Omics biomarkers to study the internal exposome: the case of air pollution

Nahid Mostafavi

Omics biomarkers to study the internal exposome: the case of air pollution
Thesis Utrecht University
ISBN: 978-94-028-1504-7

Cover design: Atefeh Nasserinejad
Layout by: Nahid Mostafavi
Printed by: Ipskamp drukkers, Enschede

Omics biomarkers to study the internal exposome: the case of air pollution

Omics biomerkers voor het bestuderen van het interne exposoom:
een uitwerkingen voor luchtverontreiniging
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de rector magnificus

prof. dr. H.R.B.M. Kummeling

ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
donderdag 4 juli 2019 des ochtends te 10.30 uur

door

Nahid Sadat Mostafavi Montazeri

geboren op 14 augustus 1985

te Teheran, Iran

Promotoren:	Prof. dr. R.C.H. Vermeulen Prof. dr. B. Brunekreef
Copromotor:	Dr. J.J. Vlaanderen

Several of the studies described in this thesis were financially supported by grants from the European Unions Seventh Framework Programmes: EnviroGenoMarkers (Grant agreement number: 226756), ESCAPE (Grant agreement number: 211250), and EXPOsOMICS (Grant agreement number: 308610).

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

Introduction

Our health is impacted by the environment in which we grow up, live, work, and sleep. In general, the majority of important chronic diseases are likely to be the result of the combination of environmental exposures and human genetics (Cui et al., 2016).
The "exposome" concept, first defined by Wild (2005) acknowledges the large role that environment can play in the causation of disease. The exposome is a paradigm involving the study of the effects of all environmental exposures and associated biological responses from conception until death. These exposures range from population-level exposures such as climate, air pollution, and urban noise to individual-level exposures such as diet, smoking, physical activity, social interactions, and stress. Once such exposures enter the human body they become part of the internal exposome, which can be measured by for example metabolic changes, protein perturbations, or epigenetic alterations (DeBord et al., 2016; Wild, 2005; Wild et al., 2013).
Investigating perturbations in the internal exposome can provide information on direct measures of exposure (e.g. specific metabolites) or on physiological perturbations (e.g. changes in DNA-methylation) that are indicative of a certain exposure. Studying the internal exposome can provide biological underpinnings for empirically observed exposure-disease associations. For example, it has been shown that smoking-induced DNA methylation markers also predict the risk of smoking related diseases such as lung cancer, shedding light on the biological pathways involved in this known association (Fasanelli et al., 2015).
In 2011, Rappaport and Smith suggested two ways to explore the exposome (Rappaport, 2011; Wild, 2005). The first is a "top-down" approach mainly interested in identifying new causes of disease by an agnostic investigation based on internal exposome. This approach utilizes methods such as metabolomics (assessment of all small molecules present in a biological matrix) or adductomics (global screening of chemical specific DNA-adducts) to generate new hypothesis on disease etiology. The second is a "bottom-up" approach that starts with a set of exposures or environmental compartments to determine the pathways or networks which lead to disease.
The aim of the work described in this thesis was to identify perturbations in the biological system (in particular proteomics, transcriptomics, and DNA-methylomics) associated with air pollution exposure and disease, which is an example of a "bottom-up" approach.

OMICS to study internal exposome

In the biological sciences the suffix -omics is used to name the large sets of cellular molecules ranging from genomics (focused on the genome) to transcriptomics (focused on large set of mRNA), proteomics (focused on large sets of proteins) and metabolomics (focused on large sets of small molecules).

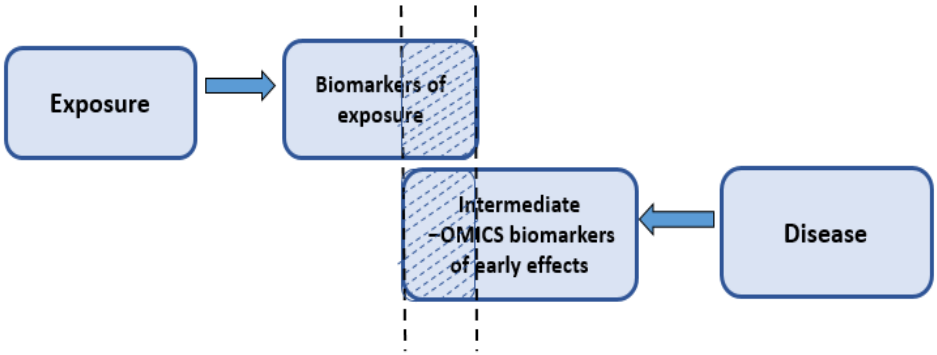

Figure 1.1: The "Meet-in-the-Middle" approach. Biological samples are characterized using OMICS platforms to identify markers that act both as markers of exposure and as marker of intermediate or early health effect. Such markers that appear to mediate the observed association between exposure and disease are used to explore which biological pathways connect exposure with disease endpoints.

OMICS technologies are relatively new biomarker discovery tools that can be applied to measure the internal exposome. The technologies aim for the detection, characterization, and quantification of (ideally) all biomolecules and complete biological processes potentially affected by external exposures. OMICS can contribute to greater understanding of the effect of environment on our biological system in several ways. First, it may directly detect biomarkers of exposure, such as metabolites or compounds in blood. OMICS may also identify which biological processes are affected by a certain exposure, such as smoking-induced changes in DNA-methylation. Optimally, OMICS may be able to generate hypotheses on the development of exposure-associated disease and yield biomarkers to predict disease development (DeBord et al., 2016; Bonassi et al., 2013). The use of OMICS in epidemiology promises to facilitate the increased use of biological evidence for the establishment of links between environmental exposure and disease causation (Kyrtopoulos, 2013). This is the bases of the so-called "Meet-in-the-Middle" concept (Figure 1.1), which starts with studying the association between exposure and disease, followed by establishing associations between exposure and OMICS markers and ultimately, investigations of the mediating link between disease and the mentioned intermediate biomarkers (Vineis et al., 2013). An example of this approach was published by Chadeau-Hyam et al. (2011) who identified a dietary colon cancer biomarker (a derivative of benzoic acid produced by fiber-digesting gut bacteria) by correlating prospectively measured metabolic profiles with both dietary fiber intake and reduced colon cancer risk.

Even though each OMICS platform only provides a partial snapshot of the internal exposome and different platforms may interact with each other, investigating each platform independently may provide an insight into the some part of the biological system. Such studies provide a basis for future explorations of the complex interactions between OMICS markers.

Two examples in Box 1.1 illustrate the idea of how the use of OMICS technology allows the investigation of biological pathway where perturbation may connect exposure to disease. Though OMICS markers have shown promises in strong environmental risk factors such as tobacco smoking, the application of OMICS markers to exposures that have lower levels or lower potencies such as air pollution is still to be demonstrated.

Box 1.1:

DNA methylation studies have shown links between smoking and lung cancer

It has been shown that smoking leads to DNA methylation changes measurable in peripheral blood (Baglietto et al., 2017). The detection of DNA methylation changes as a biomarker of tobacco smoke exposure has been highly successful, clearly identifying individuals with low levels of exposure (Joubert et al., 2014), and light smokers (Philibert et al., 2012), and also those with previous exposure such as former smokers (Guida et al., 2015; Joehanes et al., 2016; Shenker et al., 2013). Methylation changes at various loci, particularly the CpG (cytosine-guanine dinucleotide) in AHRR and F2RL3 are strongly associated with smoking (Guida et al., 2015). AHRR is the repressor of the aryl hydrocarbon which is a key regulator of the relationships between the cell and the external environment, including the effects of stressors such as dioxins and polycyclic aromatic hydrocarbons, both of which are present in tobacco smoke (Hankinson, 1995). AHRR is overexpressed in the lungs of smokers, implying a functional significance for the observed methylation changes and that upregulation of AHRR may be part of a defense against the functional consequences of AHR in toxicity pathways. In an independent study, it has been shown that hypomethylation in specific CpG sites of the AHRR and F2RL3 genes is associated with increased risk of subsequent lung cancer (Fasanelli et al., 2015). This indicates that the specific methylation alterations detected may mediate the carcinogenic effect of tobacco exposure in lung cancer aetiology.

Metabolomics studies have shown links between gut flora, diet, cardiovascular disease

Untargeted metabolomics contributed to the generation of a hypothesis for a previously unknown pathway linking dietary phosphatidylcholine (PC) to cardiovascular disease (CVD) (Wang et al., 2011b; Tang et al., 2013). In this example untargeted metabolomics was used in a prospective epidemiological study to search for blood plasma metabolic features associated with CVD risk. Three metabolites, including trimethylamine-N-oxide (TMAO), were associated with CVD and hypothesized to be derived from the major dietary lipid PC. This hypothesis was tested in an animal study by feeding mice egg-yolk PC and conducting untargeted metabolomics in the mice plasma (Koeth et al., 2013, 2014). Three mice analytes with the same mass-tocharge ratio (m/z) (76, 104 and 118) and retention times as the corresponding human metabolites increased after oral PC feeding. In further experiments, it was shown that the gut flora plays an important role in TMAO formation from dietary phosphatidylcholine (PC). In another independent epidemiological study, increasing levels of choline, TMAO, and betaine were observed all to show dose-dependent associations with the presence of CVD and multiple individual CVD phenotypes. The associations between increased risks of all CVD phenotypes monitored and elevated systemic levels of the three PC metabolites remained after adjustments for traditional cardiac risk factors and medication usage (Wang et al., 2011b). Furthermore, dietary choline or TMAO was shown to promote atherosclerosis. The use of untargeted metabolomics in an epidemiological study led to the generation of a new hypothesis: a novel pathway linking dietary lipid intake, intestinal microflora, and atherosclerosis, which was later confirmed in follow-up studies.

Focus on air pollution as high priority environmental pollutant

Air pollutants can cause significant harm to humans. According to the World Health Organization (WHO), around 7 million people die prematurely each year because of air pollution, making this a leading environmental health risk (WHO, 2018). Air pollution is recognized as a human carcinogen associated with lung cancer (Hamra et al., 2015; Loomis et al., 2013). It is also a major risk factor for other acute and chronic diseases including cardiovascular disease (CVD) and chronic respiratory diseases, such as asthma and chronic obstructive pulmonary disease (COPD) (Brunekreef and Holgate, 2002; Hoek et al., 2013). Although the exact mechanisms behind the observed associations have not clearly been delineated, oxidative stress and immune/inflammation responses are commonly considered as putative mechanisms underlying these health outcomes (Fiorito et al., 2018). It has been suggested that biomolecular changes may be detected by investigating the internal exposome (Rappaport, 2011; Wild, 2005). Here, I summarize the potential contributions of the OMICS marker in air pollution research as follows:

- OMICS markers involved in air pollution-induced biological pathways can be used to generate new hypotheses. The description of such a biological network will use insights from these biological pathways (e.g. inflammation, oxidative stress, altered coagulation) to identify air pollution induced health effects and the role of specific pollutants in these pathways.
- OMICS investigation can be used to identify whether different specific air pollutants may lead to perturbations in different biological pathways, because air pollutants are often generated as a mixture and it is difficult to disentangle which of them have the most pronounced detrimental effect on health.
- The use of OMICS to detect biomarkers of early disease can provide underpinning of the existing evidence of causal association between air pollutants and health effects.

Although many studies provided initial clues for the potential impact of air pollution on various immune markers such as C-reactive protein (CRP), fibrinogen and interleukin-6 (Hoffmann et al., 2009; Panasevich et al., 2009; Steenhof et al., 2013), evidence from other untargeted OMICS scales is scarce. Early studies have provided indications for an impact of air pollution on peripheral blood global methylation using repetitive elements in long (LINE-1) and short (ALU) interspersed nuclear element (Baccarelli et al., 2009; Bellavia et al., 2013; De Prins et al., 2013). Sequence specific analyses have yielded evidence for differential methylation in regions of genes such as inducible nitric oxide synthase (iNOS), tissue factor intercellular adhesion molecule 1, toll-like receptor 2, interferon-γ, and interleukin-6 (Bind et al., 2014b; Madrigano et al., 2012; Tarantini et al., 2009). After the introduction of OMICS, genome-wide association studies (GWAS) and peripheral blood methylome-wide comparison studies have been performed on population exposed

to different levels of air pollution (Gruzieva et al., 2017; Huang et al., 2010; Panni et al., 2016; Pettit et al., 2012; Plusquin et al., 2017). Although these studies indicate the impact of specific air pollutants on the human biological system, the results are inconclusive. This may partially be explained by small study sizes, imprecise air pollutant exposure estimates, and less than optimal statistical approaches.

Precise air pollutant exposure estimates

The first step towards a role for OMICS markers in air pollution health research is to firmly establish the impact of air pollution on OMICS markers. Since the impact of air pollution on OMICS markers is expected to be modest, high-quality exposure assessment is crucial for the detection of air pollution-OMICS associations.
What is considered high-quality exposure assessment depends on the scenario in which it is applied. In smaller studies focused on the acute effects of air pollution, personal monitoring collected in the breathing zone is considered the gold standard. Personal monitoring approaches are capable of capturing variations in exposure from factors such as different activity patterns and varying traffic intensities. Downsides of personal monitoring include the large burden to the study participants and the limited time period over which exposure measurements can feasibly be collected (Nieuwenhuijsen et al., 2006). To assess the risk of air pollution exposure in large, population-level epidemiological, one has to resort to exposure modeling. One of the increasingly applied air pollution modeling techniques is Land use regression (LUR) which is a geographical information system (GIS) and statistics based method exploiting land use, geographic and traffic characteristics to explain concentration contrasts at monitoring sites (Hoek et al., 2008). State of the art LUR models have become available in recent years that cover most of Europe and provide exposure estimates for several different air pollutants (Beelen et al., 2013; Eeftens et al., 2016). These models have been successfully applied in large air-pollution health studies (Beelen et al., 2014; Hoek et al., 2008) and are available to assess the chronic effect of air pollution on OMICS markers.

