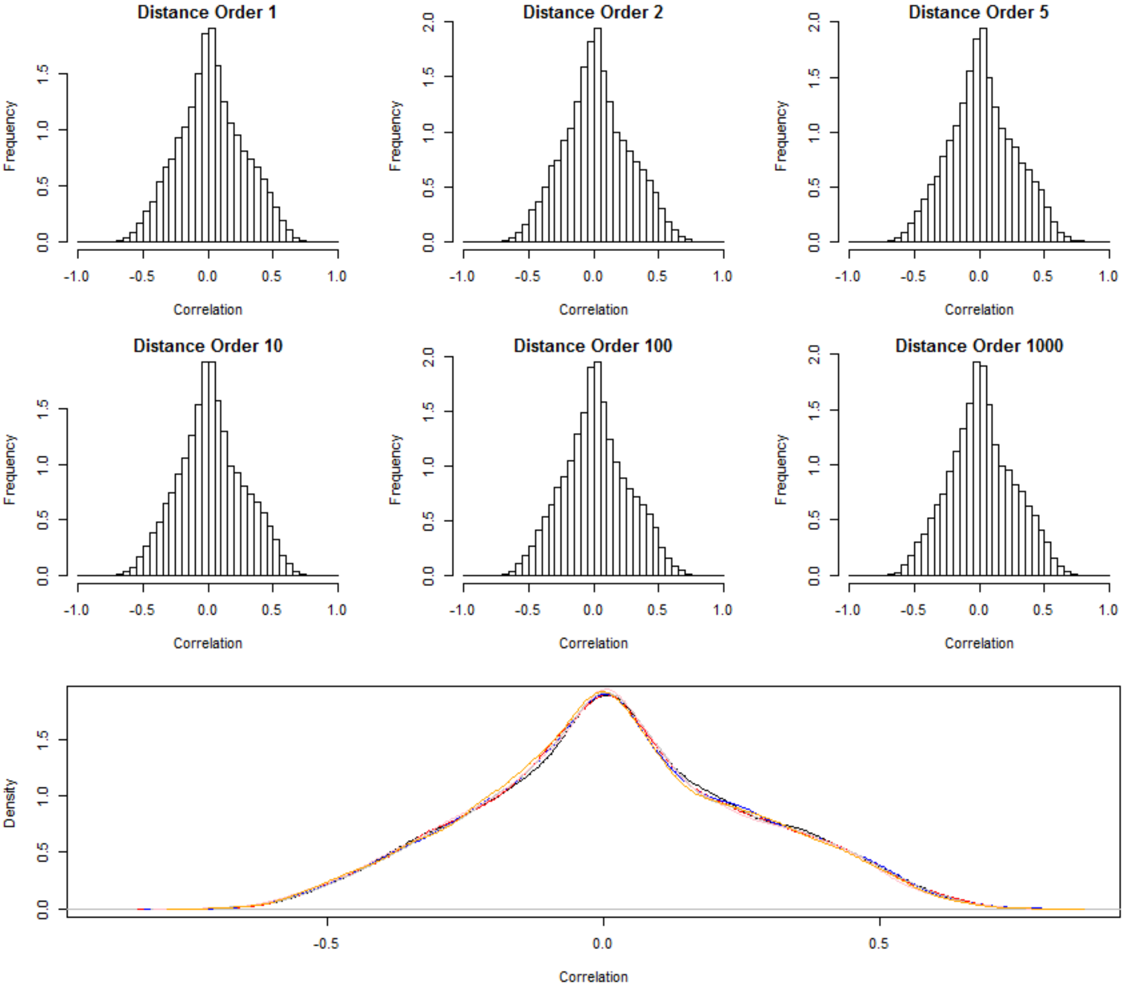

Figure 9.1: Distribution of the correlation at different distance orders across the CpGs. The bottom graph groups the density distribution of the previous six histograms.

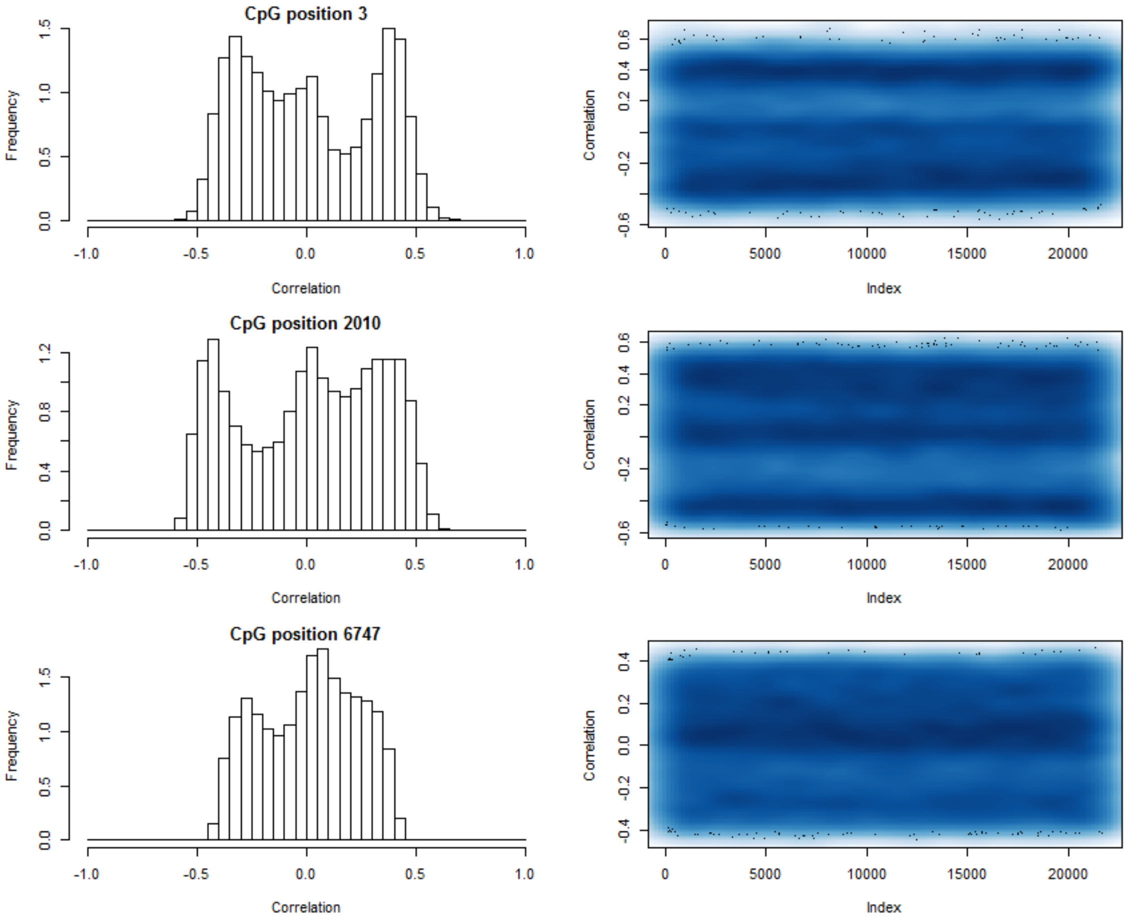

Figure 9.2: Distribution of the correlation coefficients of three random CpG sites.

all consistently influenced by a certain exposure/phenotype, exist?" This is the question that Jaffe et al. tried to answer when they applied the concept of Bump-Hunting to epigenomics [Jaffe et al., 2012]. Bump-hunting works similarly to A-clustering since it also perform a genome-screening, but after a EWA had already been performed. The basic idea is very simple; it starts taking a CpG with the corresponding estimates after the model performance, then moves to the next, and asks: is this second estimate similar to the first one? If yes then start to group them together and move to the next probe, is the new estimate similar to the previous ones? If yes, add it to the cluster, otherwise go on. And repeat it throughout the all genome. The analyst can also decide the dimension of the cluster and the other parameters according to his/her beliefs, the data and also the magnitude of the effect size in order to decide how much "similar" two probes/cluster must be to be joined. The authors propose a change in DNA methylation of 5%. The approach of Bump-Hunting, conversely than A-clustering, starts with regular EWAs and extrapolates clusters of probes after the they have already been associated with the exposure/phenotype but the focus is not to associate the probes that behave similarly, but the probes that behave similarly dependently on a specific outcome. In other words, clusters must be functional to the factor in study, that might be a disease (e.g.: tumoral vs healthy cells) or an environmental exposure (e.g.: high air pollution concentration). This constitutes a strength since this method is independent from the model and the study design, the only elements that are needed are the regression slopes per each locus. A final step, after the identification of a differentially methylated area is the performance of a permutation test in order to avoid the identification of random differentially methylated region.

Bump-Hunting is also not immune to drawbacks. The first one is the high reliability on estimates. If on one side it leaves the freedom of the model choice, on the other side regression coefficients have limitations in representing the completeness of the phenomenon in act. Smaller estimates might be more significant that larger ones and in general they are sensible to the mean, probes with an average value in the middle of the distribution might be more variables than probes at the extremes (hypo- or hyper-methylated). It seems difficult for Bump-Hunting to detect small but significant changes in DNA methylation, like the ones that are more common in environmental exposure, while it might be a very efficient technique to detect differentially methylated regions comparing strong outcomes like tumoral vs healthy cells. A second limitation is again constituted by the replication stage. It seems very unlikely that different study replicate the same regions. Like A-clustering, a limited but alternative solution would be to identify the regions in a first study and check results of other studies on the same regions.

9.3 A Separation between Hypo- and Hyper-methylated CpGs

The method we propose here represents a starting point that addresses one question: is it possible that hypo-, hemi- and hyper-methylated CpG sites behave differently? May it be more likely for a hypermethylated probe to increase its DNA methylation level than for a hypomethylated one with similar functional characteristics? This is why based on functional information a two-step probe clustering method was developed. The first step is to group CpGs annotated within the same gene and the second is to then subgroup them in hypo-, hemi- or hyper-methylated according to their average value. The threshold across these three categories would be empirically estimated based on the double peaked distribution of DNA methylation across the probes. After estimating the distribution of methylation for each subject, two derivatives have been estimated in order to detect the two flexes of the curve, after the first peak (approximately between 0.20 and 0.30) and prior to the second peak (between 0.65 and 0.75). The slope of the derivative can be arbitrarily decided and the median of the projections on the 0-1 methylation scale where the derivatives fall would be calculated across all the samples. The two medians, after the first peak and prior the second, would represent the threshold to classify the CpG sites: probes with mean lower than the first median are hypomethylated, the ones between the two medians are hemi-methylated and the ones with a mean greater than the second median would be classified hyper-methylated. We have now obtained clusters of CpGs belonging to the same gene and with a close DNA methylation output. After this selection step, all the CpGs in a group would be analyzed together with the outcome of interest with a mixed-effect model to take into account the inter-CpG correlation. In fact, even if we cannot quantify the amount of this co-variability, we expect the correlation inter-cluster to be not null: the rationale behind the method is that near CpG sites are surely not independent each other, as observed in the regional plots. This method had been tested on short- and mid-term $PM_{2.5}$ exposure and results show very low consistency between models run in the same genes but on clusters classified as different DNA methylation segment. The number of matches between significant CpGs across hypo-, hemi- and hyper-methylated clusters is presented in Figure 9.3. A strength of this method is the simplicity of the selection of the clusters. Often simplicity means lack of quality and possible repercussions are explained in the following lines, but in this case it also represents computational time saving. It is happening at two levels: one is the evident reduction of the models to run and the second one it that no screening is involved but a simpler sub-selection step: average and gene annotation are the two required parameters. Even if the mean might naïvely represent some probes, it is quite efficient for the vast majority of them which display a little variation. On the other side, an evident limitation of this method, consequence

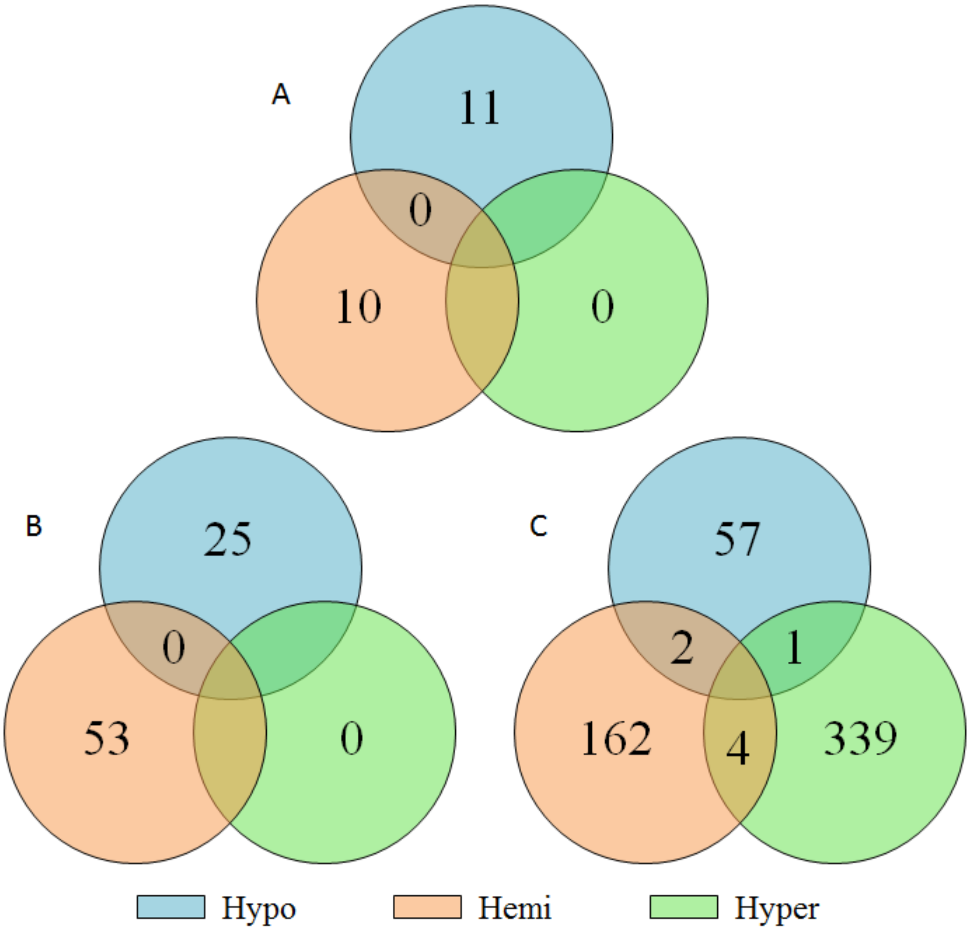

Figure 9.3: Number of significant clusters associated with $PM_{2.5}$ exposure at 2-day (Panel A), 7-day (B) and 28-day (C). Intersections show the number of matching clusters between different methylation segments. Blue: hypomethylated segment, Red: hemimethylated, Green: hypermethylated

of the simplicity of the cluster selection, is the fact that grouping a priori all the CpG sites annotated in a same gene and separating them only according to segment of the DNA methylation spectrum where their average falls, might provide imprecise results. How probes behave once they belong to the same gene is not yet clear and some CpGs will be clustered together without evidences of any association. Inter-CpGs influence has been observed in CpG Islands but evidences of a co-modulation pattern at gene level are not yet exhibited. On one side, the idea is to rely on prior information to cluster the probes before looking for an association with the outcome of interest. Therefore, across studies, this allows the selection of clusters that are not so different each other. In fact, we have observed high consistency of the DNA methylation values in different cohorts making plausible for a CpG to belong to the same methylation segment in different studies. However this problem might affect the borderline probes, those in a short around of the two thresholds.