Statistical models to analyze OMICS data

OMICS data are high-dimensional-the numbers of variables are large and can exceed the number of study subjects. This has an impact on the statistical methods that can be applied to infer which markers are perturbed. Following the Meet-in-the-Middle paradigm, in statistical analyses OMICS markers can be considered both as the outcome in exposure-OMICS analyses or the predictor in OMICS-disease analyses. Which of these two comparisons one is interested in has an impact on the statistical models that are available. Considerable work has been conducted in method development for so-called variable selection in the OMICS-disease setting and the performance of several of these approaches

has been compared (Agier et al., 2015; Chadeau-Hyam et al., 2013; Lenters et al., 2018). Examples of better-performing methods include dimension reduction and variable selection methods. Dimension-reduction methods aim to summarize the information contained in large data sets into a few synthetic variables that capture the latent data structure. Principle component analysis and partial lease square (PLS) are examples of these methods that have frequently been applied in the analysis of OMICS data (Belshaw et al., 2010; Fasoli et al., 2012; Wang et al., 2011a). Variable selection techniques pursue the selection of a subset of variables in the OMICS data that predict the outcome of interest and can be seen as extension of the dimension reduction methods. Elastic net (Zou and Hastie, 2005) and Graphical Unit Evolutionary Stochastic Search Algorithm (GUESS) are the examples of this approach (Bottolo et al., 2013).

Less work has been done on the development of methods for exposure-OMICS analyses, resulting in the continued use of univariate regression models. Univariate models are flexible and computationally efficient methods to analyze high-dimensional OMICS data. Univariate approaches assess the association between each variable in the predictor matrix (i.e. air pollution) and the outcome of interest (i.e. OMICS data) independently. Naturally, when large numbers of independent tests are performed (e.g. $\sim$450.000 in the case of DNA methylation analyses in this thesis), each with the same significance level α, the overall number of falsely rejected null hypothesis is inflated, leading to greater type I error rates. In order to control the inflated number of false-positive findings, some form of p-value correction, such as False Discovery Rate (FDR) described by Benjamini and Hochberg (1995), q-value described by Storey Storey (2002), or the Bonferroni correction (Dunn, 1961) have to be implemented. Implementation of correction techniques is, however, associated with significant losses in statistical power to detect true positive findings. One proposed way to increase statistical power in methylation associations is reducing the dimension of the methylome into clusters of genomically proximal CpG (cytosine-guanine dinucleotide) sites exhibiting correlated methylation levels; these clusters are defined as differentially methylated regions (DMR) (Wright et al., 2016).

Although confounding in these analyses can never be fully excluded, outcome-wide epidemiology, of which exposure-OMICS analyses are an example, has been suggested to be less sensitive to confounding than exposure-wide epidemiology (VanderWeele, 2017) making the use of multivariate models less crucial than in the situation of OMICS-disease analyses.

In general, there are limited guidelines on determining the 'best' model in both exposure-OMICS and OMICS-disease analyses. Instead of selecting a single model, multiple models can be complemented to improve interpretation and provide insights in method uncertainties (Agier et al., 2015; Lenters et al., 2018).

Mediating effect of OMICS

The final step of the Meet-in-the-Middle concept is to explore whether OMICS markers that have been demonstrated to be associated to both exposure and health outcome mediate the effect of exposure on the outcome. This would provide more insight into whether the identified OMICS markers are on the molecular pathway linking exposure to health outcome.

Mediation analyses for a single factor, either as a marker or exposure, in a univariate setting is well developed. The univariate regression-based mediation analysis can be grouped into three approaches: (a) causal steps, (b) difference in coefficients, and (c) product of coefficients (MacKinnon et al., 2007), all of which are based on information from the regression equations for testing the single mediator model.

While the univariate approach is widely implemented for single mediator analyses, it should be used with care as it is highly dependent on a number of strong assumptions, the measurement characteristics of the variables, and on reliable identification of causal effects (Hayes, 2009).

Methods of mediation analyses in a high-dimensional setting are less developed. One example of a method capable of conducting mediation analyses based on OMICS markers is partial least square path modeling (PLS-PM) (Sanchez, 2013). PLS-PM incorporates information from full sets of OMICS markers and exposure into a few synthetic variables, known as latent variables, which capture mediating effect of OMICS marker in the association between exposure and health outcome. PLS-PM minimizes the biases of interference, multicollinearity, missing data, and the assumption of sample distribution (Shen and Lung, 2017).

Biological Interpretation of statistical models outcome

An important last step of using OMICS markers to study the internal exposome is the biological interpretation of statistical model outcomes. Several tools are available to make the translation from interpretation on marker level to the level of biological pathways or systems.

Ontology-based tools that interrogate existing databases, such as the NIH-DAVID (Huang et al., 2009), are rich sources of information for inferring biological pathways corresponding to the candidate biomarkers identified. Specifically, gene set enrichment analyses (GSEA) assess if, and to what extent, the list of candidate transcripts are enriched for specific pathways by essentially checking if the distribution of identified transcripts is significantly different from what would be expected if candidate transcripts were chosen at random (Ashburner et al., 2000; Subramanian et al., 2005).

Biological interpretability of molecular alterations identified in high throughput profiling is highly dependent on the functional characterisation of the assayed molecules. For some molecular entities, knowing the functional role of the molecule may aid interpreta-

tion. However, interpretability can become impossible when the function of the putative biomarker is unknown.
Typically, interpretation of the results of a targeted proteomic assay is easier as the biological functionality is likely to be known or suspected. However, overall gene expression regulation may involve several transcripts and occur through complex cascade involving other genes and transcripts. Hence, direct interpretation of the results from a transcriptome-wide association study should be complemented by an investigation of all biological pathways potentially affected by differential expression. For DNA methylation data, results interpretation can be even more challenging as the connection between CpG site-specific levels to their downstream consequences is complex. Some insight into biological interpretation might be gained by statistically linking differentially methylated CpG sites to other better characterized OMICS data, such transcriptomics.

Outline of the thesis

The overall aim of this dissertation is to study three OMICS platforms and their regulatory mechanisms in order to have a better understanding of the biological processes that link air pollution exposure to health effects. The overview of the separate chapters within this thesis are as follows:
The first two chapters focus on effects of annual average estimate of nitrogen dioxide (NO_X) concentrations by comparing two relatively large prospective cohorts from Italy and Sweden, in which the same exposure assessment strategy (i.e., LUR) was applied. In **Chapter 2**, perturbation in plasma concentration of a large panel of cytokines, chemokines, and growth factors was studied; in **Chapter 3** genome-wide changes in gene expression in peripheral blood mononuclear cells was studied in relation to air pollution. The objective of **Chapters 4** is to explore the relationship between 24hr personal exposure to traffic-related particulate matter ($PM_{2.5}$) and ultrafine particles (UFP) and perturbation of genome-wide DNA-methylation in healthy non-smoking individuals from 4 European countries using a repeated measurements design. In order to reduce dimensionality and improve power to detect associations, DMRs clusters of CpGs were defined and analyzed. Additionally, we explored the association of air pollution-induced methylation alterations with gene expression and plasma-levels of immune markers, all measured in the same subjects using integrative analysis. **Chapter 5** aims to investigate to what extent inflammatory markers mediate the association between long-term exposure to air pollution and adult-onset asthma using a nested-case control study design in a prospective cohort from Switzerland. The thesis ends with **Chapter 6**, in which the main findings are reviewed in light of limitations and results from previous studies. Future perspectives and challenges in OMICS research are also discussed.

CHAPTER 2

INFLAMMATORY MARKERS IN RELATION TO LONG-TERM AIR POLLUTION

This chapter is published as: Mostafavi N, Vlaanderen J, Chadeau-Hyam M, Beelen R, Modig L, Palli D, Bergdahl I.A, Vineis P, Hoek G, Kyrtopoulos S., Vermeulen R. Inflammatory markers in relation to long-term air pollution. *Environment International*, 2015 Aug;81:1-7. doi: 10.1016/j.envint.2015.04.003.

LONG-TERM exposure to ambient air pollution can lead to chronic health effects such as cancer, cardiovascular and respiratory disease. Systemic inflammation has been hypothesized as a putative biological mechanism contributing to these adverse health effects. We evaluated the effect of long-term exposure to air pollution on blood markers of systemic inflammation.

We measured a panel of 28 inflammatory markers in peripheral blood samples from 587 individuals that were biobanked as part of a prospective study. Participants were from Varese and Turin (Italy) and Umeå (Sweden). Long-term air pollution estimates of nitrogen oxides (NO_X) were available from the European Study of Cohorts for Air Pollution Effects (ESCAPE). Linear mixed models adjusted for potential confounder were applied to assess the association between NO_X and the markers of inflammation.

Long-term exposure to NO_X was associated with decreased levels of interleukin (IL)-2, IL-8, IL-10 and tumor necrosis factor-α (TNF-α) in Italy, but not in Sweden. NO_X exposure levels were considerably lower in Sweden than in Italy (Sweden: median (5^{th}, 95^{th} percentile) 6.65 $\mu g/m^3$; (4.8, 19.7); Italy: median (5^{th}, 95^{th} percentile) 94.2 $\mu g/m^3$ (7.8, 124.5)). Combining data from Italy and Sweden we only observed a significant association between long-term exposure to NO_X and decreased levels of circulating IL-8.

We observed some indication for perturbations in the inflammatory markers due to long-term exposure to NO_X. Effects were stronger in Italy than in Sweden, potentially reflecting the difference in air pollution levels between the two cohorts.

Introduction

Epidemiologic studies have consistently shown an association between long-term exposure to ambient air pollution and cardiovascular mortality and morbidity, non-malignant respiratory diseases (Brunekreef and Holgate, 2002; Hoek et al., 2013), and lung cancer (Demetriou et al., 2012). Although the exact mechanisms behind the observed associations are not clear, long-term pulmonary oxidative stress and inflammation induced by chronic exposure to inhaled pollutants has been hypothesized to result in a systemic inflammatory state capable of activating hemostatic pathways, impairing vascular function, and accelerating atherosclerosis (Brook et al., 2010; Rückerl et al., 2007).

The link between the inflammatory process in the lung and the systemic response is thought to be mediated by markers of inflammation released by alveolar macrophage and bronchial epithelial cells in response to exposure to ambient air pollution and capable of entering the systemic circulation and stimulating the production of acute-phase proteins (Demetriou et al., 2012; Hoffmann et al., 2009). Inflammatory markers in peripheral blood might therefore reflect deregulation resulting from chronic exposure to ambient air pollution (Dubowsky et al., 2006; van Eeden et al., 2001).

A number of studies have reported on acute changes in blood markers of inflammation in response to day-to-day variability in air pollution. These studies have generally reported inconsistent results. Effects of short-term exposure to ambient air pollutants on the levels of fibrinogen have been reported, either with a negative association (in response to PM_{10} or SO_2) (Panasevich et al., 2009; Seaton et al., 1999; Steenhof et al., 2013), or a positive association (in response to PM_{10} or O_3) (Rückerl et al., 2007). Short-term exposure to particulate matter has been associated with an increase in IL-6 levels (van Eeden et al., 2001) or shown no relationship (Panasevich et al., 2009; Seaton et al., 1999; Zuurbier et al., 2011). A positive association with TNF-α levels (in response to PM_{10}, NO_2) was reported by Panasevich et al. (2009); Tsai et al. (2012), but Larsson et al. (2013) reported no association with diesel exhaust.

We identified four studies that reported on the association between markers of long-term exposure to air pollution and chronic perturbations of blood inflammatory markers (Table S2.1) (Chuang et al., 2010; Forbes et al., 2009; Hoffmann et al., 2009; Panasevich et al., 2009). All studies assessed a small set of blood markers of inflammation. Panasevich et al. (2009) assessed four markers (IL-6, TNF-α, CRP, fibrinogen) and reported significantly higher IL-6 levels and positive effect estimates of CRP levels after long-term exposure to elevated residential levels of 30-year average traffic-related NO_2. Hoffmann et al. (2009) assessed two markers of inflammation (CRP, fibrinogen) and reported a positive association between annual particulate matter ($PM_{2.5}$) and CRP and fibrinogen (an increase of 16.7% in CRP and 2.4% in fibrinogen for a unit increase in $PM_{2.5}$). Forbes et al. (2009) also explored the association between estimates of long-term exposure to ambient air pollution and fibrinogen and CRP, but observed no associations. Chuang et al. (2010) assessed two

markers of inflammation (IL-6 and neutrophils) and reported that an increase in annual average particulate air pollutant (PM_{10} and $PM_{2.5}$) and NO_2 was marginally (p-value $<$ 0.1) associated with elevated IL-6 and neutrophils.
We assessed the association between long-term exposure to nitrogen oxides (NO_X) and plasma concentration of a large panel of cytokines, chemokines, and growth factors in a sample of the general population. Our study contributes to the existing literature by combining a wide range of inflammatory markers (n=28) measured in a large number of individuals (n=587) with state of the art long-term air pollution exposure assessment. By comparing two prospective cohorts from Italy and Sweden, in which we applied the same exposure assessment strategy, we were able to study the association over a wide-range of air pollution exposure and assess the between-cohort heterogeneity of our findings. In addition, in a sensitivity analysis, we restricted our analyses to the elderly and the overweight, to assess whether we observed stronger effects among individuals that have been reported to have a higher susceptibility to develop air pollution induced cardiovascular (Bentayeb et al., 2012) and other health effects (Dubowsky et al., 2006; Rückerl et al., 2007; Simoni et al., 2015).

Materials and methods

For the current study we combined data from two existing projects: inflammatory markers from Genomics Biomarkers of Environmental Health (EnviroGenoMarkers) (Chadeau-Hyam et al., 2014; Hebels et al., 2013) and NO_X data from the European Study of Cohorts for Air Pollution Effects (ESCAPE) (Beelen et al., 2014).