9.4 Conclusion

The methods described in this chapter represent the first solutions in literature aimed to account for intra-CpG correlation. They all started from a different perspective, A-clustering is looking at CpG sites with a similar behavior, while Bump-Hunting is seeking CpG sites with a similar behavior depending on the outcome of interest and the method tested in this work simply addresses the possibility that hypo-, hemi- or hyper-methylated loci might behave differently. We also acknowledge that replication is still one of the biggest challenges that require attention in order to study the interplay that occur across different probes. The epigenome constitutes one of the biggest and most exciting novelties of the scientific research of the last couple of decades and identification of relevant differentially methylated regions (rather than simple CpG sites) may contribute to better enlighten the influence of DNA methylation in estimating the risk of common diseases and their interaction with both environment and genetics.

Part II

Socio-economic Status and Air Pollution

Chapter 10

The Problem of Confounding

The problem of confounding has been a real riddle for statisticians and epidemiologists and it will keep being so. Several techniques have been developed in order to control it, but the fact that it depends on the theme of interest, makes it always actual: confounding requires field knowledge. The reason why the mere development of techniques is not enough lies exactly in the fact that we have to know what confounds what. This is exactly why this project was started.

But what is confounding? Here is how the Oxford dictionary defines it: "mix up (something) with something else so that the individual elements become difficult to distinguish". In statistics its definition came out through other concepts like collapsibility [Whittemore, 1978] and comparability [Miettinen and Cook, 1981], inducing Nurminen to conclude: “Any attempt to clarify ’confounding’ in simple conceptual or statistical terms is destined to omit some important aspect on the topic” [Nurminen, 1997]. Showing this work epidemiological evidences, we can then refer to the definition of the Mosby’s Medical Dictionary [Inc., 2009]: 1) "Interference by a third variable so as to distort the association being studied between two other variables, because of a strong relationship with both of the other variables"; 2) "A relationship between two causal factors such that their individual contributions cannot be separated". Consequently, the analyst cannot be fully sure that the effect he/she observed depends on the variable he is studying. A simple example to emphasize it might be the following. Suppose the aim of the project is to study the effect of a new drug on a specific outcome and two samples are drawn, one for the drug and the second one for the placebo, but in the first sample there are only men and in the second only women. Whatever effect we do observe between the drug and the outcome, the conclusion cannot unambiguously declare whether it depends on the sex or the drug and the final estimate for the drug effect would clearly result biased. In other words, there is a third variable, called confounder, that alters the real estimation the researcher is interested in. The confounder, as evidenced by Mosby, must be associated with both the variables under observation and the aim of confounding correction techniques is then to delete the more possible sources of bias in the studies. A short list is here provided.

1. Adjustment. When a linear model is performed, a common way to adjust for confounding is to include all the possible confounders in the formula as covariates. Two strengths of this method are its easy application and a good degree of efficacy, however, on the other side when too many variables are included, the reduction of degrees of freedom must be compensated by a sufficient sample size.

2. Matching. It consists in forcing the comparability between the groups by including individuals with determinate characteristics. Despite the clear advantage that the matching structure create in separate the effects, a few limitations are also acknowledge such as the difficulty of selecting individuals following the rigid structure (that could alternatively lead to a residual confounding), the impossibility of measure the effect of the confounder (or matching factor) on the risk and the overmatching that leads to underrate the risk.

3. Stratification. It consists in evaluate the effect of interest intra-strata, which are homogenous according to the confounder. Compared to the matching, stratification is sensibly more informative, however a couple limitations must be taken into account. First, in order to consider several strata, some sub-strata might be affected by low sample size (and consequent difficult interpretation of stratum-specific estimates). Additionally, it only allows categorical variables to be used as correction factors.

4. Randomization. Aim of the randomization is to randomly label all the study participants into two groups which characteristics are expected to be homogenous. Randomization offers a good solution to control the confounder's effect, however it cannot be applied when we want to estimate the effect of a factor for a specific event and it is not ethically allowed to randomly label the patients for which there are solid evidences in terms of therapeutic evidence.

Purpose of this work is to provide better insights regarding the association between SES factors and air pollution exposure. Which is the role of the area based level? And the one of traffic?

Chapter 11

Preliminary Results

Preliminary analysis on KORA data has been performed with the aim of exploring the data and find possible associations. NO_2, NO_x, $PM_{2.5}$, PM_{10}, PM_{coarse} and noise were considered as exposure and smoking status, personal income, alcohol consumption, wine consumption, physical activity, BMI, educational level and occupational status as social factors. Here we briefly report some of the most interesting examples.
A multivariable statistical technique called Multiple Correspondence Analysis (MCA) has been applied. Shortly, through eigenvectors and some matrix transformations, MCA restructure the variability of the phenomenon in study creating a number of dimensions from the table of contingency. The variables must be categorical and each category of each variable receives a score for every dimension. Usually the dimensions that grab the highest amount of variability are studies together in order to look for categories that cluster together. We separated our continuous variables in three segments in order to apply MCA. Different combinations have been tested and the most interesting are reported in Figure 11.1 and Figure 11.2. Looking at results for NO_2, from Figure 11.1 on S4 we observe that noise and exposure group often together and from Figure 11.2 that never smokers seems to set along with low exposed people and middle-exposed with middle-class participants. Despite the clear potential of this method, these results couldn't guarantee much stability and the percentage of variability accounted by the first two variables was never very high (always below 35%).

One- and two-way ANOVA have been also performed associating at each exposure the social variables we wanted to look at Table 11.1.

Results are briefly summarized here. Most relevant results are reported for income and smoking. Regarding income, for NO_2 negative significant differences have been observed between the first and the second tertiles vs the third one indicating an increased exposure for wealthier subjects. The same pattern repeats also with NO_x, PM_{coarse} and noise. For PM_{10} a significant difference has been observed only for the first vs the third tertile of the distribution and for $PM_{2.5}$ no contrast looks statistically different. Regarding smoking, a positive significant contrast for NO_2 was observed between the first vs both the second

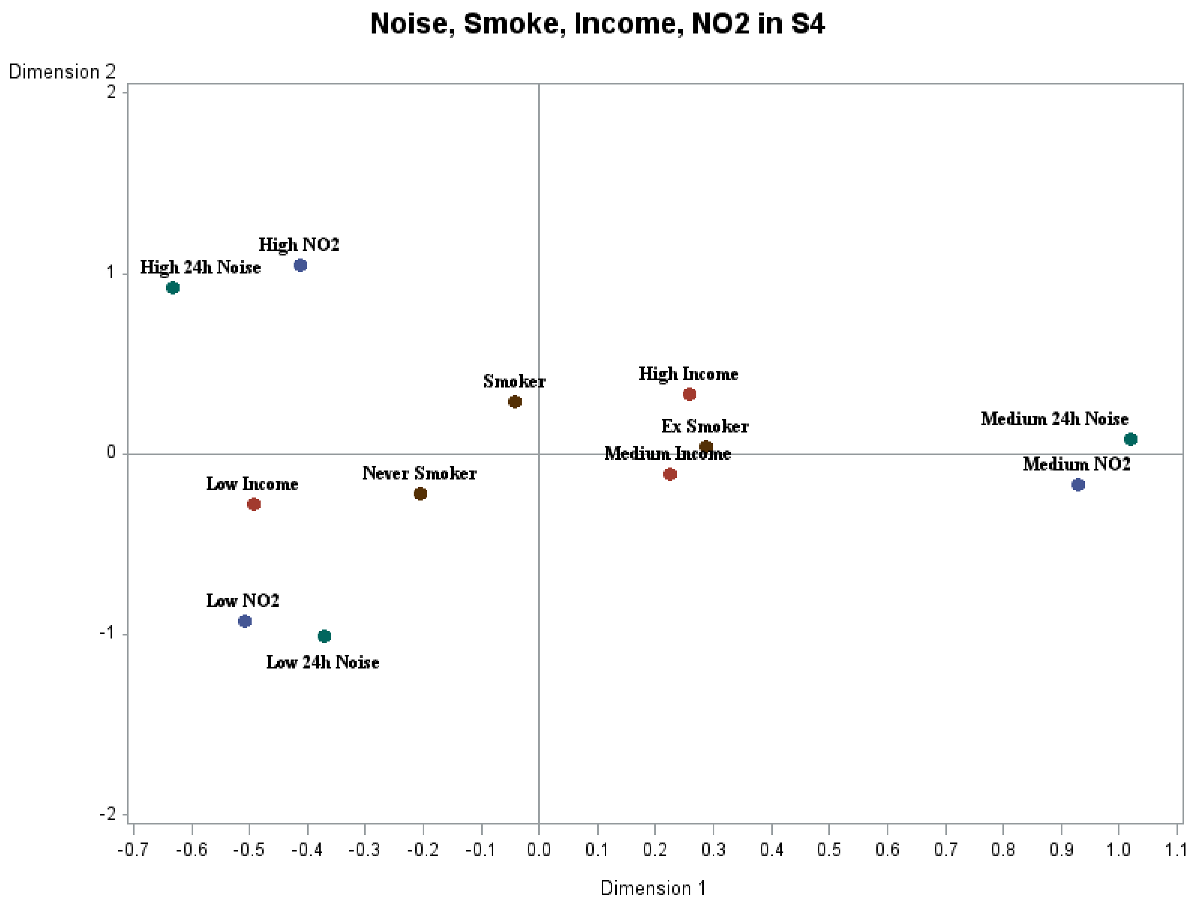

Figure 11.1: Multiple Correspondence Plot for NO_2, noise, smoke and income.

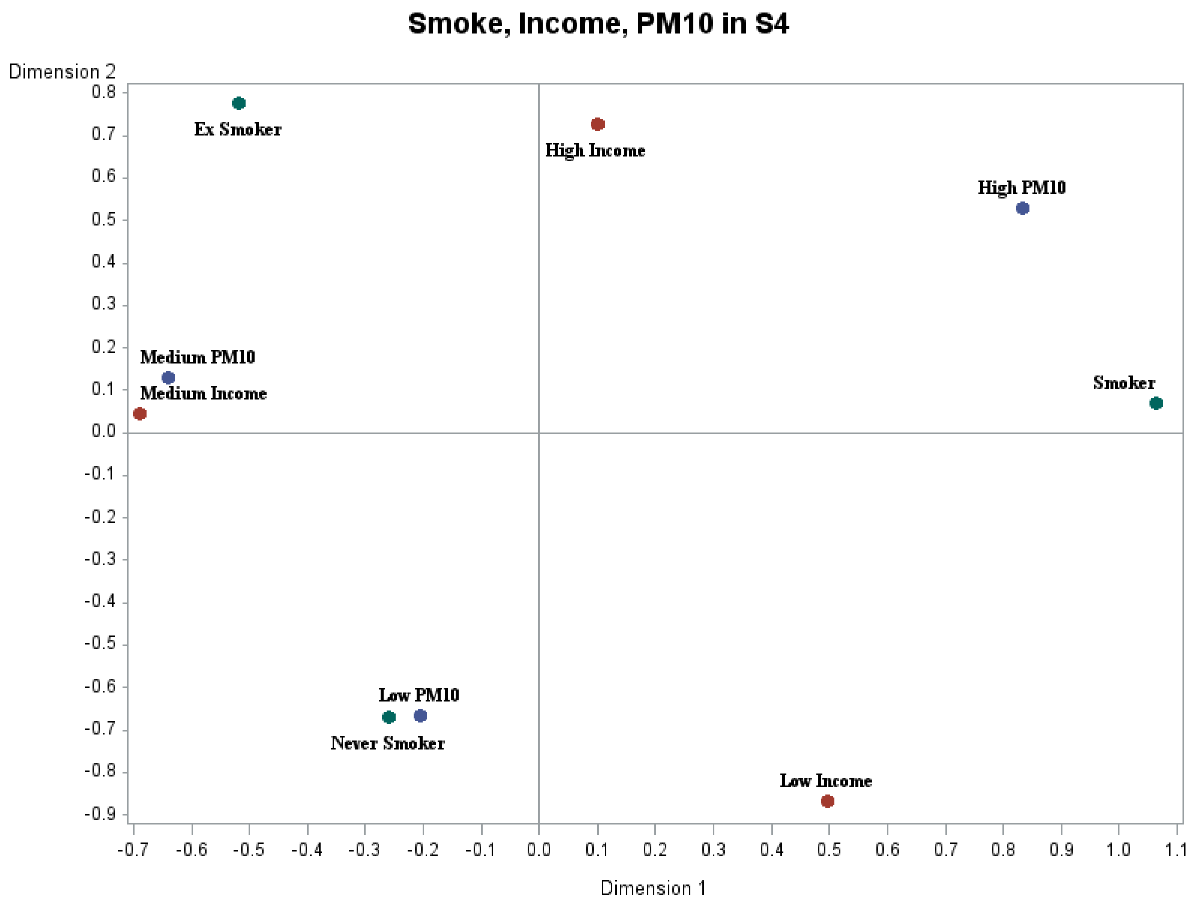

Figure 11.2: Multiple correspondence Plot for PM_{10}, smoking and income.