Study population

EnviroGenoMarkers was set up to investigate the association between a set of environmental agents (polychlorinated biphenyls, polycyclic aromatic hydrocarbons, cadmium, lead, phthalates, brominated flame retardants, ambient air pollutants and water treatment byproducts), intermediate OMICS markers (metabolomics, epigenomics, proteomics and transcriptomics), and various human diseases (Chadeau-Hyam et al., 2014). It has a nested case-control design, using biosamples of two prospective cohorts: the Italian contribution to the European Prospective Investigation into Cancer and Nutrition study (EPIC-Italy) (Palli et al., 2003), and the Northern Sweden Health and Disease Study (NSHDS) (Hallmans et al., 2003). EPIC-Italy was initiated in 1992 and completed the recruitment of 15,171 men and 32,578 women healthy middle-aged (35-70 years old) volunteers in 4 different areas of Italy: Varese, Turin, Ragusa, and Florence (Palli et al., 2003). The NSHDS cohort contains 3 sub-cohorts: the Västerbotten Intervention program (VIP), the Västerbotten Mammary Screening (MS) Program and the Northern Sweden MONICA project. It started in 1985 and the total cohort and biobank contains at present 85,000 unique healthy

individuals (Hallmans et al., 2003). At the time of recruitment (around 2002), cohort members from both studies completed a standardized questionnaire focusing on dietary and life-style habits and donated blood.
The EnviroGenoMarkers data included in the current study were collected in two phases. In the first phase 100 Non-Hodgkin's lymphoma cases, 100 breast cancer cases, were identified through local Cancer Registries (loss to follow-up $< 2\%$) and occurred on average 6 years (range 1 to 17 years) after recruitment/blood collection, and the same number of controls matched on sex, age, center and date of blood collection were included (Chadeau-Hyam et al., 2014). The lymphoma case-control data were subsequently supplemented with samples from additional cases (147 cases in NSHDS, 34 cases in EPIC-Italy) and equal number of similarly matched controls (phase 2) (Chadeau-Hyam et al., 2014; Kelly et al., 2013).
Our study population, a subset of the EnviroGenoMarkers data, includes 97 individuals (23 men, and 74 women) from two centers participating in the EPIC-Italy cohort (Turin and Varese) and 490 individuals (210 men, and 280 women) from the NSHDS cohort (Umeå). Our selection from the NSHDS cohort includes 195 future Non-Hodgkin's lymphoma cases (49 from phase 1 and 146 from phase 2), 50 future breast cancer cases, and 245 controls matched on sex, age, center and date of blood collection. Our selection from the EPIC-Italy cohort includes 38 future Non-Hodgkin's lymphoma cases (24 from phase 1 and 14 from phase 2), 12 future breast cancer cases, and 47 controls matched on sex, age, center and date of blood collection. The average time to diagnosis for breast cancer cases was 6 years (range 2 to 10 years) and for lymphoma cases was 6 years (range 2 to 16 years) after blood collection.

Cytokine measurements

Plasma levels of a panel of 28 inflammatory markers including interleukin (IL) 1β, IL-2, IL-4, IL-5, IL-6, IL-7, IL-8, IL-10, IL-12, IL-13, interferon alpha (INF-α), interferon gamma (INF-γ), tumor necrosis factor alpha (TNF-α), eotaxin, IL1 receptor antagonist (IL1ra), interferon gamma-induced protein 10 (IP10), granulocyte-macrophage colony-stimulating factor (GMCSF), epidermal growth factor (EGF), fibroblast growth factor 2 (FGF2), fms-like tyrosine kinase receptor-3 (Flt3) ligand protein (Flt3 ligand), granulocyte colony-stimulating factor (GCSF), melanoma growth stimulatory activity/growth-related oncogene (GRO), monocyte chemotactic protein-1 (MCP-1), monocyte chemotactic protein-3 (MCP-3), macrophage derived chemokine (MDC), macrophage inflammatory protein-1 alpha (MIP-1), macrophage inflammatory protein-1 beta (MIP-1β), soluble CD40 ligand (sCD40L), soluble IL-2 receptor alpha (sIL2Rα), transforming growth factor alpha (TGF-α), and vascular endothelial growth factor (VEGF) were measured in stored samples of all subjects using the milliplex hcytomag-60 19.Hum and hcytomag-60SPMX13 kits, according to the protocol described by the manufacturer (Hosnijeh et al., 2012). Quality control

samples were run in duplicate with the study samples in each batch (Hosnijeh et al., 2010). Samples from EPIC-Italy contained citrate as anticoagulant and had been stored in cryostraws in liquid nitrogen for 1119 years. Their recorded collection-to-storage times were 55347 min. Samples from NSHDS contained heparin or EDTA as anticoagulant and had been stored in plastic cryovials at $-80°C$ for 419 years. Blood samples were generally collected in the morning and their collection-to-storage time was always < 1 hour (Hebels et al., 2013). A previous analysis showed that samples not cold-stored within 2 h after blood collection had significantly different expression profiles than fresh samples, and therefore only peripheral blood mononuclear cells samples that had been placed in cold storage within 2 h after blood collection were included in the current study (Chadeau-Hyam et al., 2014).

Exposure assessment

The European Study of Cohorts for Air Pollution Effects (ESCAPE) study was set up to investigate the effect of long-term exposure to ambient air pollution on human health in 15 European countries (Beelen et al., 2014; Cyrys et al., 2012). ESCAPE measurements, includes 36 cohorts which study areas cover Europe. Exposure levels were estimated using land use regression models that were developed using exposure measurements conducted between October 2008 and April 2011 (Beelen et al., 2014). Study area specific land use regression models were used to assign estimates of long-term exposure to the cohort participants home-address. Land use regression models were developed combining exposure measurements conducted at a restricted number of sites (40 to 80) covering each study area with Geographic Information System (GIS) based predictors. Exposure to nitrogen oxides (NO_2 and NO_X) was assessed for all study areas participating in ESCAPE, while exposure to particulate matter ($PM_{2.5}$, $PM_{2.5}$ absorbance and PM_{10}) was assessed for a subset of study areas that contributed to the contrast in exposure in the ESCAPE dataset. Of the three study areas that were included in EnviroGenoMarkers (Umeå, Turin, and Varese) particulate matter estimates were only available for 13 subjects from Turin (Beelen et al., 2014). We therefore based our analyses on NO_X. In the cities included in the current analysis the squared correlations (R^2) between NO_X and NO_2 ranged between 0.94 and 0.97. The NO_2/NO_X ratio was 0.57, 0.54, 0.53 for Umeå, Turin, Varese areas, respectively (Cyrys et al., 2012).

Data analysis

For four markers of inflammation (IL-12, IL1-RA, sIL2-RA, and Flt3 ligand) more than 50% of the samples were below the limit of detection (LOD). We excluded these markers from our analyses. Data were imputed when measurements were out of range of the calibration curve (either too low: <LOD, or too high) based on a maximum likelihood estimation

procedure (Lubin et al., 2004). For imputation of samples <LOD we imputed using the empirical LOD across all plates as the upper bound. For imputation of samples with a concentration exceeding the calibration curve we used a value of the twice the highest observed concentration that was not out of range as the upper bound. Among the analytes that were retained, the maximum percentage of imputed values was 47.7% (IL-13), while 70% of the retained markers had less than 30% imputed values.

For six individuals in the Varese population unrealistically low NO_X exposure estimates (<0.16 $\mu g/m^3$) were predicted by the ESCAPE model. We set these values at the 2.5^{th} percentile of the distribution of NO_X exposure estimates in the full Varese study population also including individuals for which markers of inflammation were not assessed (7.81 $\mu g/m^3$). We assessed the sensitivity of this decision by replacing the 2.5^{th} percentile by either the 1.5^{th} or the 5^{th} percentile of the distribution and observed that our results were robust.

We conducted linear mixed-effects modeling to investigate the association between long-term exposure to NO_X and markers of inflammation. We included random intercepts for microtiter plate in the model to capture nuisance variation generated in the assessment of the inflammatory markers (clustering of the inflammatory markers measurements by plate) (McHale et al., 2011) and included fixed-effects for NO_X and for a priori selected potential confounding factors, i.e., body mass index (BMI) (Kg/m^2), age (in three categories: (30,40], (40,50], (50,75], years), sex (female, male), smoking status (never, former, current), phase (1 and 2), future disease status (lymphoma case, breast cancer case, control) and sample storage time (years) consistent with previous analyses of the EnviroGenoMarkers data (Chadeau-Hyam et al., 2014; Kelly et al., 2013). As a sensitivity analysis, we explored the potential for confounding by the remaining factors available in our dataset including socio-economic status (education level (primary, technical, secondary, university)), and physical activity (moderately inactive, moderately active, active).

Due to differences in dietary habits, lifestyle, and air pollution exposure levels between the Italian and Swedish cohorts, we stratified our analyses by cohort (Table 2.1; Figure 2.1). As there was some overlap in the exposure distributions of the two cohorts, we also conducted analyses on the combined cohorts, while adjusting for country (Figure 2.1).

As sensitivity analysis we explored the association between NO_X and markers of inflammation in two potentially susceptible strata: among overweight individuals (BMI>25; n=323) and among individuals older than age 50; n=383. In addition, we performed analyses among never smokers (n=353) to assess the effect of potential residual confounding by active smoking, and among controls only (n=292) to assess the effect of potential bias due to early manifestations of future disease.

We natural log transformed NO_X and markers of inflammation to limit the influence of high concentrations and normalize distributions. We computed unadjusted p-values and adjusted p-values for multiple testing (q-values) by controlling the false discovery rate (FDR) at 5%. We used penalized splines (P-spline) in the generalized additive mixed-

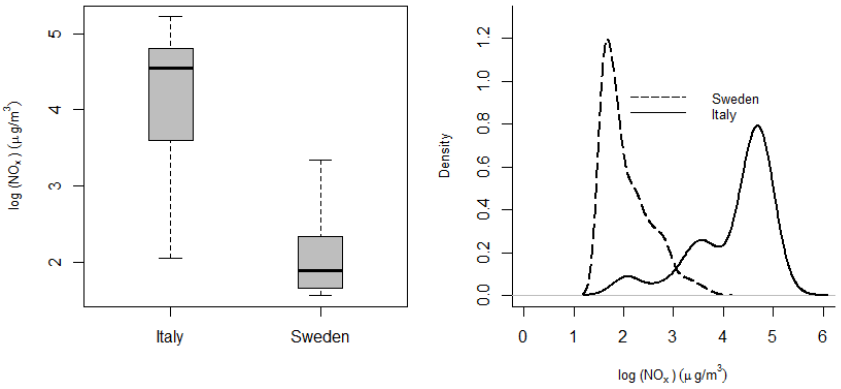

Figure 2.1: Box plot (left) and density plot (right) of log (NO_X) ($\mu g/m^3$) exposure levels observed in the Swedish and Italian cohorts.

model (GAMM) framework to assess potential non-linearity of the relationships between inflammatory markers and exposure.
Correlation coefficients between all markers of inflammation were assessed using Pearson correlation. A two-sided q-value < 0.05 was considered statistically significant. Statistical analyses were performed using R version 3.0.2 (*package lme4*).

Results

We summarized the general characteristics of the study population by cohort in Table 2.1. The Swedish cohort had a lower proportion of women than the Italian cohort (57% versus 76%), a higher proportion of current smokers (21% versus 8%) and a lower proportion of never smokers (59% versus 66%). We observed a considerable difference in the distribution of NO_X concentrations between the two cohorts (Table 2.1; Figure 2.1). The median (5^{th}, 95^{th}) concentration of NO_X estimated for the Italian cohort (94.2 $\mu g/m^3$ (7.8, 124.5)) was considerably higher than the median concentration estimated for the Swedish cohort (6.65 $\mu g/m^3$ (4.8, 19.7)) (p-value < 0.001).

Univariate mixed-effect regression analyses yielded evidence for an association between long-term exposure to NO_X and several markers of inflammation. In Table 2.2 we report the significant effects and in Table S2.2 we report the results for the complete set of inflammatory markers. Our strongest finding is for IL-8, for which we observed a significant negative association in the Italian cohort as well as in the combined cohort (q-values of 0.001 and 0.02, respectively). For an inter quartile change in NO_X concentration in the

Table 2.1: Descriptive characteristics of the study population.

Characteristics		Swedish cohort (n=490)	Italian cohort (n=97)
Sex (N (%))			
	Female	280(57%)	74(76%)
	Male	210(43%)	23(24%)
Smoking Status (N (%))			
	Current-smoker	104(21%)	8(8%)
	Former-smoker	97(20%)	25(26%)
	Never-smoker	289(59%)	7464(66%)
Age (years)(N (%))			
	<40	29(6%)	2(2%)
	40-50	143(29%)	30(31%)
	>50	318(65%)	65(67%)
BMI[b] (kg/m2) (Mean SD[a])		26.14.1	25.7(3.8)
NO_X^c (g/m3) (Median (SD[a])		6.65 (5.84)	94.2(42.4)

Note: [a] SD, Standard deviation; [b] BMI, Body mass index; [c] NO_X, Nitrogen oxide

combined cohort (10 $\mu g/m^3$) IL-8 decreased 5.1 pg/mL (17% decrease). Although a negative association between NO_X and IL-8 was observed in the Swedish cohort as well, this association was not significant (q-value = 0.655). For several other markers we observed significant negative associations with long-term exposure to NO_X in the Italian cohort, but not in the Swedish or combined cohorts: IL-2 (q-value = 0.012), TNF-α (q-value = 0.020), IL-10 (q-value = 0.017). For IL-10 the non-significant association observed in the combined cohorts was in the same direction as it was in the Italian cohort (q-value 0.083).
We observed little influence of the potential confounding factors on the effect of NO_X on the markers of inflammation (the effect size did not change after adjustment for the potential confounding factors).

Results from sensitivity analyses among never-smokers (Table 2.3 for the significant ef-

Table 2.2: Significant effects of long-term exposure to NO_X on markers of inflammation adjusted for potential confounders.

	Combined cohort (n=587)			Swedish cohort (n=490)			Italian cohort (n=97)		
Markers	β^a	p-value	q-value[c]	β^a	p-value	q-value[c]	β^a	p-value	q-value[c]
IL[d]-2	-0.058	0.623	0.703	-0.019	0.896	0.896	-0.421	0.003	0.012
IL-8	-0.179	0.008	0.020	-0.079	0.295	0.655	-0.520	0.000	0.001
IL-10	-0.256	0.028	0.083	-0.104	0.467	0.619	-0.690	0.002	0.017
TNF-α	-0.028	0.620	0.643	0.036	0.625	0.689	-0.213	0.002	0.020

Results are based on linear mixed-effects models of log-transformed dependent and independent variables with fixed-effects: body mass index (BMI) (Kg/m^2), age (in three categories: (30,40], (40,50], (50,75]; years), sex (female, male), smoking status (never, former, current), phase (1 and 2), future disease status (lymphoma case, breast cancer case, control), and sample storage time (years). And plate as random effects. [a]effect estimate per unit changes of the exposure; [c]q-value, False Discovery Rate correction for p-value, [d]Interleukin (IL)

fects and Table S2.3 for the complete set of markers) and among controls only (see supplemental material, Table S2.7) were similar, though with larger standard errors, compared to the results from our main analyses, suggesting limited impact of potential residual confounding of smoking or future (lymphoma or breast cancer) case status. We assessed the impact on our results of additional correction for further covariates available in the EnviroGenoMarkers dataset, but observed no considerable changes in effect estimates (<5% for the markers of inflammation for which we observed a significant association with NO_X; results not shown). We did not observe any evidence for deviation from linearity for the markers significantly associated with NO_X at the log scale (results not shown). The associations between NO_X and the inflammatory markers within different subgroups of the combined population are presented in Table 2.3 (cohort-specific results are presented in supplemental material, Tables S2.4, S2.5, and S2.6). In the stratified model, the size of the effect of NO_X on several inflammatory markers (especially IL-8) was often larger in overweight individuals than in non-overweight individuals. We observed smaller effect sizes of NO_X on the markers of inflammation among elderly individuals. However, we did not observe a significant interaction between NO_X and BMI or age in the unstratified models.