Exposure	**Income**		**SES**[a]		**Educational Level**		**Smoking Status**[b]	
	Contrast	**MD**[c]	**Contrast**	**MD**[c]	**Contrast**	**MD**[c]	**Contrast**	**MD**[c]
NO$_2$[d]	1 - 3	-0.77	1 - 3	-0.74	1 - 3	-0.62	1 - 3	0.8
	2 - 3	-0.7	2 - 3	-0.52			1 - 2	0.5
NO$_x$[e]	1 - 3	-1.32	2 - 3	-0.96	1 - 2	0.99		
	2 - 3	-1.38	1 - 3	-1.19			1 - 3	1.63
							2 - 3	0.63
PM$_{10}$[f]	1 - 3	-0.24					1 - 3	0.28
PM$_{coarse}$[f]	1 - 3	-0.11	1 - 3	-0.1			1 - 3	0.18
	2 - 3	-0.13						
PM$_{2.5}$[f]					1 - 3	0.16		
					1 - 3	0.11		
Noise	1 - 3	-0.74	1 - 3	-0.59			1 - 3	0.96
	2 - 3	-0.67						

[a] Combination of income and educational level

[b] 1= "smoker" , 2 = "former smoker", 3 = "never smoker"

[c] Mean Difference

[d] Nitrogen dioxide

[e] Mono-nitrogen oxides

[f] PM$_{10}$, PM$_{coarse}$, PM$_{2.5}$: particulate matter respectively smaller than 10 μm, between 2.5 and 10 μm and smaller than 2.5 μm in aerodynamic diameter

Table 11.1: Significant contrasts from the 1-way ANOVA matching pollutants level on the three-category social variables: Income, SES, Educational Level and Smoking Status.

and the third tertiles indicating increased NO_2 for smokers vs former and never smokers. For NO_x, significant differences have been observed in all the three contrasts, especially in smokers vs never smokers while for PM_{10}, PM_{coarse} and noise only between smokers and never smokers. The limit of these results is that they cannot be adjusted for other variables. The two-way ANOVA confirmed some of the previous indications and highlighted a few significant interactions: significant interaction between educational level and smoking, noise and income and noise and smoking for NO_2 and significant interaction between noise and smoking for PM_{coarse}.

Linear regression models have also been tested and stratified for occupational status. Figure 11.3 shows the regression results between NO_2 and income across the different occupational situations where employed and retired participants seemed to show a significant positive increase in NO_2 vs Income. Models were adjusted for educational level, smoking.

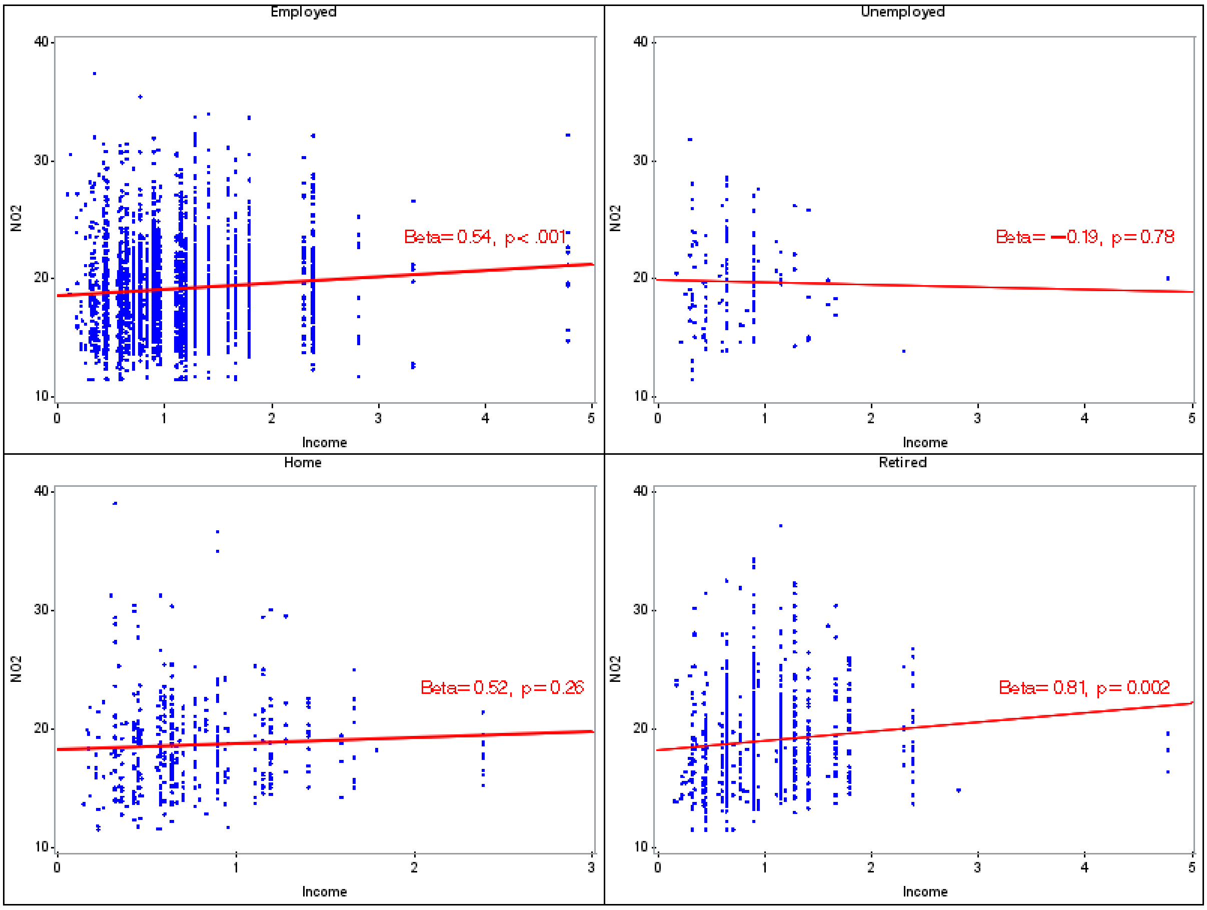

Figure 11.3: Linear regression results of NO_2 on income stratified by occupational status.

However, the most striking result was displayed applying mixed models. Performing the same model but adding a random effect for the zipcode of the participants the p-value for NO_2 vs income passed from <.0001 to 0.9647. This result was enough to firmly convince us that a normal linear regression might not be the most appropriate method to answer our questions and better understand the picture.

Chapter 12

An Alternative, the Path Model

Usual statistical techniques have been observed as incomplete tools to describe the phenomenon under consideration. In the Chapter 1, Introduction, we presented the concept of mediation as the phenomenon that happens when a third variable lies in between the two variable of interest and not only is necessary but emphasizes the effect. Aim of the path analysis is to separate the two sources of variability and once both direct and indirect effects are positive and different than zero it implies that both paths have to be taken into account.

Now, just a step back. Path Analysis was firstly introduced in 1965 by Wright with the aim of separate the origin of the co-variability between two variables [Wright, 1965]. It is based on Structural Equation Modeling (SEM), a mathematical multivariate system that allows the consideration of several equations at the same time. In this complex system, variables can be classified in two different ways: endogenous or exogenous. They recall the concept of dependent and independent variables in a usual linear regression model but on a multiple equation scale. Aim of the SEM is to list all the variables that depend on others and they will be named endogenous, irrespective of how many times they appear in other equations as the explanatory variable. On the other side, variables that never appear as dependent of any others are called exogenous. In Figure 12.1 the Directed Acyclic Graph (DAG) shows how the rationale of our analysis and evidences the only exogenous variable, the one with no income arrows: the SES factor. This is the definition that characterizes the system: a specific equation is required for any endogenous variable. The length of the equation system is determined by the number of endogenous variables (the ones with at least an income arrow in Figure 12.1). Despite several alternative schemes might be drawn, the design of the analysis must follow some priori information. Elements like temporal sequence or logical structure of the variables must be clarified at the planning stage otherwise it would be improper to even consider a mediation path. As said, the Path analysis allows the measure of the total effect between two variables that doesn't match with the real amount of association but provide most of the relevant information. The total effect (TE) is, in fact, divided into the direct (DE) and indirect effect (IE) that

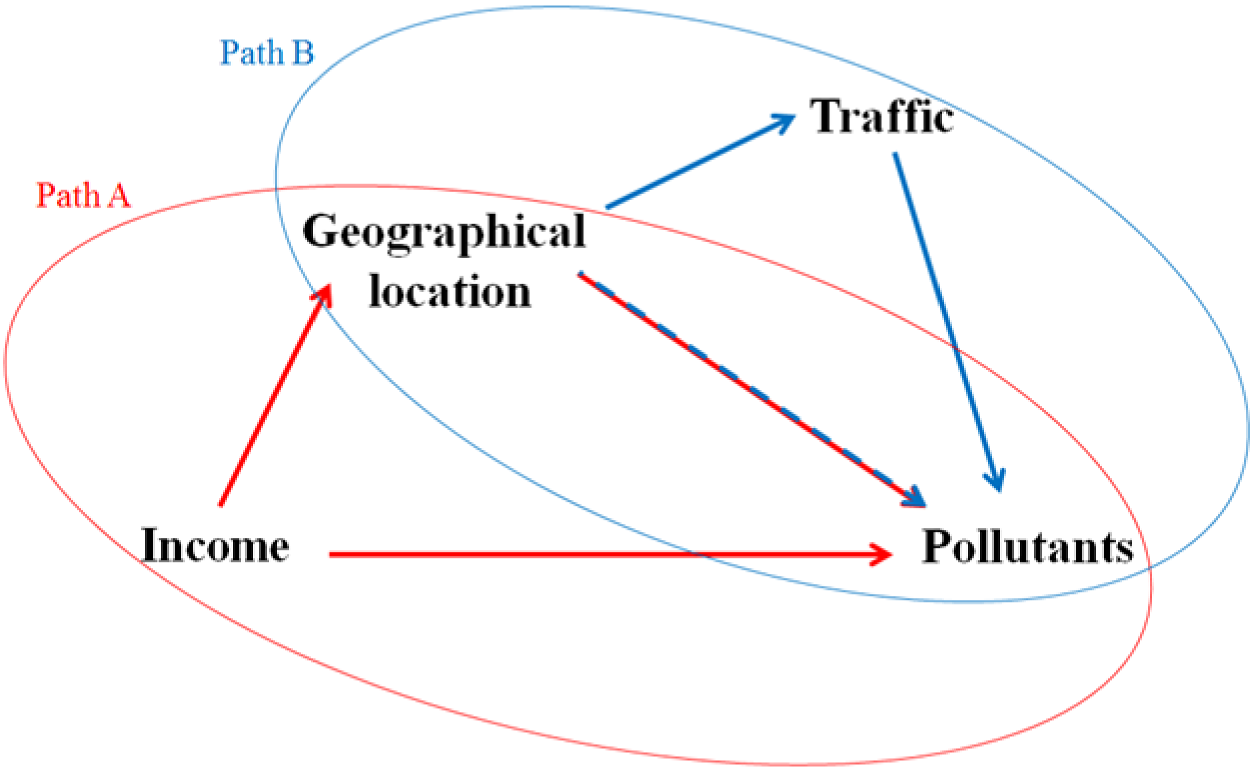

Figure 12.1: DAG of the theoretical mediation scheme. Highlighted, the two paths we considered.

coincide with the most useful information regarding the association (due to identified factors) and leaving spurious and unmeasured effects out. In other words, the total effect focuses on what can be measured and kept under control. The ratio DE/TE is then called non-mediated proportion while IE/TE represents the mediated proportion. The DE can be interpreted as the effect solely due to the pure association between two variables, while the IE (which is on turn built on two direct effects) measures the fraction of that association mediated by a third factor. Both DE and IE connect the same two variables, but for the IE there is also a mediator in play.

Figure 12.1 highlights two paths (A and B), one with number of households as mediator and one with traffic as mediator. The SES factor has been expressed at both personal and area based level and results were replicated in two European studies, one in the North (Helsinki-Turku, Finland) and one in the South (Rome, Italy). In Augsburg and Helsinki, the percentage of low income households in a 5km buffer was used as area-level SES while in Rome it was used an index of socio-economic position at census-block level resulting from a factor analysis that included several census variables such as employment and house property. As air pollution exposure the long-term parameters for NO_2, $PM_{2.5}$, and PM_{10} (estimated via LUR models as explained in Chapter 3) were included as outcomes. Households' density was used as mediator for path A and variable of interest for path B. Two traffic indices were used as mediators for path B: the intensity in the nearest road

and the load in the major roads within a 100m buffer. Both paths have been obtained as result of a single data step.
Sensitivity analysis was also performed. At first, all the path models (combinations of exposure and type of income) were performed in subsamples of employees. Besides, other social variables (BMI, smoking, alcohol consumption, wine consumption and educational year) were also included in the picture in different plausible positions in order to assess and evaluate their impact on the association between SES and pollutants. PROC CALIS from statistical software SAS (V9.3, SAS Institute) was used to perform all the analyses.
An alternative novel method to test for mediation schemes have been developed and published by Valeri and VanderWeele in 2013 [Valeri and Vanderweele, 2013]. Through a counterfactual approach, they extended the previous concepts developed by Baron & Kenny [Baron and Kenny, 1986]. The new method was denominated "mediation analysis". Once a variable have been hypothesized as causative on another variable, output of the mediation analysis is the total effect of the first variable on the second. The total effect would finally be split in natural (and controlled) direct effect and natural indirect effect with the aim of estimate the proportion mediated and non-mediated. Following the same theoretical structure identified for the path model and adjusting for educational years, intensity of traffic in the nearest road and load of traffic in a major road in a 100m buffer around the participants' home addresses, the mediation analysis have been applied to our data.

12.1 The Replication Studies

Results of Path model in KORA suggested the interest towards other European cities. Thanks to the collaboration with the groups of Francesco Forastiere in Rome (Italy) and Timo Lanki in Kuopio (Finland), our proposal was tested in other two realities. A short introduction to the two studies follows.

12.1.1 SIDRIA

The population-based study SIDRIA was conducted in Italy as extension of the ISAAC initiative. It represents a worldwide survey with the aim of determine variations in prevalence of symptoms of asthma, rhinitis and atopic eczema. Between October 1994 and March 1995, in eight medical centers in Northern and Central Italy, participants were recruited following a cross-sectional scheme. Standardized questionnaires have been filled out. For the purpose of this project, we reduced the sample to the 10,550 subjects collected in Rome between years [Anon., 1997]. Rome is located in the Tyrrhenian coast of the Italian peninsula and covers an area of 1,285 km^2 with approximately ~2,6 million inhabitants.

12.1.2 FINRISK

FINRISK is a national Finnish study was initiated in 1972 with the objective of collecting information on risk factors of chronic non communicable diseases. A stratified random sample of population aged between 25 and 64 years old has been drawn from the Population Information System every five years since then. Unique personal identity codes have been assigned to data from different population registers. For this analysis, 9,317 study participants have been considered from four cross-sectional surveys (1992, 1997, 2002 and 2007) and two study areas (Helsinki/Vantaa and Turku/Loimaa areas) located in South Finland that count for around 1.0 million inhabitants over ~2,900 km2 [Vartiainen et al., 2010].

Chapter 13

Results

Characteristics of the participants for the three studies are described in Table 13.1 and correlation coefficients between pollutants and SES factors presented in Table 13.2. The following variables resulted balanced across the studies: age (slightly lower variability in SIDRIA), BMI (not available in SIDRIA) and proportion of males. On the other side, SIDRIA and FINRISK showed a higher percentage of employees (70% and 78% respectively) than in KORA (61%). Traffic in Rome was observed more intense both in the nearest road and on major roads than in Augsburg and Helsinki. Lastly, the smallest and least dense city (lowest household density) was Augsburg.

Results from path modeling exhibit both similarities and differences across the three cohorts. Focusing on the link between SES and air pollution mediated by household density (Path A), the Goodness of Fit index resulted averagely in 92%. Cases where DE and IE displayed opposite direction where classified as non-mediated. Results with percentage of low income people as SES factor follows Table 13.3.

A larger TE was observed only for NO_2 in KORA (0.55) while the others emerged in a low to moderated range. DE and IE took then respectively 44% (proportion non-mediated) and 56% (proportion mediated) of the TE differently than the other two studies where the DE for NO_2 were too small and non-significant and, unexpectedly, the TE were negative (-0.09 in SIDRIA and -0.03 in FINRISK). $PM_{2.5}$ was the pollutant that displayed the most unpredictable and surprising results and its behavior was different in each city. In both KORA and SIDRIA the TE resulted almost null but while in SIDRIA both DE and IE were close to zero, in KORA, both DE and IE reached a similar moderate intensity but at opposite direction, exhibiting path coefficients of -0.18 and 0.17 respectively. In FINRISK, instead, we observed an even more different result since almost all the TE (-0.07) come from the DE (-0.06) taking the 84% of the co-variability. The IE, despite resulting rather small (-0.01), was still significant. Concerning PM_{10}, the pattern observed in KORA was similar (proportion mediated: 0.59%) than NO_2 but at lower intensities with a path coefficient for TE of 0.25. And while in SIDRIA the TE was zero (with DE and IE nullifying each other: 0.03 and -0.03), in FINRISK, interestingly, the proportion mediated

Study	**KORA** **(n=4,237)** **1999-2001**	**SIDRIA** **(n=10,550)** **1994-1995**	**FINRISK** **(n=9,317)** **1992/1997/2002/2007**
N(%)			
Males	2,080 (49.2)	4,970 (47.1)	4,417 (47.4)
Employed	2,570 (60.7)	7,339 (70.4)	7,282 (78.2)
Mean ± SD (5%, 50%, 95%)			
Age, years	49.2 ± 13.9 (28.0, 49.0, 71.0)	43.7 ± 6.0 (34.0, 43.0, 54.0)	48.2 ± 13.4 (27.0, 48.0, 70.0)
BMI[a], kg/m^2	27.2 ± 4.7 (20.6, 26.6, 35.7)	*	26.3 ± 4.5 (20.2, 25.8, 34.9)
Traffic Near, veh/d[b]	1,735.5 ± 3,778.5 (500.0, 500.0, 8,922.0)	2,942.8 ± 6,715.0 (500.0, 500.0, 14,928.0)	1,920.7 ± 4,697.3 (50.0, 250.0, 9,011.0)
Traffic Major, 1000*veh/d*m[c]	482.2 ± 1,156.1 (0.0, 0.0, 3,009.1)	1,399.9 ± 2,830.2 (0.0, 0.0, 6,878.3)	734.6 ± 16,125.2 (0.0, 0.0, 3,770.6)
Household Density, #/1km[d]	1,585.2 ± 1,498.7 (67.0, 1,191.0, 5,132.0)	59.5 ± 32.5 (12.4, 54.9, 112.6)	6,214.1 ± 4,944.9 (1,211.0, 4,747.8, 16,223.2)
Income, €[e]	1,005.6 ± 561.9 (341.8, 894.8, 2,300.8)	*	23,646.6 ± 6,412.2 (16,361.0, 21,596.0, 33,709.0)
Low Income Households, %[f]	29.6 ± 18.1 (2.9, 24.9, 51.9)	0.19 ± 2.0 (-2.1, -0.3, 4.0)	36.6 ± 7.3 (27.6, 35.9, 48.4)
Education (9/13/17 years), N (%)	*	4,691, 4,139, 1,544 (45.2, 39.9, 14.9)	*
Education, years	11.6 ± 2.6 (8.0, 11.0, 17.0)	*	12.6 ± 4.1 (7.0, 12.0, 19.0)

[a] BMI: Body Mass Index

[b] Traffic intensity on the nearest road

[c] Traffic load on major roads within 100m of the residence

[d] In SIDRIA it was scaled by the number of rooms and adjusted accounting for the area of both census block and buffer

[e] Monthly in KORA, yearly in FINRISK

[f] Low Income threshold: 1250€, buffer: 5km

* Data not available

Table 13.1: Descriptive statistics of the study participants in the studies KORA, Augsburg, Germany, SIDRIA, Rome, Italy and FINRISK, Helsinki/Turku, Finland.

		Mean ± SD (5%, 50%, 95%)	**Correlation**				
			Traffic Near	**Traffic Major**	**Household Density**	**Annual Income**	**Area Level Low Income**
NO_2[a]	KORA	19.01 ± 3.9 (13.79, 18.53, 26.28)	0.35	0.46	0.61	0.08	0.57
	SIDRIA	38.99 ± 9.01 (25.48, 38.00, 56.13)	0.42	0.53	0.55	NA	-0.13
	FINRISK	15.31 ± 4.89 (8.64, 14.48, 24.22)	0.22	0.39	0.57	0.12	-0.03
$PM_{2.5}$[b]	KORA	13.65 ± 0.89 (12.51, 13.49, 15.38)	0.30	0.29	0.13	-0.01	0.03
	SIDRIA	19.35 ± 1.82 (17.3, 18.91, 23.32)	0.68	0.61	0.26	NA	-0.09
	FINRISK	7.7 ± 1.13 (5.59, 7.92, 9.07)	0.13	0.34	0.30	0.05	-0.07
PM_{10}[c]	KORA	20.46 ± 2.42 (16.49, 20.52, 24.57)	0.17	0.25	0.29	0.04	0.27
	SIDRIA	36.44 ± 4.99 (31.39, 35.20, 47.1)	0.64	0.53	0.18	NA	-0.06
	FINRISK	14.07 ± 3.09 (9.59, 13.62, 19.91)	0.23	0.41	0.68	0.07	-0.11

[a] Nitrogen dioxide
[b] Particulate matter smaller than 2.5 μm in aerodynamic diameter
[c] Particulate matter smaller than 10 μm in aerodynamic diameter

Table 13.2: Distribution and Spearman correlation with covariates NO_2, $PM_{2.5}$, PM_{10} in the three studies KORA, SIDRIA, FINRISK.

		KORA	**SIDRIA**	**FINRISK**
Path A: SES → Household Density → Exposure				
% of Low Income Households[a]	**NO$_2$**	TE: 0.55**	TE: -0.09**	TE: -0.03*
		DE: 0.24** IE: 0.31**	DE: -0.01* IE: -0.09**	DE: 0.00 IE: -0.03**
		Prop. Mediated: 56%	Prop. Mediated: 92%	Prop. Mediated: 100%
	PM$_{2.5}$	TE: 0.00	TE: -0.03*	TE: -0.07**
		DE: -0.18** IE: 0.17**	DE: 0.01 IE: -0.04**	DE: -0.06** IE: -0.01**
		Prop. Mediated: ‡	Prop. Mediated: ‡	Prop. Mediated: 16%
	PM$_{10}$	TE: 0.25**	TE: 0.00	TE: -0.06**
		DE: 0.10** IE: 0.15**	DE: 0.03** IE: -0.03**	DE: -0.03** IE: -0.03**
		Prop. Mediated: 59%	Prop. Mediated: ‡	Prop. Mediated: 49%
Path B: Household Density → Traffic → Exposure				
Adjusted with % of Low Income Households[a]	**NO$_2$**	TE: 0.46**	TE: 0.56**	TE: 0.66**
		DE: 0.38** IE: 0.08**	DE: 0.49** IE: 0.07**	DE: 0.57** IE: 0.08**
		Prop. Mediated: 18%	Prop. Mediated: 13%	Prop. Mediated: 13%
	PM$_{2.5}$	TE: 0.26**	TE: 0.28**	TE: 0.26**
		DE: 0.18** IE: 0.08**	DE: 0.17** IE: 0.11**	DE: 0.18** IE: 0.08**
		Prop. Mediated: 31%	Prop. Mediated: 39%	Prop. Mediated: 31%
	PM$_{10}$	TE: 0.22**	TE: 0.19**	TE: -0.06**
		DE: 0.17** IE: 0.05**	DE: 0.10** IE: 0.09**	DE: 0.60** IE: 0.09**
		Prop. Mediated: 21%	Prop. Mediated: 48%	Prop. Mediated: 13%

[a] Low Income threshold: 1250€, buffer: 5km

‡ No mediation

** p-value <0.01

* p-value <0.05

Table 13.3: Path Coefficients for % of Low Income Households on the three studies KORA, SIDRIA and FINRISK using % of low income households as SES factor (TE= Total Effect, DE= Direct Effect, IE= Indirect Effect).

resulted almost half of the TE (49%). Despite the low intensities of the coefficients, the DE and IE quite perfectly split the TE differently than the other coefficients in Finland. To conclude, larger coefficients were observed in KORA for NO_2 and PM_{10}, also displaying a clear separation between DE and IE. In SIDRIA the IE was generally higher and in FINRISK all the TE resulted consistently negative.

Personal income was then applied to the same models as SES factor (data not available in SIDRIA) Table 13.4. Results differed quite substantially from the percentage of low income people. In KORA, despite an overall decrease of TE and in general of all the coefficients, DE suffered the strongest reduction in both NO_2 and PM_{10}. For both pollutants, path coefficients for DE turned non-significant leaving a higher percentage to the proportion mediated. Regarding $PM_{2.5}$ all the coefficients decreased drastically close to the zero. In FINRISK, instead, at least for NO_2 and PM_{10} the TE turned into positive values. The most striking results from the Finnish cohort comes from the NO_2, where a TE of 0.11 was split in a DE of 0.05 and IE of 0.06 resulting in a 56% of proportion mediated. However, for both $PM_{2.5}$ and PM_{10}, DE and IE exhibited opposite direction excluding the possibility of mediation.

Switching to Path B, where traffic was considered as possible mediator in between household density and air pollution exposure, a general trend of overall larger coefficients was observed (Table 13.3, Table 13.4). Even if the focus wasn't directly on SES, results were affected by the SES variable included in the model. Looking at results obtained when percentage of low income people was included in the path analysis, NO_2 showed a high degree of consistency across the three studies. The TE resulted between 0.46 in KORA and 0.66 in FINRISK with large percentages of proportion non-mediated (82% in KORA, 87% in SIDRIA and 86% in FINRISK). High stability, but at more modest values, was also observed for $PM_{2.5}$. In KORA and FINRISK, TE was 0.26 and 69% was the proportion non-mediated, while in SIDRIA the TE settled at 0.28 with 61% proportion non-mediated. Larger differences across the geographical location were instead observed for PM_{10}. In Finland the TE resulted very high with a value of 0.69 where only the 13% was the proportion mediated. More modest coefficients were observed in Germany (0.22) and Italy (0.19) displaying, however, different pattern in terms of mediation. If in KORA, the proportion mediated settled at 21%, in Rome was more than doubled with a value of 48%. Differently than Path A, we conclude that all the coefficients resulted positive and mediation structure was observed (at different extent) across all the pollutants and studies.

Results of sensitivity analysis on a subset of employees were mostly consistent with results in all subjects. The influence of the other social variables that were tested (wine and alcohol consumption, BMI, educational years) resulted poor as well as their own association with either SES or air pollution exposure.