Table 2.3: Significant effects of long-term exposure to NO_X on markers of inflammation within 3 sub-groups (overweight, elderly, and never-smokers).

	Overweight (n=323)			Elderly (n=383)			Never-smokers (n=353)		
Markers	β^a	p-value	q-value^c	β^a	p-value	q-value^c	β^a	p-value	q-value^c
IL^c-2	-0.073	0.613	0.843	-0.080	0.566	0.778	0.104	0.502	0.628
IL-8	-0.254	0.001	0.004	-0.039	0.619	0.847	-0.113	0.226	0.282
IL-10	-0.270	0.040	0.073	-0.301	0.026	0.071	-0.180	0.262	0.427
TNF-α	0.032	0.629	0.766	0.012	0.866	0.951	0.017	0.803	0.882

Results are based on linear mixed-effects models of log-transformed dependent and independent variables with fixed-effects: sex (female, male), phase (1 and 2), sample storage time (years), and future disease status (lymphoma case, breast cancer case, control). And plate as random effects.
[a]effect estimate per unit changes of the exposure; [c]q-value,
False Discovery Rate correction for p-value, [d]Interleukin (IL)

Discussion

This study provides some evidence for a perturbation of systemic inflammatory markers due to long-term exposure to NO_X in a cohort of healthy Italian and Swedish individuals. We observed significantly decreased levels of IL-8 in the combined cohort and for three other inflammatory markers in the Italian cohort, but observed no evidence for any association in the Swedish cohort.

Our finding of an inverse association between long-term exposure to air pollution and IL-2, IL-8, IL-10, and TNF-α is in contrast to results from most previous studies of long-term exposure to air pollution and markers of inflammation, which have generally reported positive associations (Hofmann et al., 2011; Panasevich et al., 2009). Similarly, studies focused on the biological mechanisms of these markers of inflammation have reported a primarily pro-inflammatory role. We summarize the current knowledge below. IL-2 performs critical functions for the elimination of diseased cells, including promotion of T and natural killer (NK) cells, cytolytic activities, and regulation of naive T cell differentiation into Th1 and Th2 subsets upon exposure to antigens (Liao et al., 2013). IL-8 is produced under inflammation stimulation and is attracting and activating neutrophilic granulocytes (Zarogoulidis et al., 2014). TNF-α is involved in the innate immune response (Clark, 2007). IL-10 is an immunoregulatory cytokine mainly secreted by macrophages, but also by T helper 1 (Th1) and Th2 lymphocytes, dendritic cells, cytotoxic T cells, B lymphocytes, monocytes and mast cells. IL-10 inhibits the capacity of monocytes and macrophages to present antigen to T cells and therefore downregulates the expression of IL-1β, IL-6, IL-8, IL-12 and TNF-α (Trifunović et al., 2015). Although we do observe downregulation of IL-8 and TNF-α in our study, this finding does not correspond to the observation that IL-10 is downregulated as well. Interestingly, a study by Forbes et al. (2009) also observed negative associations between long-term exposure to air pollution and markers of inflammation (fibrinogen and C-reactive protein), triggering Forbes et al. (2009) to hypothesize that the health effects of chronic outdoor air pollution are not mediated by systemic inflammation. Some short-term air pollution studies have also reported inverse associations with selected markers of inflammation (e.g. Seaton et al. (1999) (Fibrinogen) and Rückerl et al. (2007) (IL-6)). Reasons for these inverse associations remain unclear. Furthermore, similar inverse patterns have been observed in relation to other exposures such as endotoxin (Lauw et al., 2000) and carbon monoxide (Morse et al., 2003) and can be seen as indicative of a de-regulation of the inflammatory system.

Our study was subject to some (potential) limitations. One potential limitation is related to the air pollution exposure metric we used in our analyses, as only NO_X exposure estimates were available. Although the correlation of spatial variation in NO_X and particulate matter is often high (and was high ($\rho = 0.83$) in the ESCAPE study from which we derived our exposure estimates (Eeftens et al., 2012b)), generally stronger associations with chronic health effects have been observed for particulate matter (especially $PM_{2.5}$) than for NO_X (Beelen et al., 2014) and particulate matter has been suggested to be a better marker for ambient air pollution than NO_X (Hoek et al., 2013). Another limitation of our exposure metric, related to the design of ESCAPE, is the fact that current exposure levels were used to assign exposure for historical home addresses (in some cases more than 20 years retrospectively), which might have introduced a certain degree of measurement error. However, several studies have documented that the land use regression models can be applied successfully to estimate air pollution concentrations several years forwards or

backwards in time. Although the absolute level of exposure to air pollution has generally decreased over time, spatial contrasts for NO_2 have been shown to remain stable over long periods of time (10 years and longer) (Cesaroni et al., 2012; Gulliver et al., 2013; Wang et al., 2013; Eeftens et al., 2011). Note, we performed a sensitivity analyses within the Swedish part of the study using back-extrapolated NO_X exposure levels at time of the blood draw using the methodology described in (Beelen et al., 2014). Results of these analyses did not provide any different results. The markers of inflammation that we included in our analysis are subject to variability over time (e.g. diurnal and seasonal variation), likely contributing to increased standard errors (Forbes et al., 2009). Several studies have shown examples of considerably larger between-person variability in markers of inflammation than within-person variability, suggesting that single measurements of circulating inflammatory markers provide information about long-term perturbation of the inflammatory markers (Hofmann et al., 2011; Navarro et al., 2012).

Due to the nested case-control design of the EnviroGenoMarkers study our study population included future cancer cases (individuals that were diagnosed with lymphoma or breast cancer in the years following biological sampling) and matched controls. A potential implication of this design for our study is a perturbation of the inflammatory markers due to pre-diagnostic manifestation of the disease among cases. However our results did not change significantly when we restricted our analyses to controls only. In addition, the fact that our study population cannot be viewed as a random sample from the general population might have reduced the external validity of our findings (but should not have any impact on the internal validity).

We observed stronger associations in the Italian cohort than in the Swedish cohort. Although the exposure assessment strategy in both cohorts was the same, the cohorts differed with regards to certain aspects. The difference in the air pollution exposure level between the two cohorts likely contributed considerably to the level of statistical significance (but not the direction) of the observed association. Although most of the individuals in our analyses were included from the Swedish cohort, absolute exposure levels and the exposure contrast in that cohort were low (median 6.65 $\mu g/m^3$, SD 5.8), compared to the Italian cohort (median 94.21 $\mu g/m^3$, SD 43.0). Furthermore NO_X exposure levels observed in the Swedish cohort were the lowest observed among the 36 European regions for which exposure to NO_X was estimated in the ESCAPE study, while exposure levels observed for the Italian cohort (especially the Turin cohort) were among the highest (Cyrys et al., 2012). If we restricted the analyses of the Swedish cohort to subjects exposed to NO_X levels higher than the lowest 5th percentile (7.8 $\mu g/m^3$) of the Italian cohort, we did not observe a more homogeneous picture.

Differences between the cohorts also existed with regards to age, sex, BMI, and smoking behavior. As we controlled for these aspects in our analyses, a large impact on our results is unlikely. Further (not a priori selected) factors available in our dataset including, socio-economic status, dietary patterns, and physical activity, did not significantly

explain variability in markers of inflammation within the two cohorts independently and are therefore unlikely responsible for the observed differences between the cohorts. The cohorts differed with regards to the anticoagulant that was used in the blood samples (EDTA in Sweden, citrate in Italy). Although the use of different anticoagulants results in absolute differences in levels of inflammatory markers (See Table S2.8 for the variance of the concentration of each inflammation marker for the Swedish and Italian cohorts separately), correlations between measurements in split samples simultaneously treated with heparin, citrate, and EDTA have shown to be highly correlated (Hebels et al., 2013; Wong et al., 2008). The difference between the Swedish and the Italian cohort with respect to the anticoagulant that was used, for which we corrected in the statistical models on the combined dataset by adjusting for cohort, is therefore an unlikely explanation for the observed between-cohort differences in our results.
Our analysis among two groups of potentially susceptible populations (individuals with a BMI > 25 or individuals older than age 50) provided some indication (though not significant) for a larger effect among the overweight, but not among the individuals older than age 50. As individuals with conditions associated with chronic inflammation such as older or overweight individuals have been shown to have enhanced susceptibility for air pollution related health effects (Dubowsky et al., 2006; Rückerl et al., 2007), we *a priori* assumed a larger effect size in these subgroups compared to the overall population. For the overweight this was indeed the case. However, among the elderly we observed a smaller effect size compared to the effect size observed in the overall population. Considering the unexpected inverse relationship that we observed between exposure to air pollution and cytokine production, further biological interpretation of these patterns in the effect sizes is not evident.

Conclusion

Our results suggested some indication of an inverse association between long-term NO_X exposure and four systemic inflammatory markers: IL-2, IL-8, IL10, and TNF-α. These results might contribute to a future elucidation of the pathways through which long-term exposure to air pollution induces adverse health effects.

Supplemental material

Table S2.1: Summary of the results of related studies.

Studies	Study Design	Exposures	Biomarkers	Marker(s) included in current study	Direction of the effect
Forbes et al. 2009	Data included adults participating in 3 representative cross-sectional studies of the English population in 1994, 1998, and 2003. Measures of outdoor air pollution estimated for each postcode sector of residence using models incorporating information on pollutantemissions from multiple sources, and atmospheric dispersion and processing, in 3 surveys.	PM_{10} NO_2 SO_2 O_3	Fibrinogen C-reactive protein (CRP)	no	(↓) not significant (↑) not significant
Hoffman et al.2009	Baseline data from the German Heinz Nixdorf Recall Study have been used, a population-based, prospective cohort study of 4,814 participants that started in 2000.	$PM_{2.5}$	Fibrinogen C-reactive protein (CRP)	no	(↑) significant for both
S.Panasevich et al. 2009 (12)	The study group consisted of a population sample of 1028 men and 508 women aged 4570 years from Stockholm.	$NO_2(Traffic-related)$ SO_2 (heating-related)	IL-6 TNF-α CRP (C-reactive protein) Fibrinogen	Yes (IL-6, TNF-α)	(↑) significant for both pollutants. (↑) not significant TNF-α and NO_2 (↓) not significant TNF-α and SO_2 (↓) not significant Fibrinogen with both pollutants
Chuang et al. 2011	Author conducted secondary analyses of data on blood pressure and blood biochemistry markers from the Social Environment and Biomarkers of Aging Study in Taiwan and air pollution data from the Taiwan Environmental Protection Administration in 2000 Environmental Protection Administration in 2000 Environmental Protection Administration in 2000	PM_{10} $PM_{2.5}$ $O3_2$ NO_2 SO_2 CO	IL-6 Neutrophils	Yes (IL-6)	(↑) significant for both markers with PM_{10}, $PM_{2.5}$, NO_2

Note: [a] SD, Standard deviation; [b] BMI, Body mass index; [c] NO_X, Nitrogen oxide

Table S2.2: Effects of long-term exposure to NO_X on markers of inflammation adjusted for potential confounders

	Combined cohort (n=587)				Swedish cohort (n=490)				Italian cohort (n=97)			
Markers	β^a	SE[b]	p-value	q-value [c]	β^a	SE[b]	p-value	q-value[c]	β^a	SE [b]	p-value	q-value [c]
EGF[d]	-0.015	0.136	0.912	0.912	0.044	0.159	0.784	0.863	-0.237	0.296	0.423	0.870
eotaxin	-0.006	0.041	0.883	0.883	0.048	0.049	0.330	0.412	-0.088	0.083	0.286	0.449
FGF[e].2	0.011	0.132	0.934	0.934	0.111	0.163	0.498	0.687	-0.276	0.247	0.263	0.469
fractalkine	0.008	0.135	0.955	0.962	0.102	0.175	0.559	0.804	-0.177	0.165	0.283	0.520
GCSF[f]	0.045	0.085	0.597	0.970	0.141	0.111	0.202	0.445	-0.107	0.102	0.292	0.604
GRO[g]	-0.021	0.039	0.590	0.590	-0.041	0.046	0.377	0.461	-0.022	0.084	0.796	0.969
INF[h]-α	-0.049	0.234	0.834	0.959	-0.209	0.293	0.475	0.821	0.216	0.412	0.599	0.811
IP[i]-10	-0.034	0.050	0.495	0.660	-0.023	0.064	0.724	0.889	-0.059	0.074	0.421	0.772
MCP[j]-1	0.020	0.049	0.680	0.742	0.020	0.064	0.758	0.871	0.003	0.053	0.955	0.956
MCP-3	-0.014	0.147	0.927	0.927	0.122	0.183	0.507	0.727	-0.296	0.252	0.240	0.441
MDC[k]	0.012	0.049	0.805	0.805	0.038	0.058	0.508	0.559	-0.126	0.102	0.215	0.520
MIP[l]-1α	0.056	0.115	0.626	0.751	0.090	0.145	0.534	0.653	-0.131	0.184	0.478	0.679
MIP-1β	-0.003	0.093	0.973	0.973	0.060	0.122	0.624	0.687	-0.131	0.112	0.243	0.535
sCD40L[m]	-0.067	0.065	0.302	0.518	-0.064	0.084	0.446	0.674	-0.129	0.091	0.159	0.386
TGF[n]-α	0.076	0.153	0.618	0.674	-0.013	0.187	0.944	0.944	0.042	0.289	0.884	0.985
VEGF[o]	-0.016	0.142	0.908	0.908	0.084	0.174	0.630	0.694	-0.255	0.261	0.329	0.782
IL[p]-1β	-0.120	0.145	0.405	0.666	0.021	0.171	0.904	0.904	-0.640	0.252	0.011	0.061
IL-2	-0.058	0.118	0.623	0.703	-0.019	0.144	0.896	0.896	-0.421	0.143	0.003	0.012
IL-4	-0.227	0.133	0.088	0.211	-0.241	0.160	0.132	0.291	-0.425	0.196	0.030	0.166
IL-5	-0.080	0.112	0.475	0.712	-0.046	0.139	0.741	0.988	-0.273	0.147	0.064	0.234
IL-6	-0.087	0.114	0.448	0.489	0.041	0.140	0.769	0.851	-0.488	0.192	0.011	0.080
IL-7	0.083	0.115	0.471	0.802	0.077	0.153	0.615	0.794	-0.039	0.088	0.657	0.803
IL-8	-0.179	0.068	0.008	0.020	-0.079	0.076	0.295	0.655	-0.520	0.133	0.000	0.001
IL-10	-0.256	0.116	0.028	0.083	-0.104	0.142	0.467	0.619	-0.690	0.218	0.002	0.017
IL-13	-0.027	0.151	0.856	0.856	0.147	0.183	0.423	0.491	-0.564	0.219	0.010	0.056
INF-γ	-0.152	0.163	0.351	0.468	-0.108	0.202	0.593	0.722	-0.379	0.279	0.174	0.618
GMCSF[q]	0.005	0.117	0.968	0.968	0.090	0.150	0.546	0.649	-0.355	0.135	0.009	0.094
TNF[r]-α	-0.028	0.057	0.620	0.643	0.036	0.073	0.625	0.689	-0.213	0.069	0.002	0.020

Note: [a] Effect estimate per unit changes of the exposure; [b] SE, standard error of β; [c] Q-value, false discovery rate correction for p-value,

[d] epidermal growth factor (EGF), [e] fibroblast growth factor (FGF), [f] granulocyte colony-stimulating factor (GCSF), [g] melanoma growth stimulatory activity/growth-related oncogene (GRO), [h] interferon (INF),

[i] interferon gamma-induced protein 10 (IP10), [j] monocyte chemotactic protein (MCP), [k] macrophage derived chemokine (MDC), [l] macrophage inflammatory protein (MIP), [m] soluble CD40 ligand (sCD40L),

[n] transforming growth factor (TGF), [o] vascular endothelial growth factor (VEGF), [p] Interleukin (IL), [q] granulocytemacrophage colony-stimulating factor (GMCSF), and [r] tumor necrosis factor (TNF).

Table S2.3: Effects of long-term exposure to NO_X on markers of inflammation within 3 sub-groups (overweight, elderly, and never-smokers)

	Overweight(n=323)				Elderly (n=383)				Never-smokers (n=353)			
Markers	β^a	SE[b]	p-value	q-value[c]	β^a	SE [b]	p-value	q-value [c]	β^a	SE [b]	p-value	q-value [c]
EGF[d]	- 0.121	0.162	0.455	0.679	- 0.202	0.162	0.212	0.389	0.170	0.180	0.343	0.491
eotaxin	- 0.054	0.052	0.301	0.414	- 0.031	0.047	0.509	0.522	0.011	0.049	0.829	0.829
FGF[e].2	0.074	0.171	0.664	0.664	- 0.012	0.165	0.941	0.941	- 0.108	0.173	0.531	0.759
fractalkine	0.057	0.176	0.748	0.768	0.017	0.162	0.915	0.972	- 0.198	0.178	0.264	0.440
GCSF[f]	- 0.077	0.099	0.440	0.508	0.040	0.101	0.690	0.908	0.095	0.110	0.385	0.883
GRO[g]	- 0.026	0.050	0.601	0.601	- 0.007	0.047	0.880	0.880	- 0.007	0.051	0.897	0.897
INF[h]-α	- 0.198	0.295	0.502	0.864	0.055	0.259	0.833	0.833	- 0.134	0.308	0.663	0.829
IP[i]-10	- 0.024	0.066	0.718	0.877	- 0.004	0.062	0.952	0.957	- 0.047	0.060	0.435	0.534
MCP[j]-1	0.042	0.062	0.492	0.676	0.034	0.061	0.580	0.580	- 0.002	0.057	0.966	0.966
MCP-3	- 0.014	0.187	0.939	0.939	- 0.098	0.182	0.590	0.852	- 0.101	0.195	0.603	0.862
MDC[k]	- 0.051	0.061	0.398	0.487	- 0.041	0.054	0.449	0.539	0.025	0.065	0.705	0.783
MIP[i]-1α	0.014	0.144	0.922	0.967	- 0.059	0.143	0.678	0.790	0.033	0.149	0.824	0.824
MIP-1β	- 0.034	0.119	0.775	0.812	- 0.028	0.109	0.799	0.879	- 0.063	0.121	0.606	0.757
sCD40L[m]	- 0.086	0.087	0.322	0.586	- 0.054	0.082	0.511	0.702	- 0.046	0.075	0.539	0.674
TGF[n]-α	0.038	0.193	0.843	0.921	- 0.121	0.188	0.521	0.792	- 0.004	0.202	0.984	0.984
VEGF[o]	- 0.078	0.179	0.663	0.880	- 0.126	0.169	0.457	0.718	- 0.123	0.188	0.513	0.732
IL[p]-1β	- 0.183	0.177	0.302	0.831	- 0.039	0.162	0.809	0.890	0.144	0.194	0.458	0.640
IL-2	- 0.073	0.144	0.613	0.843	- 0.080	0.139	0.566	0.778	0.104	0.154	0.502	0.628
IL-4	- 0.319	0.159	0.045	0.123	- 0.192	0.151	0.205	0.563	- 0.065	0.175	0.708	0.941
IL-5	- 0.124	0.135	0.359	0.915	0.014	0.126	0.913	0.986	- 0.015	0.146	0.917	0.962
IL-6	- 0.095	0.131	0.468	0.643	- 0.052	0.124	0.678	0.829	0.095	0.154	0.536	0.640
IL-7	0.055	0.138	0.690	0.817	- 0.042	0.135	0.757	0.926	0.132	0.155	0.397	0.893
IL-8	- 0.254	0.078	0.001	0.004	- 0.039	0.078	0.619	0.847	- 0.113	0.093	0.226	0.282
IL-10	- 0.270	0.131	0.040	0.073	- 0.301	0.135	0.026	0.071	- 0.180	0.160	0.262	0.427
IL-13	- 0.184	0.174	0.292	0.428	- 0.084	0.163	0.607	0.910	0.015	0.207	0.944	0.944
INF-γ	- 0.184	0.199	0.353	0.777	- 0.196	0.175	0.263	0.547	- 0.123	0.227	0.587	0.837
GMCSF[q]	0.105	0.143	0.464	0.843	0.065	0.134	0.628	0.767	0.105	0.156	0.499	0.623
TNF[r]-α	0.032	0.067	0.629	0.766	0.012	0.069	0.866	0.951	0.017	0.068	0.803	0.882

Note: [a] Effect estimate per unit changes of the exposure; [b] SE, standard error of β; [c] Q-value, false discovery rate correction for p-value,

[d] epidermal growth factor (EGF), [e] fibroblast growth factor (FGF), [f] granulocyte colony-stimulating factor (GCSF), [g] melanoma growth stimulatory activity/growth-related oncogene (GRO), [h] interferon (INF), [i] interferon gamma-induced protein 10 (IP10), [j] monocyte chemotactic protein (MCP), [k] macrophage derived chemokine (MDC), [l] macrophage inflammatory protein (MIP), [m] soluble CD40 ligand (sCD40L), [n] transforming growth factor (TGF), [o] vascular endothelial growth factor (VEGF), [p] Interleukin (IL), [q] granulocytemacrophage colony-stimulating factor (GMCSF), and [r] tumor necrosis factor (TNF).

Table S2.4: Effects of long-term exposure to NO_X on markers of inflammation adjusted for potential confounders in overweight sub-population

	Sweden (n=276)				Italy (n=47)			
Markers	β^a	SE^b	p-value	q-value^c	β^a	SE^b	p-value	q-value^c
EGF^d	-0.011	0.209	0.957	0.957	-0.376	0.308	0.222	0.630
eotaxin	0.015	0.068	0.825	0.853	-0.145	0.095	0.124	0.207
FGF^e.2	0.187	0.228	0.411	0.514	-0.150	0.270	0.578	0.775
fractalkine	0.156	0.243	0.520	0.650	-0.070	0.190	0.712	0.791
GCSF^f	-0.012	0.137	0.932	0.932	-0.184	0.101	0.068	0.281
GRO^g	-0.109	0.064	0.090	0.179	0.088	0.094	0.346	0.544
INF^h-α	-0.559	0.388	0.149	0.497	0.478	0.543	0.379	0.912
IP^i-10	-0.015	0.090	0.872	0.927	-0.010	0.077	0.896	0.896
MCP^j-1	0.058	0.085	0.500	0.625	0.041	0.055	0.450	0.562
MCP-3	0.149	0.248	0.548	0.783	-0.159	0.318	0.618	0.997
MDC^k	-0.036	0.079	0.645	0.693	-0.143	0.114	0.209	0.417
MIP^l-1α	0.053	0.193	0.783	0.947	-0.126	0.220	0.567	0.822
MIP-1β	0.017	0.164	0.919	0.929	-0.143	0.116	0.218	0.546
sCD40L^m	-0.054	0.116	0.641	0.682	-0.149	0.144	0.299	0.699
TGF^n-α	0.014	0.255	0.955	0.965	-0.061	0.327	0.851	0.851
VEGF^0	0.167	0.238	0.485	0.774	-0.434	0.301	0.148	0.495
IL^p-1β	0.169	0.219	0.440	0.673	-0.690	0.280	0.014	0.046
IL-2	0.030	0.178	0.864	0.960	-0.478	0.198	0.016	0.040
IL-4	-0.287	0.204	0.159	0.377	-0.415	0.200	0.038	0.191
IL-5	-0.081	0.174	0.643	0.869	-0.220	0.157	0.163	0.271
IL-6	0.073	0.170	0.669	0.744	-0.299	0.181	0.099	0.198
IL-7	0.106	0.190	0.578	0.642	-0.178	0.088	0.042	0.123
IL-8	-0.171	0.091	0.060	0.198	-0.393	0.144	0.006	0.031
IL-10	-0.092	0.170	0.588	0.839	-0.626	0.197	0.001	0.005
IL-13	-0.045	0.220	0.837	0.896	-0.563	0.221	0.011	0.054
INF-γ	-0.016	0.262	0.952	0.952	-0.326	0.343	0.343	0.733
GMCSF^q	0.322	0.192	0.093	0.249	-0.315	0.145	0.030	0.235
TNF^r-α	0.130	0.091	0.152	0.379	-0.153	0.073	0.037	0.122

Note: ^{a} Effect estimate per unit changes of the exposure; ^{b} SE, standard error of β; ^{c} Q-value, false discovery rate correction for p-value, ^{d} epidermal growth factor (EGF), ^{e} fibroblast growth factor (FGF), ^{f} granulocyte colony-stimulating factor (GCSF), ^{g} melanoma growth stimulatory activity/growth-related oncogene (GRO), ^{h} interferon (INF), ^{i} interferon gamma-induced protein 10 (IP10), ^{j} monocyte chemotactic protein (MCP), ^{k} macrophage derived chemokine (MDC), ^{l} macrophage inflammatory protein (MIP), ^{m} soluble CD40 ligand (sCD40L), ^{n} transforming growth factor (TGF), ^{o} vascular endothelial growth factor (VEGF), ^{p} Interleukin (IL), ^{q} granulocytemacrophage colony-stimulating factor (GMCSF), and ^{r} tumor necrosis factor (TNF).

Table S2.5: Effects of long-term exposure to NO_X on markers of inflammation adjusted for potential confounders in elderly sub-population

	Sweden (n=318)				Italy (n=65)			
Markers	β^a	SE [b]	p-value	q-value[c]	β^a	SE[b]	p-value	q-value [c]
EGF[d]	-0.083	0.192	0.666	0.740	-0.542	0.316	0.086	0.287
eotaxin	0.015	0.057	0.797	0.797	-0.098	0.080	0.225	0.321
FGF[e].2	0.067	0.203	0.741	0.823	-0.324	0.291	0.266	0.467
fractalkine	0.089	0.207	0.667	0.880	-0.093	0.194	0.632	0.790
GCSF[f]	0.175	0.129	0.173	0.434	-0.235	0.138	0.087	0.627
GROg	-0.032	0.055	0.562	0.703	-0.077	0.096	0.421	0.702
INF[h]-α	-0.078	0.314	0.805	0.805	0.228	0.502	0.650	0.955
IP[i]-10	0.021	0.079	0.788	0.876	-0.065	0.095	0.493	0.980
MCP[j]-1	0.047	0.080	0.557	0.557	0.004	0.063	0.950	0.950
MCP-3	-0.039	0.225	0.863	0.959	-0.352	0.303	0.245	0.409
MDCk	0.026	0.063	0.679	0.755	-0.229	0.118	0.053	0.178
MIP[l]-1α	0.067	0.179	0.709	0.709	-0.371	0.225	0.099	0.493
MIP-1β	0.050	0.141	0.725	0.873	-0.221	0.122	0.070	0.174
sCD40L[m]	-0.031	0.102	0.762	0.847	-0.222	0.130	0.087	0.291
TGF[n]-α	-0.108	0.227	0.632	0.904	-0.229	0.365	0.529	0.814
VEGF[o]	0.084	0.206	0.683	0.683	-0.643	0.311	0.039	0.195
IL[p]-1β	0.034	0.186	0.855	0.855	-0.579	0.293	0.049	0.243
IL-2	-0.061	0.162	0.708	0.787	-0.485	0.184	0.008	0.041
IL-4	-0.158	0.173	0.362	0.633	-0.488	0.240	0.042	0.212
IL-5	0.005	0.154	0.972	0.972	-0.217	0.152	0.154	0.386
IL-6	0.042	0.147	0.776	0.776	-0.524	0.222	0.018	0.173
IL-7	-0.042	0.175	0.810	0.923	-0.196	0.097	0.043	0.142
IL-8	-0.042	0.087	0.631	0.901	-0.284	0.136	0.037	0.092
IL-10	-0.255	0.167	0.126	0.421	-0.663	0.205	0.001	0.012
IL-13	0.022	0.185	0.904	0.946	-0.708	0.249	0.004	0.022
INF-γ	-0.234	0.211	0.266	0.598	-0.342	0.333	0.305	0.763
GMCSF[q]	0.170	0.166	0.308	0.529	-0.346	0.143	0.015	0.076
TNF[r]-α	0.067	0.089	0.450	0.899	-0.195	0.065	0.003	0.009

Note: [a] Effect estimate per unit changes of the exposure; [b] SE, standard error of β; [c] Q-value, false discovery rate correction for p-value, [d] epidermal growth factor (EGF), [e] fibroblast growth factor (FGF), [f] granulocyte colony-stimulating factor (GCSF), [g] melanoma growth stimulatory activity/growth-related oncogene (GRO), [h] interferon (INF), [i] interferon gamma-induced protein 10 (IP10), [j] monocyte chemotactic protein (MCP), [k] macrophage derived chemokine (MDC), [l] macrophage inflammatory protein (MIP), [m] soluble CD40 ligand (sCD40L), [n] transforming growth factor (TGF), [o] vascular endothelial growth factor (VEGF), [p] Interleukin (IL), [q] granulocytemacrophage colony-stimulating factor (GMCSF), and [r] tumor necrosis factor (TNF).