Finally, as shown in Table 13.5, results from mediation analysis developed with Valeri

		KORA	**FINRISK**
Path A: SES → Household Density → Exposure			
Personal Income[a]	**NO_2**	TE: 0.08** DE: 0.02 IE: 0.06** Prop. Mediated: 75%	TE: 0.11** DE: 0.05** IE: 0.06** Prop. Mediated: 54%
	$PM_{2.5}$	TE: -0.01 DE: -0.02 IE: 0.01** Prop. Mediated: ‡	TE: -0.04** DE: -0.06** IE: 0.02** Prop. Mediated: ‡
	PM_{10}	TE: 0.03* DE: 0.00 IE: 0.03** Prop. Mediated: 100%	TE: 0.04** DE: -0.02** IE: 0.06** Prop. Mediated: ‡
Path B: Household Density → Traffic → Exposure			
Adjusted with Personal Income[a]	**NO_2**	TE: 0.62** DE: 0.53** IE: 0.09** Prop. Mediated: 15%	TE: 0.65** DE: 0.57** IE: 0.09** Prop. Mediated: 14%
	$PM_{2.5}$	TE: 0.14** DE: 0.07** IE: 0.08** Prop. Mediated: 57%	TE: 0.27** DE: 0.19** IE: 0.08** Prop. Mediated: 30%
	PM_{10}	TE: 0.29** DE: 0.24** IE: 0.05** Prop. Mediated: 17%	TE: 0.69** DE: 0.61** IE: 0.09** Prop. Mediated: 13%

[a] Monthly in KORA, yearly in FINRISK, not available in SIDRIA

‡ No mediation

** p-value <0.01

* p-value <0.05

Table 13.4: Path Coefficients for Personal Income on the three studies KORA, SIDRIA and FINRISK using personal income as SES factor (TE= Total Effect, DE= Direct Effect, IE= Indirect Effect).

and VanderWeele method applied on Path A with percentage of how income households displayed a considerable consistency with results of path model in terms of proportion mediated and non-mediated.

	KORA	**SIDRIA**	**FINRISK**
SES → Household Density → Exposure			
NO_2	TE: 0.91 DE: 0.47 IE: 0.45 Prop. Mediated: 49%	TE: -0.08 DE: 0.01 IE: -0.09 Prop. Mediated: ‡	TE: 0.00 DE: 0.04 IE: -0.04 Prop. Mediated: ‡
$PM_{2.5}$	TE: -0.17 DE: -0.33 IE: 0.16 Prop. Mediated: ‡	TE: -0.01 DE: 0.02 IE: -0.03 Prop. Mediated: ‡	TE: -0.04 DE: -0.01 IE: -0.03 Prop. Mediated: 16%
PM_{10}	TE: 0.37 DE: 0.21 IE: 0.16 Prop. Mediated: 43%	TE: 0.03 DE: 0.05 IE: -0.02 Prop. Mediated: ‡	TE: -0.06 DE: 0.00 IE: -0.06 Prop. Mediated: 100%

‡ No mediation

Table 13.5: Results of mediation analysis on the three studies KORA, SIDRIA and FINRISK using % of low income households as SES factor.

Chapter 14

Discussion and Conclusion

We collected data from three independent population-based studies in Europe, one South (SIDRIA, Rome, Italy), one Central (KORA, Augsburg, Germany) and one North (FINRISK, Helsinki/Turku, Finland). Objective of the analysis was to study the association between individual and area-level SES factors and identifying and quantifying the influential factors. Preliminary analysis showed an inconsistency of the results and more importantly evidenced a lack of pertinence in the methods. This is why it was hypothesized that the household density (indicator of geographic area) may mediate the link between pollutants (NO_2, $PM_{2.5}$, PM_{10} were considered in this analysis) and SES factors. An alternative statistical technique has been proposed: path analysis, which is rather uncommon within the field of environmental epidemiology. Interestingly, our results unveiled differences across the two levels, exposures and cities. Nevertheless, the hypothesized scheme seems to work in several circumstances but not in all the cases and all the pollutants. However, path model, splitting the total effect in direct and indirect effect, was proved as a more consistent method to try to explain the complexity of the phenomenon in act.
At first, it was acknowledged that the TE regarding NO_2 differed on Path A across the studies. The higher value observed in Augsburg may be explained by the stratification due to the different SES levels across the people and that this pattern is more closely associated with the spatial distribution of NO_2 Table 13.3: Augsburg is the least dense city of the three considered. What can this structure suggest? A possibility might be that, displaying both DE and IE a moderate association in KORA, despite the trend of people with close SES is to group together, their choice keep being variable. This variability is relevant and the positive association may reflect the higher NO_2 exposure in subjects belonging to areas with an increased percentage of low income people. The mediation structure is confirmed by the IE which reflects the association between the 1km density and the SES, reckoning for one part of the variability. Conversely, the sign of the TE in the Finnish and Italian cohorts resulted negative (actually in Helsinki almost null) and it might reflect the weight of a high-SES part of the inhabitants that keeps living in the city

center.

At a second stage, we observed a very different pattern for $PM_{2.5}$. While in Augsburg path coefficients displayed a surprising structure, in Rome and Helsinki they showed low negative intensity. In KORA, DE and IE resulted equal in absolute value but opposite in the sign, presenting a null TE. Trying to get to the bottom of these results, it can be seen that higher proportions of low-income people reflected a dual patter: on one side they are associated with lower pollution concentrations but on the other side are also positively associated with household density, which, in turn, correlates positively with $PM_{2.5}$. We can conclude that lower SES levels may be positively associated with air pollution only through the mediation of the household density. The path model, separating the two effects, could give us this overview of the picture, but at the same time we have to acknowledge that the mediation structure doesn't properly fit this scenario.

Thirdly, the behavior of PM_{10} in KORA was similar to NO_2 but displaying lower coefficients. However, the substantial difference between the two pollutants is their source: NO_2 is more related to traffic while PM_{10} reflects the local road traffic emissions as well as the long-range transported particulate air pollution.

Keeping the scheme of Path A, percentage of low income people was replaced by personal income as SES factor where available (KORA and FINRISK). Results changed drastically. In Augsburg we assisted a consistent decrease of the DE towards the null value across the pollutants. We assume then, that individual income is not as appropriate as the area-based indicator to describe the variability of the path we are interested in. Personal income and low income percentage in KORA display a 0.11 correlation coefficient, considerably lower than the -0.72 observed in FINRISK. While on one side the difference between these correlation values surely explains the substantial decrease of TE in KORA, on the other side it suggests that in FINRISK the population is stratified by SES level but it doesn't match with the uneven spatial distribution of the pollutants. In fact, the TE resulted to be pretty low.

Evidence of a positive association between low SES and increased pollutant concentration has already been assessed [Gray et al., 2014, Hajat et al., 2013]. From the analysis of this work, consistent results have been observed in KORA but not in SIDRIA and FINRISK (TE was negative). Despite the low intensity of the path coefficients, subjects that can afford to live in the city center are only the more affluent. This might come as a consequence of the fact that the percentage of households with income below a certain value was indicating low-SES. Therefore the distribution of high-SES people cannot be fully addressed using this approach and it cannot be inferred as inverse of the low-income individuals. Finally, the differences observed in megacities of developing countries cannot be revealed from our results, since additionally, they also suggest a city-specific pattern.

Looking at the results for path B, including traffic as element in between household density and exposures, a clear patter can be suddenly detected across the cities: a consistent

increment of the DE that took the majority of the TE. Additionally, the TE for NO_2 resulted always larger than for $PM_{2.5}$ and PM_{10}, except in Finland. We might conclude that even if traffic was a major element, its effect wasn't carrying more than 20% for NO_2 and at most 40% and 50% for $PM_{2.5}$ and PM_{10} respectively, of the TE. Consider other sources of emissions can surely help to understand more extensively the whole path.
Within the framework of the ESCAPE study, the confounder selection followed a priori staged approach. A conceptualization of the main model for each outcome was assessed and area-based SES was included later as sensitivity analysis [Cesaroni et al., 2014, Lanki et al., 2015, Stafoggia et al., 2014]. The impact of area-based SES on air pollution estimates observed in meta-analysis was very low but sometimes it implicated a reduction in heterogeneity. Results of the analysis performed in this work showed a clear city-specific pattern of area level SES and it seems feasible to expect, despite producing similar estimates in meta-analysis, a reduction of between-city heterogeneity (co-variability) by adjusting for area-based SES. From our results, the link between area level SES and particle concentration might be mediated or at least partially explained by the density of the households within a certain area. Even if lies in between in the causal association, it cannot be assumed as responsible of the whole conjunct variability.

14.1 Strengths and Limitations

The possibility offered by the working group of Francesco Forastiere in Rome and Timo Lanki in Koupio to replicate our results in a southern and northern European city clearly strengthen the validity of this work. Secondly, the methodology behind the measurement and the assessment of the exposures followed a shared and verified approach. Finally, geographical variables were comparable across the studies and estimated on the same buffer or spatial grid. Going to limitations, not all the relevant elements, such as residential wood burning, involved in the link between SES, population density and air pollution could be included. In the second place, it is acknowledged that in SIDRIA, area-level SES was differently estimated than in KORA and FINRISK. Lastly, pollutant concentrations were not measured but rather estimated via Land-use Regression models, whose predictors (e.g. traffic indicators) might on turn be associated to SES levels on their own (Details in Table 14.1).

14.2 Conclusion

In summary, the results of this analysis highlight a stronger association of area-based SES to annual average air pollution exposure than individual SES indicators. Overall, the role of area-based indicators was larger for NO_2 than $PM_{2.5}$ and PM_{10}. Additionally, substantial differences were observed regarding the impact of household density and traffic

		LUR model[a]
NO_2	**KORA**	TRAFLOAD_50, INTMAJORINVDIST, ROADLENGTH_50, POP_5000,MAJORROADLENGTH_50, HLDRES_500
	SIDRIA	POP_100, ROADLENGTH_1000, DISTINVNEAR2, INDUSTRY_5000, URBGREEN_1000, TRAFLOAD_50, MAJORROADLENGTH_100
	FINRISK	TRAFLOAD_25, TRAFLOAD_25_1000, ROADLENGTH_25, ROADLENGTH_25_300, URBGREEN_500
$PM_{2.5}$	**KORA**	MAJORROADLENGTH_50, ROADLENGTH_300, URBGREEN_5000, TRAFMAJORLOAD_1000
	SIDRIA	TRAFLOAD_25, ROADLENGTH_100
	FINRISK	NATURAL_500, TRAFMAJORLOAD_50
PM_{10}	**KORA**	MAJORROADLENGTH_50, NATURAL_100, ROADLENGTH_50
	SIDRIA	TRAFLOAD_25, ROADLENGTH_50
	FINRISK	TRAFMAJORLOAD_50, HHOLD_100, NATURAL_500

[a] Variables that are buffered with _X indicating the radius of the buffer in meters: urban green space (URBGREEN_X), natural land (NATURAL_X); population (POP_X), number of households (HHOLD_X), total length of all road (ROADLENGTH_X), all major road segments (MAJORROADLENGTH_X); the product of inverse distance to the nearest major road and the traffic intensity on nearest major roads (vehicles·day^{-1}m^{-1})(INTMAJORINVDIST); the sum of (traffic intensity × length of all road segments) within a buffer (_X) (vehicles·day^{-1}·m) for all roads (TRAFLOAD_X), for major roads (TRAFMAJORLOAD_X)

Table 14.1: Composition of the LUR models for NO_2, $PM_{2.5}$ and PM_{10} for the three studies

covariates, plausibly reflecting the diversity between larger and smaller cities as well as dissimilar types of city across Europe. Path analysis might be a useful tool in order to better enlighten and extensively comprehend the underlying correlation structure. Finally, multi-city analyses require considering the role of different confounding structure by area-based SES over Europe, often also dependent on the size of the cities/areas.

Summary

Ambient air pollution exposure still remains a huge public health problem in the recent decades. Epidemiological evidence of increased risks in major adverse health outcomes such as hospitalization and premature deaths has been observed following increased exposure to air pollution. This work will investigate two topics: the interplay between ambient air pollution exposure and epigenetics and the implication of socio-economic status and other social factors.

Epigenetics rose as promising field in filling the holes left by modern genetics being unable to fully explain the factors responsible for diseases risks. DNA methylation, its most accessible marker, has been found to be linked with both endogenous environmental exposure and adverse health effects. Blood samples were collected in three independent cohorts (KORA F3 and KORA F4 from Germany and Normative Aging Study from the United States) and genome-wide DNA methylation was measured with the Illumina 450k Beadchip. Three different PM2.5 trailing averages prior the visit day (2-, 7- and 28-day) were considered as exposure, representing short- and mid-term variations and an Epigenome-wide Analysis was performed. The model was developed based on previous knowledge and results meta-analyzed across the three studies. Twelve Bonferroni significant CpGs have been identified (one at 2-, one at 7- and ten at 28-day average), and of them, nine displayed increased methylation. Four of them also showed homogeneity across the studies: cg19963313 (on gene NSMAF, chromosome 8), cg23276912 (C1orf212, chr. 1), cg11046593 (MSGN1, chr. 2), and cg26003785 (NXN, chr. 17). Applying False Discovery Rate, 7 additional CpGs were highlighted at 7-day and 1,819 at 28-day average. Finally, sensitivity analysis revealed that one of the associated loci was attributable to long-term effects. In conclusion, PM-related CG targets identified in this study suggest novel plausible pathways between air pollution exposure and adverse health effect. Further studies are needed to better understand possible pathophysiological implications of the detected biological pathways.