Table S2.6: Effects of long-term exposure to NO_X on markers of inflammation adjusted for potential confounders in never-smokers sub-population

	Sweden (n=289)				Italy (n=64)			
Markers	β^a	SE^b	p-value	q-value^c	β^a	SE ^{b}	p-value	q-value^c
EGF^d	0.012	0.209	0.953	0.985	0.352	0.400	0.379	0.602
eotaxin	0.032	0.058	0.584	0.738	0.024	0.108	0.827	0.827
FGF^e.2	-0.140	0.215	0.513	0.723	-0.181	0.330	0.585	0.723
fractalkine	-0.206	0.235	0.381	0.490	-0.194	0.196	0.322	0.451
GCSF^f	0.204	0.146	0.161	0.628	-0.058	0.136	0.668	0.751
GRO^g	-0.047	0.061	0.445	0.555	0.017	0.103	0.869	0.869
INFh-α	-0.555	0.385	0.149	0.392	0.301	0.562	0.593	0.777
IP^i-10	-0.063	0.076	0.410	0.566	-0.101	0.097	0.299	0.448
MCP^j-1	-0.022	0.075	0.772	0.869	0.001	0.066	0.994	0.994
MCP-3	-0.071	0.246	0.774	0.948	-0.139	0.345	0.687	0.884
MDC^k	0.011	0.077	0.889	0.899	-0.032	0.146	0.828	0.947
MIP^l-1α	-0.025	0.188	0.895	0.895	-0.043	0.258	0.868	0.868
MIP-1β	-0.127	0.161	0.430	0.909	0.039	0.142	0.784	0.872
sCD40L^m	-0.102	0.099	0.307	0.460	-0.013	0.068	0.851	0.959
TGF^n-α	-0.130	0.246	0.599	0.874	0.133	0.405	0.743	0.995
VEGF^o	-0.266	0.233	0.254	0.570	0.031	0.350	0.928	0.928
IL^p -1β	0.172	0.230	0.454	0.506	-0.525	0.340	0.122	0.275
IL-2	0.103	0.192	0.592	0.665	-0.519	0.174	0.003	0.009
IL-4	-0.220	0.212	0.299	0.609	-0.507	0.271	0.061	0.276
IL-5	-0.035	0.185	0.851	0.921	-0.355	0.195	0.069	0.207
IL-6	0.290	0.187	0.120	0.359	-0.589	0.274	0.032	0.152
IL-7	0.084	0.212	0.691	0.901	-0.038	0.116	0.742	0.954
IL-8	0.006	0.104	0.954	0.954	-0.634	0.182	0.000	0.002
IL-10	-0.056	0.197	0.777	0.783	-0.658	0.305	0.031	0.140
IL-13	0.078	0.261	0.766	0.821	-0.675	0.270	0.012	0.037
INF-γ	-0.270	0.285	0.343	0.446	-0.127	0.381	0.738	0.772
GMCSF^q	0.104	0.202	0.607	0.904	-0.408	0.197	0.039	0.116
TNF^r-α	0.086	0.089	0.334	0.399	-0.217	0.088	0.014	0.045

Note: ^{a} Effect estimate per unit changes of the exposure; ^{b} SE, standard error of β; ^{c} Q-value, false discovery rate correction for p-value, ^{d} epidermal growth factor (EGF), ^{e} fibroblast growth factor (FGF), ^{f} granulocyte colony-stimulating factor (GCSF), ^{g} melanoma growth stimulatory activity/growth-related oncogene (GRO), ^{h} interferon (INF), ^{i} interferon gamma-induced protein 10 (IP10), ^{j} monocyte chemotactic protein (MCP), ^{k} macrophage derived chemokine (MDC), ^{l} macrophage inflammatory protein (MIP), ^{m} soluble CD40 ligand (sCD40L), ^{n} transforming growth factor (TGF), ^{o} vascular endothelial growth factor (VEGF), ^{p} Interleukin (IL), ^{q} granulocytemacrophage colony-stimulating factor (GMCSF), and ^{r} tumor necrosis factor (TNF).

Table S2.7: Effects of long-term exposure to NO_X on markers of inflammation adjusted for potential confounders in future control sub-population

	Combined cohort (n=292)				Swedish cohort (n=245)				Italian cohort (n=47)			
Markers	β^a	SE[b]	p-value	q-value[c]	β^a	SE[b]	p-value	q-value [c]	β^a	SE[b]	p-value	q-value [c]
EGF[d]	-0.240	0.207	0.247	0.411	-0.265	0.234	0.256	0.329	-0.170	0.511	0.739	0.915
eotaxin	0.014	0.064	0.831	0.922	0.048	0.078	0.536	0.603	0.033	0.119	0.779	0.816
FGF[e].2	-0.256	0.197	0.193	0.445	-0.250	0.231	0.278	0.443	-0.374	0.425	0.378	0.798
fractalkine	-0.321	0.206	0.120	0.240	-0.310	0.257	0.227	0.341	-0.319	0.277	0.250	0.390
GCSF[f]	0.048	0.134	0.722	0.874	0.160	0.168	0.341	0.515	-0.171	0.187	0.361	0.732
GRO[g]	-0.125	0.059	0.035	0.058	-0.145	0.068	0.033	0.060	-0.101	0.140	0.471	0.963
INF[h]-α	-0.779	0.354	0.028	0.140	-0.744	0.431	0.084	0.378	-0.805	0.652	0.217	0.625
IP[i]-10	-0.046	0.073	0.530	0.589	-0.038	0.091	0.678	0.763	-0.096	0.126	0.447	0.575
MCP[j]-1	0.003	0.072	0.964	0.964	-0.014	0.091	0.874	0.874	-0.004	0.090	0.964	0.964
MCP-3	-0.270	0.215	0.210	0.451	-0.165	0.259	0.523	0.673	-0.638	0.404	0.114	0.342
MDC[k]	-0.025	0.073	0.731	0.841	-0.055	0.084	0.509	0.763	-0.002	0.160	0.991	0.991
MIP[l]-1α	-0.045	0.175	0.797	0.897	-0.009	0.212	0.966	0.966	-0.175	0.317	0.581	0.837
MIP-1α	-0.114	0.144	0.428	0.853	-0.141	0.180	0.433	0.846	0.054	0.193	0.780	0.780
sCD40L[m]	-0.066	0.098	0.505	0.561	-0.053	0.119	0.654	0.735	-0.210	0.174	0.226	0.509
TGF[n]-α	-0.092	0.229	0.687	0.871	-0.135	0.271	0.618	0.835	-0.103	0.482	0.831	0.907
VEGF[o]	-0.294	0.216	0.174	0.248	-0.358	0.255	0.160	0.230	-0.171	0.464	0.712	0.952
IL[p]-1α	-0.272	0.219	0.214	0.714	-0.571	0.248	0.021	0.096	-0.086	0.386	0.823	0.940
IL-2	-0.332	0.181	0.066	0.133	-0.565	0.212	0.008	0.024	-0.369	0.215	0.086	0.387
IL-4	-0.136	0.211	0.520	0.732	-0.310	0.247	0.208	0.468	-0.247	0.316	0.435	0.739
IL-5	-0.122	0.173	0.481	0.740	-0.398	0.204	0.051	0.230	0.183	0.232	0.430	0.660
IL-6	-0.001	0.165	0.994	0.994	-0.057	0.196	0.770	0.816	-0.145	0.289	0.616	0.792
IL-7	-0.166	0.174	0.340	0.680	-0.346	0.223	0.121	0.544	0.008	0.144	0.958	0.958
IL-8	-0.148	0.102	0.145	0.277	-0.190	0.110	0.084	0.376	-0.318	0.210	0.130	0.292
IL-10	-0.368	0.184	0.046	0.114	-0.487	0.227	0.032	0.097	-0.218	0.287	0.448	0.740
IL-13	-0.178	0.231	0.442	0.689	-0.249	0.268	0.354	0.557	-0.555	0.351	0.114	0.515
INF-γ	-0.260	0.236	0.271	0.677	-0.509	0.278	0.067	0.301	0.148	0.469	0.753	0.978
GMCSF[q]	-0.161	0.186	0.387	0.450	-0.283	0.231	0.220	0.426	-0.268	0.226	0.235	0.675
TNF[r]-α	-0.077	0.086	0.371	0.499	-0.074	0.110	0.498	0.747	-0.097	0.109	0.372	0.837

Note: [a] Effect estimate per unit changes of the exposure; [b] SE, standard error of β; [c] Q-value, false discovery rate correction for p-value,

[d] epidermal growth factor (EGF), [e] fibroblast growth factor (FGF), [f] granulocyte colony-stimulating factor (GCSF), [g] melanoma growth stimulatory activity/growth-related oncogene (GRO), [h] interferon (INF), [i] interferon gamma-induced protein 10 (IP10), [j] monocyte chemotactic protein (MCP), [k] macrophage derived chemokine (MDC), [l] macrophage inflammatory protein (MIP), [m] soluble CD40 ligand (sCD40L), [n] transforming growth factor (TGF), [o] vascular endothelial growth factor (VEGF), [p] Interleukin (IL), [q] granulocytemacrophage colony-stimulating factor (GMCSF), and [r] tumor necrosis factor (TNF).

Table S2.8: Variance of log-transformed concentrations for each inflammatory mediator

Inflammatory Markers	Sweden	Italy
EGF[a]	3.06	5.11
eotaxin	0.34	0.46
FGF[b].2	3.36	3.95
fractalkine	4.20	1.84
GCSF[c]	1.42	0.70
GRO[d]	0.48	0.50
INF[e]-α	11.21	8.74
IP[f]-10	0.50	0.32
MCP[g]-1	0.56	0.17
MCP-3	4.01	4.40
MDC[h]	0.48	0.69
MIP[i]-1	3.79	4.03
MIP-1	1.81	0.93
sCD40L[j]	0.90	0.47
TGF[k]-α	5.70	6.58
VEGF[l]	4.07	4.24
IL[m]-1	5.66	4.13
IL-2	3.68	4.08
IL-4	6.14	2.79
IL-5	3.06	3.35
IL-6	2.84	2.21
IL-7	3.29	0.48
IL-8	0.84	2.61
IL-10	2.73	2.69
IL-13	5.73	4.04
INF-γ	4.80	4.09
GMCSF[n]	3.56	1.44
TNF[o]-α	0.63	0.32

[a] epidermal growth factor (EGF), [b] fibroblast growth factor (FGF), [c] granulocyte colony-stimulating factor (GCSF), [d] melanoma growth stimulatory activity oncogene (GRO), [e] interferon (INF), [f] interferon gamma-induced protein 10 (IP10), [g] monocyte chemotactic protein (MCP), [h] macrophage derived chemokine (MDC), [i] macrophage inflammatory protein (MIP), [j] soluble CD40 ligand (sCD40L), [k] transforming growth factor (TGF), [l] vascular endothelial growth factor (VEGF), [m] Interleukin (IL), [n] granulocytemacrophage colony-stimulating factor (GMCSF), and [o] tumor necrosis factor (TNF).

CHAPTER 3

ASSOCIATIONS BETWEEN GENOME-WIDE GENE EXPRESSION AND AMBIENT NITROGEN OXIDES

This chapter is published as: Mostafavi N, Vlaanderen J, Portengen L, Chadeau-Hyam M, Modig L, Palli D, Bergdahl IA, Brunekreef B, Vineis P, Hebels DGAJ, Kleinjans JCS, Krogh V, Hoek G, Georgiadis P, Kyrtopoulos S, Vermeulen R. Associations between genome-wide gene expression and ambient nitrogen oxides. *Epidemiology*, 2017 May;28(3):320-328. doi: 10.1097/EDE.0000000000000628.

WE hypothesize that biological perturbations due to exposure to ambient air pollution are reflected in gene expression levels in peripheral blood mononuclear cells (PBMC).

We assessed the association between exposure to ambient air pollution and genome-wide gene expression levels in PBMC collected from 550 healthy subjects participating in cohorts from Italy and Sweden. Annual air pollution estimates of nitrogen oxides (NO_X) at time of blood collection (1990 to 2006) were available from the ESCAPE study. In addition to univariate analysis and two variable selection methods to investigate the association between probe-level expression and long-term exposure to NO_X, we applied gene set enrichment analysis to assess overlap between our most perturbed genes and gene sets hypothesized to be related to air pollution and cigarette smoking. Finally, we assessed associations between NO_X and CpG island methylation at the identified genes.

Annual average NO_X exposure in the Italian and Swedish cohorts was 94.2 $\mu g/m^3$, and 6.7 $\mu g/m^3$, respectively. Long-term exposure to NO_X was associated with seven probes (A_23_P252075 (*AHCYL2*), A_24_P406830 (*MTMR2*), A_32_P175313, A_32_P44961 (*LARP1B*), A_32_P156373, A_32_P61298, and A_23_P217280 (*NOX1*)) in the Italian cohort and one probe A_23_P89380 (*IL-8*) in the Swedish (and combined) cohorts. For genes *AHCYL2* and *MTMR2* changes were also seen in the methylome. Genes hypothesised to be downregulated due to cigarette smoking were enriched among the most strongly downregulated genes from our study.

This study provides evidence of subtle changes in gene expression related to exposure to long-term NO_X. On a global level the observed changes in the transcriptome may indicate similarities between air pollution and tobacco induced changes in the transcriptome.

Introduction

Epidemiologic studies have consistently shown associations between long-term exposure to ambient air pollution and incidence and prevalence of chronic diseases such as respiratory and cardiovascular disease (Brunekreef and Holgate, 2002; Hoek et al., 2013). Although the exact mechanisms responsible for these adverse health effects are unclear, several studies have suggested pollutant-induced oxidative stress and systemic inflammation as potential intermediate biological responses to air pollutants (Ghio et al., 2012; Li et al., 2003; Demetriou et al., 2012). It has been hypothesized that (early) systemic effects of long-term exposure to air pollution can be detected by assessing genome-wide gene expression profiles in peripheral blood mononuclear cells (PBMC) (Chuang et al., 2010; Pettit et al., 2012).

Although exposure to air pollutants has been shown to induce changes in gene expression in animal and in vitro experiments (Li et al., 2004; Huang et al., 2011; Araujo et al., 2008), evidence from human studies is scarce (van Leeuwen et al., 2008; Van Leeuwen et al., 2006). Relevant evidence comes from a study by Wittkopp et al. (2016). In this panel study, week-long exposure assessment of ambient air pollutants was combined with the assessment of expression levels of 35 candidate genes from 10 biological pathways relevant to air pollution exposure responses (Wittkopp et al., 2016). Positive associations were observed between traffic-related pollutants elemental carbon, black carbon, primary organic carbon, polycyclic aromatic hydrocarbons (PAHs) in particulate matter (PM), and nitrogen oxides (NO_X) and the *Nrf2* gene (*NFE2L2*), and *Nrf2*-mediated genes, *HMOX1*, *NQO1*, and *SOD2*.

Three short-term (1-2 hours) inhalation studies reported changes in the expression of genes involved in inflammation, tissue growth and host defense, including *IGF-1* signaling, insulin receptor signaling and *NRF2*-mediated oxidative stress response pathway in response to exposure to ultrafine particles (Huang et al., 2010); increased expression levels of genes involved in vascular inflammation and hemostasis (e.g. *IL8RA*, *TNFAIP6*, and *VEGF*) in response to 2 hours exposure to diesel exhaust (Pope III et al., 2011); and genes involved in oxidative stress, protein degradation, and coagulation (e.g. *PLAU*, *F2R*, *CBL*, *UBR1*) in response to 1 hour of exposure to diesel exhaust (Pettit et al., 2012). To date studies are largely inconclusive and have not resulted in clear gene expression profiles associated with air pollution. This may in part be explained by relatively small study sizes and modest exposure contrasts.

Considering the similarities between tobacco smoke (a combustion product) and air pollution (mostly combustion products), studies of smoking and gene expression might also provide some insight into potential gene expression targets of air pollution (Pope III et al., 2011). An example of such a study is Beineke et al. (2012) in which 4214 genes from biological pathways known to be affected by both smoking and air pollution (i.e., apoptosis and cellular death, immune system development, leukocyte activation, haematopoiesis, stress

response, and alterations in platelet activity) were correlated with self-reported smoking status.
In this hypothesis generating study, we assessed the association between annual average estimate of NO_X concentrations and genome-wide changes in gene expression in PBMC in a large population using state-of-the-art exposure assessment methods (Beelen et al., 2013). We assessed the overlap between genes associated to NO_X and gene sets hypothesized to be related to air pollution exposure and cigarette smoking. Moreover, for genes which expression levels were associated to NO_X, we assessed the role of DNA methylation in the regulation of the gene expression at potentially cis-acting CpG sites. In addition, we assessed the interaction between the top-ranked probes and four inflammatory markers (IL-2, IL-8, IL-10, TNF-α) which we previously observed to be associated with NO_X in the same study population (Mostafavi et al., 2015).

Material and methods

We combined data from two existing projects: gene expression profiles from the "Genomics Biomarkers of Environmental Health" (EnviroGenoMarkers) (Chadeau-Hyam et al., 2014; Hebels et al., 2013) and long-term average NO_X concentrations at residential address from the European Study of Cohorts for Air Pollution Effects (ESCAPE) (Beelen et al., 2013).
Study design and data collection procedures have been previously described in detail (Mostafavi et al., 2015).

Study population

The EnviroGenoMarkers study was based on analyses of PBMCs of participants from two prospective cohorts: the Italian contribution to the European Prospective Investigation into Cancer and Nutrition study (EPIC-Italy, 95 individuals (22 men, and 73 women)) and the Northern Sweden Health and Disease Study (NSHDS, 455 individuals (202 men, and 253 women)). In both cohorts, blood samples were prospectively collected from healthy subjects at enrolment (around 1990-2006) and cohort members were asked to complete a standardized questionnaire focusing on dietary and life style.
Our study population, a subset of the EnviroGenoMarkers data, was collected in two phases and comprised in total 221 Non-Hodgkins lymphoma cases and 58 breast cancer cases, identified through local Cancer Registries (loss to follow-up $< 2\%$), and the same number of controls matched on sex, age, center and date of blood collection were included (Mostafavi et al., 2015). Cases were diagnosed on average 6 years (range 2 to 16 years) after recruitment/blood collection.

Ethics statement

This study was approved by the committees on research ethics in Umea and Florence at the relevant institutions. All participants provided written consent at recruitment (EPIC-Italy 1993-1998; NSHDS 19902006).

Exposure assessment

Annual modeled outdoor concentrations of NO_X at the study participant's home-address were available from the ESCAPE project (Beelen et al., 2013; Cyrys et al., 2012). Exposure to particulate matter ($PM_{2.5}$, $PM_{2.5}$ absorbance and PM_{10}) was only available for a subset of our study population (13 subjects). We therefore restricted our analyses to NO_X (Mostafavi et al., 2015). We natural log transformed the distribution of the NO_X concentration to limit the influence of high concentrations and normalize the distribution.

Gene expression assessment

Total RNA was extracted from PBMC samples stored within 2 hours of collection at $-80°C$. RNA from each sample was used to generate cDNA for array hybridization. The cDNA was then labeled with cyanine 3. The labeled cDNA was hybridized to Agilent whole human genome (4×44 K) arrays, containing 43376 probes representing 29846 genes. Subsequently, the hybridized slides were washed and scanned by using an Agilent Technologies G2565CA DNA Microarray scanner. Measurements for both phases were performed at Maastricht University. Technical performance and quality of the microarrays has been described in detail previously (Chadeau-Hyam et al., 2014; Hebels et al., 2013). In short, microarray scan images were visually evaluated before and after within- and between-array normalization (LOESS and A-quantile, respectively). Good probes were identified based on the number of pixels, mean/median intensity ratio, saturation, or foreground/background intensity ratio. A total of 29662 probes, representing 15216 genes, were selected based on these criteria. We imputed missing values in Gene Pattern (version 3.1) using the k nearest neighbors approach (k = 15, Euclidian metric). When known, annotation of probes are provided in Italic within parentheses.

Data analysis

We performed univariate analyses to identify transcript concentration levels associated with long-term average exposure to NO_X. We complemented univariate analysis with two additional variable selection approaches (Elastic-Net regression and the Graphical Unit Evolutionary Stochastic Search Algorithm (GUESS)) that are capable of capturing the correlation among genes (Chadeau-Hyam et al., 2013). We call all probes that were identified in any of the statistical approaches "noteworthy probes".

Table 3.1: Characteristics of the study population.

Characteristics	Swedish cohort (n=455)	Italian cohort (n=95)
Sex (N (%))		
Female	253 (56%)	73 (77%)
Male	202 (44%)	22 (23%)
Smoking Status (N (%))		
Current- smoker	97 (21%)	8 (8%)
Former- smoker	92 (20%)	24 (25%)
Never-smoker	266 (59%)	63 (66%)
Age (years) (N (%))		
<40	28 (6%)	2 (2%)
40-50	134 (29%)	29 (31%)
>50	293 (65%)	64(67%)
Future disease status (N (%))[a]		
Breast cancer	46 (10%)	12 (13%)
Lymphoma	183 (40%)	38 (40%)
Control	226 (50%)	45 (47%)
BMI ($\mu kg/m^2$) (Mean SD[b])	26.13.8	25.73.7
NO_X ($\mu g/m^3$) (Median (SD[b]))	6.7 (5.8)	94.2 (42.5)

Note: [a]SD, Standard deviation. [b]Non-Hodgkin's lymphoma and breast cancer cases, identified through local Cancer Registries (loss to follow-up $< 2\%$.) occurred on average 6 years (range 1 to 17 years) after recruitment/blood collection.

Within the ESCAPE project (Beelen et al., 2013; Cyrys et al., 2012) the Swedish cohort was among the cohorts with the lowest levels of air pollution, while the Italian cohort was among the highest. We therefore stratified all statistical analyses by cohort (Table 3.1 ; Figure 3.1). As there was some overlap in the exposure distributions of the two cohorts, we also conducted analyses on the combined cohorts, while adjusting for country.
All statistical analyses were performed using R version 3.0.2 (packages: lme4 (De Boeck et al., 2011), glmnet (Friedman et al., 2010), c060 (Sill et al., 2014), and R2Guess (Liquet et al., 2016)).

Univariate mixed-effects model

We conducted linear mixed-effects modeling to investigate the association between probe-level expression and long-term exposure to NO_X. To account for potential technical noise (nuisance variation), we incorporated the dates of three main steps of sample processing (i.e. RNA isolation, hybridization, and dye labeling) as random-effects in the models. Exposure to NO_X and a priori selected potential confounding factors were included as fixed-effects in the models. These confounding factors were body mass index (BMI) (kg/m^2), age (years) in three categories: (30-40, 41-50, 51-60), sex, smoking status (never-

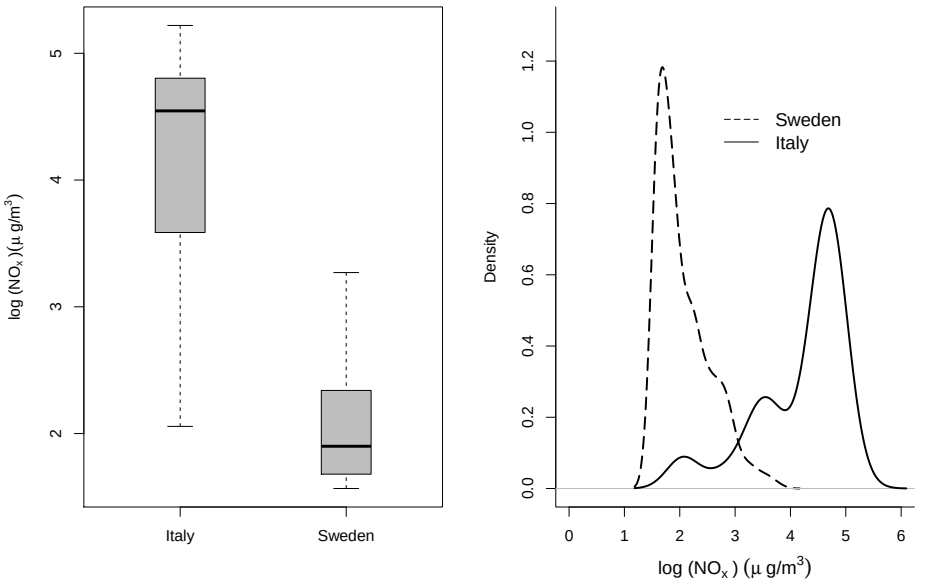

Figure 3.1: Box plot (left) and density plot (right) of log(NO_X) ($\mu g/m^3$) concentration for Swedish and Italian participants. Log(NO_X) consentrations are shown on the Y axis of the box plots and on the X axis of the density plot.

smoker, former-smoker, current-smoker), phase (1 or 2), future disease status (lymphoma case, breast cancer case, control), and sample storage time (years), consistent with previous analyses of the EnviroGenoMarkers data (Chadeau-Hyam et al., 2014; Kelly et al., 2013).

To assess how sensitive our findings were to variations in the confounder model, we conducted a set of additional analyses. We ran a minimally adjusted model (only age and sex included as covariates), a model in which smoking and BMI were excluded from the primary set of covariates, and a model in which we added education level (primary, technical, secondary, and university), and physical activity (moderately inactive, moderately active, and active) as covariates. In a further sensitivity analysis, we assessed the impact on our findings of adjusting our regression models for estimated cell-type composition (Georgiadis et al., 2016) (available from an epigenome-wide analysis on a subset of the subjects).

We assessed the output from the univariate analysis using two approaches. First, we followed an agnostic approach using the Benjamini-Hochberg false discovery rate (BH-FDR) correction (Benjamini and Hochberg, 1995) to control for false positives. A BH-FDR < 0.2 was used to classify probes as noteworthy.

Second, we followed a candidate gene approach, including 35 candidate genes from

10 biological pathways relevant to air pollution exposure responses (coagulation, *Klf2*-mediated immune response, *NF-B* signaling, acute phase response, *Nrf2*-mediated oxidative stress response, endoplasmic reticulum stress (*UPR*), glutathione metabolism, phase I and phase II metabolism, endogenous reactive oxygen species (*ROS*) production, and cytokine signaling) (Wittkopp et al., 2016) (biological candidate genes) augmented with genes that were associated to air pollution in epidemiological studies (OMICS and gene-environment interaction) published since 2006 (empirical candidate genes) (Table S3.1). Empirical candidate genes were selected if an association in the same direction was observed in at least two previous studies. We assessed the overlap (strength and direction of the association) between the genes identified from the literature and results from our univariate analysis. In these analyses we used a p-value of 0.05 to classify probes as noteworthy.

Variable selection methods

Elastic-Net is a form of penalized multiple regression in which parameter estimates are achieved by using a combination of Ridge and Lasso penalties (Zou and Hastie, 2005). To control the number of falsely selected predictors by Elastic-Net, we applied the stability selection technique proposed by Meinshausen et al., (2010) (Meinshausen and Bühlmann, 2010). We accounted for multiple testing by setting family-wise error rate (FWER) to 0.05. As a sensitivity analysis, we also set the FWER at 0.2, relaxing the type one error. We set the threshold of selection probability (probability of selecting a predictor by algorithm; π) to 0.6.