Socio-economic status (SES) has been so far considered as effect modification in air pollution studies. The approach followed in this work is the identification of plausible mediators that play a role in the link between SES and pollutants' exposure. Long-term air pollution data (on NO_2, $PM_{2.5}$ and PM_{10}) from three European Cohorts (KORA, Germany, SIDRIA, Italy and FINRISK, Finland) were collected. Households' density and traffic were proposed as possible mediators between air pollution and SES factors (considered as

Percentage of low Income People in a 5km buffer and Personal Income) and Path Model was applied. A clear mediation pattern for percentage of low income people was observed in KORA for NO_2 and PM_{10} (proportion mediated: 56% and 59%, respectively) and in FINRISK for PM10 (proportion mediated: 49%). With personal income, a proportion mediated of 49% was observed in FINRISK for NO_2. The results of this study indicate that area-based SES factors are more related to SES factors than personal indicators and observed differences across the studies revealed a city-specific effect. In order to consider the role of confounding, multi-city analysis need to be performed, in order to account for the differences in size and structure of the urban areas.

Zusammenfassung

Die Belastung durch Luftverschmutzung stellt ein großes Problem für die öffentliche Gesundheit dar. Epidemiologische Daten zeigen ein erhöhtes Risiko für negative gesundheitliche Folgen, wie eine erhöhte Anzahl an Krankenhausaufenthalten und eine geringere Lebenserwartung nach erhöhtem Kontakt mit verschmutzter Umgebungsluft. In dieser Arbeit werden zwei Themenkomplexe bearbeitet: Erstens das Zusammenspiel zwischen Epigenetik und dem Kontakt/der Exposition mit verschmutzter Umgebungsluft und zweitens die Auswirkungen des sozioökonomischen Status und anderer sozialer Faktoren.

Mithilfe epigenetischer Prozesse können Risikofaktoren für umweltbedingte Erkrankungen begründet werden, bei denen die moderne Genetik an ihre Grenzen stößt. So wurde die DNA-Methylierung, als leicht analysierbarer Regulierungsmechanismus, mit endogenen Umweltbelastungen und nachteiligen Effekten auf die Gesundheit assoziiert. Im Rahmen dreier Kohorten (KORA F3 und KORA F4, Deutschland und die Normative Aging Study, USA) wurden Blutproben akquiriert und genomweite DNA-Methylierungsmuster mit dem Illumina 450K Beadchip ermittelt. Drei $PM_{2.5}$ Durchschnittswerte, 2, 7 oder 28 Tage der Probennahme vorausgehend, wurden als Expositionswerte herangezogen um kurz- und mittelfristige Variationen in einer genomweiten Analyse darzustellen. Das Model wurde basierend auf Vorkenntnissen entwickelt und Ergebnisse über die drei Studien metaanalysiert. Zwölf nach Bonferroni-Analyse signifikante CpG-Loci konnten identifiziert werden (ein Locus für den 2-, ein Locus für den 7- und zehn für den 28-Tagesdurchschnitt), von diesen zeigten wiederum neun eine erhöhte Methylierung. Vier dieser Loci wiesen Homogenität zwischen den Studien auf: cg19963313 (auf Gen NSMAF, Chromosom 8), cg23276912 (C1orf212, Chromosom 1), cg11046593 (MSGN1, Chromosom 2), und cg26003785 (NXN, Chromosom 17). Durch Anwendung der False Discovery Rate konnten sieben zusätzliche CpGs für den 7-Tage-Durchschnitt und 1.819 CpGs für den 28-Tage-Durchschnitt ermittelt werden. Schließlich zeigte die Sensitivitätsanalyse, dass einem der assoziierten Loci Langzeiteffekte zugeschrieben werden können. Zusammenfassend wurden in dieser Studie PM-abhängige CpG-Loci identifiziert, die neue, plausible Verbindungen zwischen der Belastung durch Luftverschmutzung und nachteiligen Folgen für die Gesundheit herstellen. Weitere Studien sind notwendig, um mögliche pathophysiologische Prozesse innerhalb der aufgezeigten biologischen Signalwege besser zu verstehen.

Bibliography

[Al-Hakim et al., 2010] Al-Hakim, A., Escribano-Diaz, C., Landry, M., O'Donnell, L., Panier, S., Szilard, R., and Durocher, D. (2010). The ubiquitous role of ubiquitin in the dna damage response. *DNA Repair (Amst)*, 9:1229–40.

[Andrysik et al., 2011] Andrysik, Z., Vondracek, J., Marvanova, S., Ciganek, M., Neca, J., Pencikova, K., Mahadevan, B., Topinka, J., Baird, W. M., Kozubik, A., and Machala, M. (2011). Activation of the aryl hydrocarbon receptor is the major toxic mode of action of an organic extract of a reference urban dust particulate matter mixture: the role of polycyclic aromatic hydrocarbons. *Mutat Res*, 714(1-2):53–62.

[Anon., 1994] Anon. (1994). Active and passive tobacco exposure: a serious pediatric health problem. a statement from the committee on atherosclerosis and hypertension in children, council on cardiovascular disease in the young, american heart association. *Circulation*, 90(5):2581–90.

[Anon., 1997] Anon. (1997). Asthma and respiratory symptoms in 6-7 yr old italian children: gender, latitude, urbanization and socioeconomic factors. sidria (italian studies on respiratory disorders in childhood and the environment). *Eur Respir J*, 10(8):1780–6.

[Aref et al., 2013] Aref, S. F., Ibrahim, L., Morkes, H., Azmy, E., and Ebrahim, M. (2013). Meningioma 1 (mn1) expression: refined risk stratification in acute myeloid leukemia with normal cytogenetics (cn-aml). *Hematology.*, 18:277–83.

[Baccarelli and Bollati, 2009] Baccarelli, A. and Bollati, V. (2009). Epigenetics and environmental chemicals. *Curr Opin Pediatr.*, 21:243–51.

[Baccarelli et al., 2010] Baccarelli, A., Rienstra, M., and Benjamin, E. J. (2010). Cardiovascular epigenetics: basic concepts and results from animal and human studies. *Circ Cardiovasc Genet*, 3(6):567–73.

[Baccarelli et al., 2009] Baccarelli, A., Wright, R. O., Bollati, V., Tarantini, L., Litonjua, A. A., Suh, H. H., A., Z., D., S., S., V. P., and J., S. (2009). Rapid dna methylation changes after exposure to traffic particles. *Am J Respir Crit Care Med.*, 179:572–8.

[Baron and Kenny, 1986] Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, 51(6):1173–82.

[Beelen et al., 2013] Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., and Tsai, M. (2013). Development of no2 and nox land use regression models for estimating air pollution exposure in 36 study areas in europe e the escape project. *Atmospheric Environment*, (72):10e23.

[Bell et al., 1972] Bell, B., Rose, C., and Damon, A. (1972). The normative aging study: an interdisciplinary and longitudinal study of health and aging. *Aging Hum Dev*, pages 83–126.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Bestor, 1998] Bestor, T. H. (1998). The host defence function of genomic methylation patterns. *Novartis Found Symp*, 214:187–95; discussion 195–9, 228–32.

[Bind et al., 2012] Bind, M. A., Baccarelli, A., Zanobetti, A., Tarantini, L., Suh, H., Vokonas, P., and Schwartz, J. (2012). Air pollution and markers of coagulation, inflammation, and endothelial function: associations and epigene-environment interactions in an elderly cohort. *Epidemiology*, 23(2):332–40.

[Bind et al., 2014] Bind, M. A., Lepeule, J., Zanobetti, A., Gasparrini, A., Baccarelli, A., Coull, B. A., Tarantini, L., Vokonas, P. S., Koutrakis, P., and Schwartz, J. (2014). Air pollution and gene-specific methylation in the normative aging study: association, effect modification, and mediation analysis. *Epigenetics.*, 9:448–58.

[Blanco-Becerra et al., 2014] Blanco-Becerra, L. C., Miranda-Soberanis, V., Barraza-Villarreal, A., Junger, W., Hurtado-Diaz, M., and Romieu, I. (2014). Effect of socioeconomic status on the association between air pollution and mortality in bogota, colombia. *Salud Publica Mex*, 56(4):371–8.

[Bossé et al., 1984] Bossé, R., Ekerdt, D. J., and Silbert, J. E. (1984). The veterans administration normative aging study. *Handbook of longitudinal research*, 2:273–289.

[Breton et al., 2009] Breton, C. V., Byun, H. M., Wenten, M., Pan, F., Yang, A., and Gilliland, F. D. (2009). Prenatal tobacco smoke exposure affects global and gene-specific dna methylation. *Am J Respir Crit Care Med*, 180(5):462–7.

[Brook and Rajagoapalan, 2010] Brook, R. D. and Rajagoapalan, S. (2010). Particulate matter air pollution and atherosclerosis. *Curr Atheroscler Rep.*, 12:291–300.

[Brook et al., 2010] Brook, R. D., Rajagopalan, S., Pope, C. A., r., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., J., Whitsel, L., and Kaufman, J. D. (2010).

Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association. *Circulation*, 121(21):2331–78.

[Byon et al., 2008] Byon, C. H., Javed, A., Dai, Q., Kappes, J. C., Clemens, T. L., Darley-Usmar, V. M., McDonald, J. M., and Chen, Y. (2008). Oxidative stress induces vascular calcification through modulation of the osteogenic transcription factor runx2 by akt signaling. *J Biol Chem*, 283(22):15319–27.

[Cassee et al., 2013] Cassee, F. R., Héroux, M. E., E., G.-N. M., and J., K. F. (2013). Particulate matter beyond mass: recent health evidence on the role of fractions, chemical constituents and sources of emission. *Inhal Toxicol.*, 25:802–12.

[Cesaroni et al., 2014] Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z. J., Badaloni, C., Beelen, R., Caracciolo, B., de Faire, U., Erbel, R., Eriksen, K. T., Fratiglioni, L., Galassi, C., Hampel, R., Heier, M., Hennig, F., Hilding, A., Hoffmann, B., Houthuijs, D., Jockel, K. H., Korek, M., Lanki, T., Leander, K., Magnusson, P. K., Migliore, E., Ostenson, C. G., Overvad, K., Pedersen, N. L., J, J. P., Penell, J., Pershagen, G., Pyko, A., Raaschou-Nielsen, O., Ranzi, A., Ricceri, F., Sacerdote, C., Salomaa, V., Swart, W., Turunen, A. W., Vineis, P., Weinmayr, G., Wolf, K., de Hoogh, K., Hoek, G., Brunekreef, B., and Peters, A. (2014). Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 european cohorts from the escape project. *Bmj*, 348:f7412.

[Chanda et al., 2006] Chanda, S., Dasgupta, U. B., Guhamazumder, D., Gupta, M., Chaudhuri, U., Lahiri, S., Das, S., Ghosh, N., and Chatterjee, D. (2006). Dna hypermethylation of promoter of gene p53 and p16 in arsenic-exposed people with and without malignancy. *Toxicol Sci*, 89(2):431–7.

[Christensen et al., 2008] Christensen, B. C., Godleski, J. J., Marsit, C. J., Houseman, E. A., Lopez-Fagundo, C. Y., Longacker, J. L., Bueno, R., Sugarbaker, D. J., Nelson, H. H., and Kelsey, K. T. (2008). Asbestos exposure predicts cell cycle control gene promoter methylation in pleural mesothelioma. *Carcinogenesis*, 29(8):1555–9.

[Christensen and Marsit, 2012] Christensen, B. C. and Marsit, C. J. (2012). Epigenomics in environmental health. *Front Genet.*, page 2:84.

[Cyrys et al., 2012] Cyrys, J., Eeftens, M., Heinrich, J., Ampe, C., Armengaud, A., Beelen, R., Bellander, T., Beregszaszi, T., Birk, M., and Cesaroni, G. (2012). Variation of no 2 and no x concentrations between and within 36 european study areas: results from the escape study. *Atmospheric Environment*, 62:374–390.

[Dai et al., 2014] Dai, L., Zanobetti, A., Koutrakis, P., and Schwartz, J. D. (2014). Associations of fine particulate matter species with mortality in the united states: a multicity time-series analysis. *Environ Health Perspect*, 122(8):837–42.

[Dedeurwaerder et al., 2011] Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the infinium methylation 450k technology. *Epigenomics.*, 3:771–84.

[Di Meglio et al., 2014] Di Meglio, P., Duarte, J. H., Ahlfors, H., Owens, N. D., Li, Y., Villanova, F., Tosi, I., Hirota, K., Nestle, F. O., Mrowietz, U., Gilchrist, M. J., and Stockinger, B. (2014). Activation of the aryl hydrocarbon receptor dampens the severity of inflammatory skin conditions. *Immunity*, 40(6):989–1001.

[Dubowsky et al., 2006] Dubowsky, S. D., Suh, H., Schwartz, J., Coull, B. A., and Gold, D. R. (2006). Diabetes, obesity, and hypertension may enhance associations between air pollution and markers of systemic inflammation. *Environ Health Perspect*, 114(7):992–8.

[Dudbridge and Gusnanto, 2008] Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32(3):227–34.

[Eeftens et al., 2012] Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., de Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Grazuleviciene, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Molter, A., Nador, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M. Y., Yli-Tuomi, T., Varro, M. J., Vienneau, D., Klot, S., Wolf, K., Brunekreef, B., and Hoek, G. (2012). Development of land use regression models for pm(2.5), pm(2.5) absorbance, pm(10) and pm(coarse) in 20 european study areas; results of the escape project. *Environ Sci Technol*, 46(20):11195–205.

[Fior et al., 2012] Fior, R., Maxwell, A. A., Ma, T. P., Vezzaro, A., Moens, C. B., Amacher, S. L., Lewis, J., and Saúde, L. (2012). The differentiation and movement of presomitic mesoderm progenitor cells are controlled by mesogenin 1. *Development*, 139:4656–65.

[Forastiere et al., 2007] Forastiere, F., Stafoggia, M., Tasco, C., Picciotto, S., Agabiti, N., Cesaroni, G., and Perucci, C. A. (2007). Socioeconomic status, particulate air pollution, and daily mortality: differential exposure or differential susceptibility. *Am J Ind Med*, 50(3):208–16.

[Fustinoni et al., 2012] Fustinoni, S., Rossella, F., Polledri, E., Bollati, V., Campo, L., Byun, H. M., Agnello, L., Consonni, D., Pesatori, A. C., Baccarelli, A., and Bertazzi,

P. A. (2012). Global dna methylation and low-level exposure to benzene. *Med Lav*, 103(2):84–95.

[Gardiner-Garden and Frommer, 1987] Gardiner-Garden, M. and Frommer, M. (1987). Cpg islands in vertebrate genomes. *J Mol Biol*, 196(2):261–82.

[Gray et al., 2013] Gray, S. C., Edwards, S. E., and Miranda, M. L. (2013). Race, socioeconomic status, and air pollution exposure in north carolina. *Environ Res*, 126:152–8.

[Gray et al., 2014] Gray, S. C., Edwards, S. E., Schultz, B. D., and Miranda, M. L. (2014). Assessing the impact of race, social factors and air pollution on birth outcomes: a population-based study. *Environ Health*, 13(1):4.

[Guo et al., 2014] Guo, L., Byun, H. M., Zhong, J., Motta, V., Barupal, J., Zheng, Y., Dou, C., Zhang, F., McCracken, J. P., Diaz, A., Marco, S. G., Colicino, S., Schwartz, J., Wang, S., Hou, L., and Baccarelli, A. A. (2014). Effects of short-term exposure to inhalable particulate matter on dna methylation of tandem repeats. *Environ Mol Mutagen.*, 55:322–35.

[Hajat et al., 2013] Hajat, A., Diez-Roux, A. V., Adar, S. D., Auchincloss, A. H., Lovasi, G. S., O'Neill, M. S., Sheppard, L., and Kaufman, J. D. (2013). Air pollution and individual and neighborhood socioeconomic status: evidence from the multi-ethnic study of atherosclerosis (mesa). *Environ Health Perspect*, 121(11-12):1325–33.

[Hallows et al., 1996] Hallows, K. R., Law, F. Y., Packman, C. H., and Knauf, P. A. (1996). Changes in cytoskeletal actin content, f-actin distribution, and surface morphology during hl-60 cell volume regulation. *J Cell Physiol*, 167(1):60–71.

[Hedges and Deininger, 2007] Hedges, D. J. and Deininger, P. L. (2007). Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res*, 616(1-2):46–59.

[Holle et al., 2005] Holle, R., Happich, M., Lowel, H., and Wichmann, H. E. (2005). Kora–a research platform for population based health research. *Gesundheitswesen*, 67 Suppl 1:S19–25.

[Holm, 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

[Houseman et al., 2012] Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.*, 13:86.

[Hu et al., 2004] Hu, L., Lee, M., Campbell, W., Perez-Soler, R., and Karpatkin, S. (2004). Role of endogenous thrombin in tumor implantation, seeding, and spontaneous metastasis. *Blood*, 104:2746–51.

[Inc., 2009] Inc., M. (2009). Mosby's dictionary of medicine, nursing & health professions.

[Jaffe et al., 2012] Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–209.

[Jana and Pahan, 2007] Jana, A. and Pahan, K. (2007). Oxidative stress kills human primary oligodendrocytes via neutral sphingomyelinase: implications for multiple sclerosis. *J Neuroimmune Pharmacol*, 2(2):184–93.

[Jones and Takai, 2001] Jones, P. A. and Takai, D. (2001). The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–70.

[Koensgen et al., 2007] Koensgen, D., Mustea, A., Klaman, I., Sun, P., Zafrakas, M., Lichtenegger, W., Denkert, C., Dahl, E., and Sehouli, J. (2007). Expression analysis and rna localization of pai-rbp1 (serbp1) in epithelial ovarian cancer: association with tumor progression. *Gynecol Oncol.*, 107:266–73.

[Laird, 2005] Laird, P. W. (2005). Cancer epigenetics. *Hum Mol Genet*, page 14.

[Lanki et al., 2015] Lanki, T., Hampel, R., Tiittanen, P., Andrich, S., Beelen, R., Brunekreef, B., Dratva, J., De Faire, U., Fuks, K. B., Hoffman, B., Imboden, M., Jousilahti, P., Koenig, W., Mahabadi, A. A., Kunzli, N., Pedersen, N. L., Penell, J., Pershagen, G., Probst-Hensch, N. M., Schaffner, E., Schindler, C., Sugiri, D., Swart, W. J., Tsai, M. Y., Turunen, A. W., Weinmayr, G., Wolf, K., Yli-Tuomi, T., and Peters, A. (2015). Air pollution from road traffic and systemic inflammation in adults: A cross-sectional analysis in the european escape project. *Environ Health Perspect.*

[Lee et al., 2007] Lee, J., Beliakoff, J., and Sun, Z. (2007). The novel pias-like protein hzimp10 is a transcriptional co-activator of the p53 tumor suppressor. *Nucleic Acids Res.*, 35:4523–34.

[Liang and Zeger, 1986] Liang, K. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, pages 13–22.

[Liao et al., 2013] Liao, L., Zhou, Q., Song, Y., Wu, W., Yu, H., Wang, S., Chen, Y., Ye, M., and Lu, L. (2013). Ceramide mediates ox-ldl-induced human vascular smooth muscle cell calcification via p38 mitogen-activated protein kinase signaling. *PLoS One*, 8(12):e82379.

[Liu et al., 2010] Liu, T., Jankovic, D., Brault, L., Ehret, S., Baty, F., Stavropoulou, V., Rossi, V., Biondi, A., and Schwaller, J. (2010). Functional characterization of high levels of meningioma 1 as collaborating oncogene in acute leukemia. *Leukemia.*, 24:601–12.

[MacKinnon et al., 2007] MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annu Rev Psychol.*, 58:593.

[Madjid et al., 2004] Madjid, M., Awan, I., Willerson, J. T., and Casscells, S. W. (2004). Leukocyte count and coronary heart disease: implications for risk assessment. *J Am Coll Cardiol*, 44(10):1945–56.

[Marabita et al., 2013] Marabita, F., Almgren, M., Lindholm, M. E., Ruhrmann, S., Fagerstrom-Billai, F., Jagodic, M., Sundberg, C. J., Ekstrom, T. J., Teschendorff, A. E., Tegner, J., and Gomez-Cabrero, D. (2013). An evaluation of analysis pipelines for dna methylation profiling using the illumina humanmethylation450 beadchip platform. *Epigenetics*, 8(3):333–46.

[McConnell et al., 2015] McConnell, R., Shen, E., Gilliland, F. D., Jerrett, M., Wolch, J., Chang, C. C., Lurmann, F., and Berhane, K. (2015). A longitudinal cohort study of body mass index and childhood exposure to secondhand tobacco smoke and air pollution: the southern california children's health study. *Environ Health Perspect*, 123(4):360–6.

[Miettinen and Cook, 1981] Miettinen, O. S. and Cook, E. F. (1981). Confounding: essence and detection. *Am J Epidemiol*, 114(4):593–603.

[Montfort et al., 2010] Montfort, A., Martin, P. G., Levade, T., Benoist, H., and Ségui, B. (2010). Fan (factor associated with neutral sphingomyelinase activation), a moonlighting protein in tnf-r1 signaling. *J Leukoc Biol.*, 88:897–903.

[Mor et al., 2011] Mor, I., Cheung, E. C., and Vousden, K. H. (2011). Control of glycolysis through regulation of pfk1: old friends and recent additions. In *Cold Spring Harbor symposia on quantitative biology*, volume 76, pages 211–216. Cold Spring Harbor Laboratory Press.

[Nurminen, 1997] Nurminen, M. (1997). On the epidemiologic notion of confounding and confounder identification. *Scand J Work Environ Health*, 23(1):64–8.

[Orphanides and Reinberg, 2002] Orphanides, G. and Reinberg, D. (2002). A unified theory of gene expression. *Cell*, 108(4):439–51.

[Ou et al., 2008] Ou, C. Q., Hedley, A. J., Chung, R. Y., Thach, T. Q., Chau, Y. K., Chan, K. P., Yang, L., Ho, S. Y., Wong, C. M., and Lam, T. H. (2008). Socioeconomic disparities in air pollution-associated mortality. *Environ Res*, 107(2):237–44.

[Ovrevik et al., 2014] Ovrevik, J., Lag, M., Lecureur, V., Gilot, D., Lagadic-Gossmann, D., Refsnes, M., Schwarze, P. E., Skuland, T., Becher, R., and Holme, J. A. (2014). Ahr and arnt differentially regulate nf-kappab signaling and chemokine responses in human bronchial epithelial cells. *Cell Commun Signal*, 12:48.

[Peters, 2012] Peters, A. (2012). Epidemiology: air pollution and mortality from diabetes mellitus. *Nat Rev Endocrinol.*, 8:706–7.

[Ponticiello et al., 2015] Ponticiello, B. G., Capozzella, A., Di Giorgio, V., Casale, T., Giubilati, R., Tomei, G., Tomei, F., Rosati, M. V., and Sancini, A. (2015). Overweight and urban pollution: preliminary results. *Sci Total Environ*, 518-519:61–4.

[Raaschou-Nielsen et al., 2013] Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., Hoffmann, B., Fischer, P., Nieuwenhuijsen, M. J., Brunekreef, B., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Sommar, J., Forsberg, B., Modig, L., Oudin, A., Oftedal, B., Schwarze, P. E., Nafstad, P., De Faire, U., Pedersen, N. L., Ostenson, C. G., Fratiglioni, L., Penell, J., Korek, M., Pershagen, G., Eriksen, K. T., Sorensen, M., Tjonneland, A., Ellermann, T., Eeftens, M., Peeters, P. H., Meliefste, K., Wang, M., Bueno-de Mesquita, B., Key, T. J., de Hoogh, K., Concin, H., Nagel, G., Vilier, A., Grioni, S., Krogh, V., Tsai, M. Y., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Trichopoulou, A., Bamia, C., Vineis, P., and Hoek, G. (2013). Air pollution and lung cancer incidence in 17 european cohorts: prospective analyses from the european study of cohorts for air pollution effects (escape). *Lancet Oncol*, 14(9):813–22.

[Rakowski et al., 2013] Rakowski, L. A., Garagiola, D. D., Li, C. M., Decker, M., Caruso, S., Jones, M., Kuick, R., Cierpicki, T., Maillard, I., and Chiang, M. Y. (2013). Convergence of the zmiz1 and notch1 pathways at c-myc in acute t lymphoblastic leukemias. *Cancer Res.*, 73:930–41.

[Reinius et al., 2012] Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlen, S. E., Greco, D., Soderhall, C., Scheynius, A., and Kere, J. (2012). Differential dna methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One*, 7(7):e41361.

[Robciuc et al., 2012] Robciuc, A., Hyotylainen, T., Jauhiainen, M., and Holopainen, J. M. (2012). Hyperosmolarity-induced lipid droplet formation depends on ceramide production by neutral sphingomyelinase 2. *J Lipid Res*, 53(11):2286–95.

[Rogers et al., 2013] Rogers, L. M., Riordan, J. D., Swick, B. L., Meyerholz, D. K., and Dupuy, A. J. (2013). Ectopic expression of zmiz1 induces cutaneous squamous cell malignancies in a mouse model of cancer. *J Invest Dermatol*, 133(7):1863–9.

[Ruckerl et al., 2011] Ruckerl, R., Schneider, A., Breitner, S., Cyrys, J., and Peters, A. (2011). Health effects of particulate air pollution: A review of epidemiological evidence. *Inhal Toxicol*, 23(10):555–92.

[Rückerl et al., 2007] Rückerl, R., Greven, S., Ljungman, P., Aalto, P., Antoniades, C., Bellander, T., Berglind, N., Chrysohoou, C., Forastiere, F., Jacquemin, B., von Klot, S., Koenig, W., Küchenhoff, H., Lanki, T., Pekkanen, J., Perucci, C. A., Schneider, A., Sunyer, J., Peters, A., and Group., A. S. (2007). Air pollution and inflammation (interleukin-6, c-reactive protein, fibrinogen) in myocardial infarction survivors. *Environ Health Perspect.*, 115:1072–80.

[Sandoval et al., 2011] Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M. (2011). Validation of a dna methylation microarray for 450,000 cpg sites in the human genome. *Epigenetics.*, 6:692–702.

[Schwartz, 2000] Schwartz, J. (2000). Assessing confounding, effect modification, and thresholds in the association between ambient particles and daily deaths. *Environ Health Perspect.*, 108:563–8.

[Scrivo et al., 2011] Scrivo, R., Vasile, M., Bartosiewicz, I., and Valesini, G. (2011). Inflammation as "common soil" of the multifactorial diseases. Autoimmun Rev(10):369–374.

[Sen et al., 2015] Sen, A., Cingolani, P., Senut, M., Land, S., Mercado-Garcia, A., Tellez-Rojo, M. M., Baccarelli, A. A., Wright, R. O., and Ruden, D. M. (2015). Lead exposure induces changes in 5-hydroxymethylcytosine clusters in cpg islands in human embryonic stem cells and umbilical cord blood. *Epigenetics*, 10(7):607–621.

[Serce et al., 2012] Serce, N. B., Boesl, A., Klaman, I., von Serényi, S., Noetzel, E., Press, M. F., Dimmler, A., Hartmann, A., Sehouli, J., Knuechel, R., Beckmann, M. W., Fasching, P. A., and Dahl, E. (2012). Overexpression of serbp1 (plasminogen activator inhibitor 1 rna binding protein) in human breast cancer is correlated with favourable prognosis. *BMC Cancer*, 12:597.

[Shah et al., 2014] Shah, S., McRae, A. F., Marioni, R. E., Harris, S. E., Gibson, J., Henders, A. K., Redmond, P., Cox, S. R., Pattie, A., and Corley, J. (2014). Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome research*, 24(11):1725–1733.

[Soberanes et al., 2012] Soberanes, S., Gonzalez, A., Urich, D., Chiarella, S. E., Radigan, K. A., Osornio-Vargas, A., Joseph, J., Kalyanaraman, B., Ridge, K. M., Chandel, N. S., Mutlu, G. M., De Vizcaya-Ruiz, A., and Budinger, G. R. (2012). Particulate matter air pollution induces hypermethylation of the p16 promoter via a mitochondrial ros-jnk-dnmt1 pathway. *Sci Rep*, 2:275.

[Sofer et al., 2013] Sofer, T., Schifano, E. D., Hoppin, J. A., Hou, L., and Baccarelli, A. A. (2013). A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics*, page btt498.

[Soler et al., 2008] Soler, G., Radford-Weiss, I., Ben-Abdelali, R., Mahlaoui, N., Ponceau, J. F., Macintyre, E. A., Vekemans, M., Bernard, O. A., and Romana, S. P. (2008). Fusion of zmiz1 to abl1 in a b-cell acute lymphoblastic leukaemia with a t(9;10)(q34;q22.3) translocation. *Leukemia.*, 22:1278–80.

[Stafoggia et al., 2014] Stafoggia, M., Cesaroni, G., Peters, A., Andersen, Z. J., Badaloni, C., Beelen, R., Caracciolo, B., Cyrys, J., de Faire, U., de Hoogh, K., Eriksen, K. T., Fratiglioni, L., Galassi, C., Gigante, B., Havulinna, A. S., Hennig, F., Hilding, A., Hoek, G., Hoffmann, B., Houthuijs, D., Korek, M., Lanki, T., Leander, K., Magnusson, P. K., Meisinger, C., Migliore, E., Overvad, K., Ostenson, C. G., Pedersen, N. L., Pekkanen, J., Penell, J., Pershagen, G., Pundt, N., Pyko, A., Raaschou-Nielsen, O., Ranzi, A., Ricceri, F., Sacerdote, C., Swart, W. J., Turunen, A. W., Vineis, P., Weimar, C., Weinmayr, G., Wolf, K., Brunekreef, B., and Forastiere, F. (2014). Long-term exposure to ambient air pollution and incidence of cerebrovascular events: results from 11 european cohorts within the escape project. *Environ Health Perspect*, 122(9):919–25.

[Tarantini et al., 2009] Tarantini, L., Bonzini, M., Apostoli, P., Pegoraro, V., Bollati, V., Marinelli, B., Cantone, L., Rizzo, G., Hou, L., Schwartz, J., Bertazzi, P. A., and Baccarelli, A. (2009). Effects of particulate matter on genomic dna methylation content and inos promoter methylation. *Environ Health Perspect*, 117(2):217–22.

[Teschendorff et al., 2013] Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics.*, 29:189–96.

[Touleimat and Tost, 2012] Touleimat, N. and Tost, J. (2012). Complete pipeline for infinium(®) human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation. *Epigenomics.*, 4:325–41.

[Ukena et al., 2010] Ukena, C., Mahfoud, F., Kindermann, M., Kindermann, I., Bals, R., Voors, A. A., van Veldhuisen, D. J., and Bohm, M. (2010). The cardiopulmonary continuum systemic inflammation as ‘common soil’ of heart and lung disease. *Int J Cardio*, 145:172–176.

[Ulrich and Walden, 2010] Ulrich, H. D. and Walden, H. (2010). Ubiquitin signalling in dna replication and repair. *Nat Rev Mol Cell Biol.*, 11:479–89.

[Valeri and Vanderweele, 2013] Valeri, L. and Vanderweele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with sas and spss macros. *Psychol Methods*, 18(2):137–50.

[van Voorhis et al., 2013] van Voorhis, M., Knopp, S., Julliard, W., Fechner, J. H., Zhang, X., Schauer, J. J., and Mezrich, J. D. (2013). Exposure to atmospheric particulate matter enhances th17 polarization through the aryl hydrocarbon receptor. *PLoS One*, 8(12):e82545.

[Vartiainen et al., 2010] Vartiainen, E., Laatikainen, T., Peltonen, M., Juolevi, A., Mannisto, S., Sundvall, J., Jousilahti, P., Salomaa, V., Valsta, L., and Puska, P. (2010). Thirty-five-year trends in cardiovascular risk factors in finland. *Int J Epidemiol*, 39(2):504–18.

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

[Warde-Farley et al., 2010] Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., and Morris, Q. (2010). The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, 38(Web Server issue):W214–20.

[Whittemore, 1978] Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 328–340.

[Wichmann et al., 2005] Wichmann, H. E., Gieger, C., and Illig, T. (2005). Kora-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen.*, 67:S26–30.

[Wieczorek et al., 2009] Wieczorek, G., Asemissen, A., Model, F., Turbachova, I., Floess, S., Liebenberg, V., Baron, U., Stauch, D., Kotsch, K., Pratschke, J., Hamann, A., Loddenkemper, C., Stein, H., Volk, H. D., Hoffmuller, U., Grutzkau, A., Mustea, A., Huehn, J., Scheibenbogen, C., and Olek, S. (2009). Quantitative dna methylation analysis of foxp3 as a new method for counting regulatory t cells in peripheral blood and solid tissue. *Cancer Res*, 69(2):599–608.

[Wittler et al., 2007] Wittler, L., Shin, E. H., Grote, P., Kispert, A., Beckers, A., Gossler, A., Werber, M., and Herrmann, B. G. (2007). Expression of msgn1 in the presomitic mesoderm is controlled by synergism of wnt signalling and tbx6. *EMBO Rep.*, 8:784–9.

[Wright, 1965] Wright, S. (1965). Path coefficients and path regressions: Alternative or complementary concepts?

[Wu et al., 2010] Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*, 11(5):R53.

[Xiang et al., 2013] Xiang, L., Li, M., Liu, Y., Cen, J., Chen, Z., Zhen, X., Xie, X., Cao, X., and Gu, W. (2013). The clinical characteristics and prognostic significance of mn1 gene and mn1-associated microrna expression in adult patients with de novo acute myeloid leukemia. *Ann Hematol.*, 92:1063–9.

[Yang and Omaye, 2009] Yang, W. and Omaye, S. T. (2009). Mutation research/genetic toxicology and environmental mutagenesis. *Oxidative Stress and Mechanisms of Environmental Toxicity*, 674(1-2):45–54.

[Yap et al., 2013] Yap, P. S., Gilbreath, S., Garcia, C., Jareen, N., and Goodrich, B. (2013). The influence of socioeconomic markers on the association between fine particulate matter and hospital admissions for respiratory conditions among children. *Am J Public Health*, 103(4):695–702.

[Yauk et al., 2008] Yauk, C., Polyzos, A., Rowan-Carroll, A., Somers, C. M., Godschalk, R. W., Van Schooten, F. J., Berndt, M. L., Pogribny, I. P., Koturbash, I., Williams, A., Douglas, G. R., and Kovalchuk, O. (2008). Germ-line mutations, dna damage, and global hypermethylation in mice exposed to particulate air pollution in an urban/industrial location. *Proc Natl Acad Sci U S A*, 105(2):605–10.

[Yi et al., 2012] Yi, W., Clark, P. M., Mason, D. E., Keenan, M. C., Hill, C., Goddard, W. A., Peters, E. C., Driggers, E. M., and Hsieh-Wilson, L. C. (2012). Phosphofructokinase 1 glycosylation regulates cell growth and metabolism. *Science*, 337(6097):975–980.

[Zeger et al., 2000] Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspect*, 108(5):419–26.

[Zeilinger et al., 2013] Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A., Strauch, K., Waldenberger, M., and Illig, T. (2013). Tobacco smoking leads to extensive genome-wide changes in dna methylation. *PLoS One*, 8:e63812.

[Zhao et al., 2013] Zhao, J., Gao, Z., Tian, Z., Xie, Y., Xin, F., Jiang, R., Kan, H., and Song, W. (2013). The biological effects of individual-level pm(2.5) exposure on systemic immunity and inflammatory response in traffic policemen. *Occup Environ Med*, 70(6):426–31.

[Zhu et al., 2010] Zhu, Z. Z., Hou, L., Bollati, V., Tarantini, L., Marinelli, B., Cantone, L., Yang, A. S., Vokonas, P., Lissowska, J., Fustinoni, S., Pesatori, A. C., Bonzini, M., Apostoli, P., Costa, G., Bertazzi, P. A., Chow, W. H., Schwartz, J., and Baccarelli, A. (2010). Predictors of global methylation levels in blood dna of healthy subjects: a combined analysis. *Int J Epidemiol.*, 41:126–39.

Acknowledgments

The first person I must acknowledge is my supervisor **Annette**. Thanks for having me ("the other Italian") as a Graduate Student and for the "ocean of patience" you needed to correct my very bad writing style. I never felt underestimated and never quit a meeting disappointed but rather animated. For that, I would like to thank you for your dynamism, motivation and enthusiasm. However, one more important thing I learned (or at least I will ideally try to bear in mind): to look at facts from different perspectives, not just directly, but also from the sides and all these perspectives may hide unexpected, surprising and challenging positive news.

Then, it's the moment of my group leader **Alex**. I should say many things but I want to focus on one: your fully openness and helpfulness. Just one example are worth more than many words. Last summer, after a series of questions I wrote to you "may I ask one last question?" to which you replied "you can ask me whatever you want". I stared at the monitor for several seconds. Such gratuity cannot be taken for granted.

Key part of my working development was played by Prof. **Andrea Baccarelli**. Aware or not, you introduced me to science, a fascinating world, that I leave with a struggle. Through my first experience at HSPH, I could see the unique sparkles of the scientific research, as also paradigm of life, and it left an everlasting impression on me.

Keeping on the other side of the ocean, many thanks to Prof. **Joel Schwartz**. The KORA-NAS collaboration was very fruitful (and I hope it will continue) and I am glad and honored of being the operative link.

I would like to further thank **Hanna Kirchmair**. Her unconditional and friendly welcome throughout the time has been priceless. Your gratuity of caring about me even when I thought I didn't need it left me speechless.

Simone Wahl. First of all, it was precious to share a project with you. Your determination, dedication and detail-orientation are unique. Then, my thoughts go to the Stats Club. We started it with many doubts and uncertainties but now I clearly see it as an elite virtuous example of sharing knowledge, experience and passion.

Special thanks to **Stefanie Lanzinger**, who started her PhD with me. I could see your experience grow with time. I also always appreciated your commitment and preciseness in your tasks that inspired me in my laziness and negligence.

Several thanks to the all EnRi group: **Ute**, **Katja**, **Kathi**, **Regina**, **Regina**, **Chang**, **Susi**, **Josef** and **Jianwei**. Thanks for your patience and for keeping reminding me of things I should have already known and for standing with me even if sometimes my brain was clearly at the "Hawaii". I also appreciated very much your continuous invitations at your place to share extra-work life.

My long-time office-mates, **Uta Geruschkat** and **Bernadette Lex**. Not only did you stand with me but you kept making our working-time very pleasant and enjoyable and supporting me in my difficulties, often due to my "rudimentär Deutsch". That was priceless. Thanks also to **Brigitte** and **Hanni**.

Several thanks to other collaborators, **Melanie Waldenberger** and the methylation group (especially **Sonja**, **Lilli**, **Jenni** and **Eva**). It was very fruitful sharing with you the biological side of my work. Thanks for answering all the time to my very basic questions. Special thanks also to **Rui Wang-Sattler** and her group (**Tao**, **Markus**, **Jonni**, **Stefan** and **Martina**) for the useful discussions and our shared work and consideration.

I would further like to thank Prof. **Konstantin Strauch**, also member of my Thesis Committee. Thanks for your ever-useful comments and inputs and also for letting me work with the LINUX server. Along with that I also thank all the group of people involved in its maintenance. It was paramount for me and my research.

Many thanks for the productive time shared with Prof. **Holger Schulz** and Dr. **Stefan Karrasch**. You were very dedicated and determined and it was a pleasure for me to work together.

Lena Riess and **Tonia de la Heras Gala**. First thanks for your Spanish touch, then thanks for the very good time we shared and working on the WBC project and the Stats Club. It's not granted to meet people as open-minded as you are.

Cavin Ward-Caviness. Thanks, instead, for your American touch. Also thanks for your several advices; you have been a "stable island within the sea waves".

Clemens Baumbach. I cannot say how precious for me was your help with LINUX but even more is the possibility of having met such a good person as you are.

Thanks a lot to all the people who were always very nice to me and helped at the various Epi Institutes: **Gabi**, **Angie**, **Heike**, **Anja**, prof. **Ladwig**, Prof. **Thorand**, **Jürgen**, **Herbert**, **Sanne**, **Susanne**, **Jens**, **Astrid**, **Birgit**, **Brigitte**, **Elaine** and **Iana**. I want to also thank the students who brought a touch of youth: **Hanna** and **Manuel**, my office neighbors and **Margarethe** and **Tatiana** from EnRi.

Thanks a lot to my second family: the containers. **Anna**, **Julia**, **Patricia**, **Inke**, **Inga**, **Conni**, **Christina**, **Nina** and Prof. **Linseisen**, thanks a lot for a friendly welcome in

"Gebäude" 60. Breaking the routine just before the end might be a pain but I never experienced such a feeling. Instead you accompanied me very discretely (that doesn't mean leaving me alone!) to a new step of my life. I appreciated that.

Thanks a lot to the Graduate School Office: **Monika**, **Gaby**, **Julia** and **Andrea**. I always appreciated your help and patience regarding my bureaucratic requests. But most of all, thanks for all your work with HELENA, for your effort to improve it and enrich the PhD Students' life.

A special thanks goes to the **Mensa**. Have a break during the day and look and meet other people is a healthy input for our lives.

Thanks a lot to my German teacher **Rainald** who I first met at HMGU. Thanks for all your support regarding my ironical effort to speak good German and helping me to feel more at home. This was a clear example to show that differences make people stronger.

Last special acknowledgment: **DINI**. Well, I don't have much to "say" apart than a huge massive "thank you". Thanks for your dedication, your free time, your commitment, your interest, your care, your desire and whatever else made the group united even in moments of contrast. Thanks for having gone beyond the social characterization "a group of Students" towards a broader "group of Persons".

Eidesstattliche Versicherung

Panni, Tommaso

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Thema

Environmental Epigenomics

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 26.01.16

Ort, Datum

Unterschrift Doktorandin/Doktorand