GUESS is a Bayesian variable selection approach that uses an advanced stochastic search Markov Chain Monte Carlo (MCMC) algorithm (Bottolo et al., 2013). GUESS fits a range of models containing varying combinations of probes (predictors) to the data. Noteworthy probes were selected based on the Marginal Posterior Probability of Inclusion (MPPI), which provides a model-averaged measure of importance of each probe with respect to the models that were fit to the data. We ran 5 different chains in GUESS for 90k iterations and discarded the first 30k iterations as burn-in. Expected and standard deviation of the model size were set to 3 and 5, respectively. We used a MPPI of 10% to call a probe noteworthy (technical details in Supplemental material (Appendix 1)).

DNA methylation at relevant loci

We assessed the association between long-term exposure to NO_X and degree of methylation for all CpG islands (n=74) that were in cis with the noteworthy genes identified in the gene expression analysis. Methylation data generated using the Infinium HumanMethylation450 BeadChip (450K) was available for a subset (466 out of 550) subjects (Georgiadis

et al., 2016). We used the same univariate mixed-effects models as described above for our main gene expression analyses.

Correlation between noteworthy probes and inflammatory plasma markers

To explore whether our noteworthy transcripts and markers of inflammation in plasma (IL-2, IL-8, IL-10, TNF-α) which we previously observed to be associated to air pollution in this cohort (Mostafavi et al., 2015) potentially operate in similar biological pathways, we assessed the interrelationships (Pearson correlation) between identified transcripts and inflammatory markers.

Gene set enrichment analysis (GSEA)

We used GSEA (Subramanian et al., 2005) to assess whether the gene expression pattern associated with NO_X in our data has similarities to patterns associated with cigarette smoking or air pollution responses. We assessed whether the distribution of overlap between two sets of genes - associated with either "exposure to cigarette smoke" or "biological responses due to exposure to air pollution" and all genes included in our study was random, or whether this overlap primarily occurred among our most up- or downregulated genes (i.e. enrichment). The first gene set comprised 4214 genes whose expression in peripheral blood was associated with smoking status in a study by Beineke et al. (2012) (smoking set). The second gene set comprised 35 genes from 10 biological pathways relevant to air pollution responses (Wittkopp et al., 2016) (air pollution set). A similar approach using GSEA, was described by Wang et al. Wang et al. (2015) demonstrating enrichment of cigarette smoke (Wang et al., 2015) related genes among genes affected by indoor air pollution (technical details in Supplemental material (Appendix 2)).
To enhance interpretability, we conducted GSEA separately for upregulated genes (all genes with positive t-statistic in our univariate analysis) and downregulated genes (all genes with negative t-statistic). We compared our upregulated genes with the upregulated genes in the smoking and air pollution sets and compared our downregulated genes with the downregulated genes in the smoking set (no downregulated genes were included in the air pollution set). A p-value <0.05 was used as statistical cut-off for enrichment.

Results

Table 3.1 summarizes the baseline characteristics of the study participants. The Swedish cohort has a lower proportion of women than the Italian cohort (56% versus 77%), a higher proportion of current smokers (21% versus 8%), and a lower proportion of never smokers (59% versus 66%).

Table 3.2: Selected associations between long-term exposure to NO_X and transcript levels based on agnostic approach (q-value <0.2) in the Italian, Swedish and combined population.

		Italian cohort		Swedish cohort		Combined cohort		
Agilent ID	Gene name	β^a(SE)	q-value^b	β^a(SE)	q-value^b	β^a(SE)	q-value^b	interaction^c
A_23_P252075	*AHCYL2*[d]	0.23 (0.05)	0.12	0.01 (0.04)	0.99	0.10 (0.03)	0.99	$3*10^{-4}$
A_24_P406830	*MTMR2*[e]	-0.20 (0.05)	0.12	0.02 (0.03)	0.99	-0.05 (0.03)	0.99	$2*10^{-4}$
A_32_P156373	Unknown	0.31 (0.07)	0.17	-0.06 (0.07)	0.99	0.06 (0.06)	0.99	$3*10^{-4}$
A_32_P175313	Unknown	0.30 (0.07)	0.17	-0.05 (0.06)	0.99	0.07 (0.05)	0.99	$1*10^{-3}$
A_32_P44961	*LARP1B*[f]	0.33 (0.07)	0.02	0.04 (0.06)	0.99	0.13 (0.05)	0.99	$6*10^{-1}$
A_32_P61298	Unknown	0.38 (0.09)	0.17	0.01 (0.09)	0.99	0.07 (0.07)	0.99	$8*10^{-1}$

Note: [a] Effect estimate per unit changes of the exposure., [b] Q-value, false discovery rate correction for p-value,
[c] P-value, for the interaction between country and NOx in the combined population, [d] Adenosylhomocysteinase-like2,
[e] Myotubularin related protein 2, [f] La ribonucleoprotein domain family member 1B.

We observed a considerable difference in the distribution of NO_X concentrations between the two countries (Table 3.1 and Figure 3.1). The median (5^{th} percentile, 95^{th} percentile) concentration of NO_X estimated for the Italian cohort (94.2 $\mu g/m^3$ (7.8, 124.6)) was considerably higher than the median concentration estimated for Sweden (6.7 $\mu g/m^3$ (4.8, 19.5)) (p-value < 0.001).

Univariate mixed-effects model

Agnostic approach

Following an agnostic approach, we identified six noteworthy probes that were associated (BH-FDR <0.2) to long-term average exposure to NO_X in the Italian cohort (Table 3.2). These probes are: A_23_P252075 (*AHCYL2*) (q-value 0.12), A_24_P406830 (*MTMR2*) (q-value 0.12), A_32_P175313 (q-value 0.17), A_32_P44961 (*LARP1B*) (q-value 0.018), A_32_P156373 (q-value 0.17), and A_32_P61298 (q-value 0.17). In Table 3.2, we show the results from the Swedish and combined cohorts for the top-ranked probes (based on the q-value) in the Italian cohort. Perturbations of the noteworthy probes in Italy were all in the same direction in the combined cohort and for three probes (A_23_P252075 (*AHCYL2*), A_32_P44961 (*LARP1B*), A_32_P61298) in the Swedish cohort. However, these differences did not reach the threshold (BH-FDR<0.2) in either the Swedish or the combined cohort.

To formally explore heterogeneity between cohorts, we conducted analyses in the combined population, while including an interaction term between cohort and NO_X. We observed an interaction for four out of six noteworthy probes that were associated to NO_X in the Italian cohort, but not in the Swedish cohort (Table 3.2) (complete results (by cohort and combined) are available in Supplemental material (Appendix 3).

Sensitivity analyses

To assess how sensitive our findings were to variations in the confounder model, we conducted a set of additional analyses. Applying a model that was only adjusted for sex and age, resulted in 8 additional probes being associated (BH-FDR < 0.2) with NO_X in the Italian cohort. One of the original findings, probe A_32_P61298, was not retained in these analyses. Applying a model in which smoking and BMI were excluded as covariates, four of the six probes identified in our primary univariate analyses were retained (all except A_32_P175313 and A_32_P61298) and four additional probes were associated (BH-FDR $<$ 0.2) with NO_X in the Italian cohort. Applying a model in which we added education and physical activity as covariates, we observed three probes (A_32_P44961 (*LARP1B*), A_32_P175313, A_23_P252075 (*AHCYL2*)) to be associated (BH-FDR < 0.2) to NO_X in the Italian cohort. All three probes were included in our primary model. Applying a model in which we corrected for cell-type composition resulted in 10 additional probes being associated (BH-FDR < 0.2) to NO_X in the Italian cohort. One of the original findings, probe A_32_P175313 (q-value = 0.226), no longer met the cut-off (BH-FDR$<$0.2) for noteworthiness after correction for cell-type composition.

Candidate gene approach

We included 36 probes in our candidate gene approach. Eight probes were selected on empirical grounds and 30 probes (corresponding to 26 genes that are overlapping between our study and study by Wittkopp Wittkopp et al. (2016)) were selected based on biological motivation. *IL-6* and *HMOX1* were included in both lists of candidate genes. Studies from which these genes were selected are listed in Table S3.1. We present parameter estimates and associated p-values from univariate mixed-effects regression in Table 3.3. We observed a positive association of NO_X with *NOX1* in the Italian cohort and with *IL-8* in the Swedish (and combined) cohort. The direction of these effects was in agreement with what has been reported in the literature. One gene was negatively associated with exposure to NO_X in our analysis (*SELP* in the combined cohort), but the direction of this effect was not in agreement with what has been reported in the literature.

Variable selection method

Application of Elastic-Net with stability selection yielded one probe (A_32_P44961 (*LARP1B*)) that was associated (FWER$<$0.05) with long-term average exposure to NO_X in the Italian cohort. By increasing the FWER cut-off to 20%, Elastic-Net selected three more probes (A_32_P156373, A_32_P175313, and A_23_P252075 (*AHCYL2*)). All probes were also identified using univariate analysis. We observed no evidence for an association between long-term exposure to NO_X and gene expression in the combined and Swedish cohort for

Table 3.3: Parameter estimate and standard error from univariate regression for candidate genes (both emperical and biological) previously associated with air pollution in the epidemiological literature.

		Italian cohort			Swedish cohort			Combined cohort		
Agilent ID	Gene Name	β^a	SE^b	p-value	β^a	SE^b	p-value	β^a	SE^b	p-value
Empirical genes										
A_23_P502464	*NOS2A*[c]	- 0.03	0.08	0.75	0.03	0.06	0.58	0.02	0.05	0.76
A_32_P50123	*SRGAP2*	-0.16	0.08	0.05	0.01	0.08	0.93	- 0.06	0.06	0.34
A_24_P357869		- 0.01	0.07	0.89	- 0.08	0.05	0.13	- 0.06	0.04	0.12
A_23_P200829		- 0.002	0.06	0.98	0.04	0.04	0.36	0.02	0.03	0.53
A_23_P115407	*GSTM1*	- 0.06	0.08	0.41	- 0.05	0.07	0.42	- 0.03	0.04	0.48
A_23_P202658	*GSTP1*	- 0.084	0.08	0.32	0.04	0.06	0.53	- 0.01	0.05	0.76
Empirical and Biological										
A_23_P120883	*HMOX1*	0.07	0.08	0.40	0.07	0.06	0.19	0.06	0.05	0.18
A_23_P71037	*IL-6*	0.004	0.14	0.98	0.11	0.10	0.26	0.05	0.08	0.56
Biological genes										
A_23_P103996	*GCLM*	0.06	0.06	0.31	- 0.04	0.04	0.32	- 0.01	0.03	0.86
A_32_P177953		0.06	0.06	0.31	- 0.04	0.05	0.47	- 0.003	0.04	0.94
A_23_P105138	*CAT*	0.01	0.06	0.82	0.01	0.05	0.83	0.01	0.04	0.77
A_23_P119196	*KLF2*	0.04	0.08	0.58	0.04	0.07	0.61	0.05	0.05	0.40
A_24_P151305		- 0.02	0.05	0.74	- 0.04	0.05	0.51	- 0.01	0.04	0.81
A_23_P120933	*ATF4*	-0.10	0.07	0.18	- 0.05	0.05	0.26	- 0.07	0.04	0.07
A_23_P120941		- 0.01	0.04	0.75	- 0.07	0.04	0.05	- 0.05	0.03	0.07
A_23_P137697	*SELP*	-0.15	0.11	0.19	-0.12	0.08	0.15	-0.13	0.07	0.04
A_23_P154840	*SOD1*	- 0.01	0.06	0.86	0.02	0.04	0.59	0.02	0.03	0.60
A_23_P163402	*CYP1A1*	0.10	0.08	0.23	0.08	0.07	0.24	0.09	0.05	0.09
A_23_P202658	*GSTP1*	- 0.08	0.08	0.32	0.04	0.06	0.53	- 0.01	0.05	0.76
A_23_P204581	*TXNRD1*	- 0.01	0.06	0.92	- 0.09	0.11	0.40	- 0.08	0.08	0.33
A_23_P209625	*CYP1B1*	0.13	0.11	0.22	0.08	0.07	0.29	0.07	0.06	0.24
A_23_P215566	*AHR*	- 0.01	0.11	0.96	0.03	0.06	0.68	- 0.002	0.05	0.97
A_23_P217280	*NOX1*	0.26	0.13	0.04	0.05	0.11	0.68	0.10	0.09	0.23
A_23_P250671	*GPX1*	- 0.01	0.05	0.87	0.02	0.04	0.65	0.001	0.03	0.99
A_23_P352879	*GCLC*	0.11	0.07	0.15	0.01	0.05	0.88	0.04	0.04	0.24
A_23_P5761	*NFE2L2*	- 0.01	0.06	0.85	- 0.05	0.04	0.25	- 0.04	0.04	0.28
A_23_P62907	*ATF6*	0.09	0.07	0.17	0.01	0.05	0.86	0.03	0.04	0.42
A_23_P7144	*CXCL1*	-0.18	0.13	0.18	-0.13	0.13	0.32	-0.16	0.10	0.12
A_23_P79518	*IL1B*	- 0.06	0.18	0.75	-0.13	0.14	0.38	- 0.05	0.12	0.69
A_23_P89380	*IL8*	0.11	0.11	0.29	0.25	0.09	0.01	0.23	0.07	0.001
A_23_P89431	*CCL2*	0.13	0.11	0.23	0.01	0.09	0.93	0.05	0.07	0.47
A_24_P379413	*IL6R*	- 0.02	0.08	0.79	- 0.03	0.06	0.65	- 0.02	0.05	0.60
A_24_P77008	*PTGS2*	-0.24	0.18	0.18	-0.13	0.16	0.41	-0.17	0.12	0.16
A_24_P935819	*SOD2*	-0.21	0.12	0.09	-0.11	0.13	0.40	-0.13	0.10	0.17
A_24_P936444	*NFE2L2*	0.05	0.06	0.33	- 0.08	0.04	0.08	- 0.02	0.04	0.49
A_32_P13728	*HSPA8*	- 0.02	0.09	0.84	0.046	0.05	0.38	0.03	0.04	0.55

[a] Effect estimate per unit changes of the exposure. [b] Standard error of the effect estimates.
[c] Gene abbreviations are listed in Supplementary Table S3.1.

either FWER cut-off values.
GUESS did not identify any probe that exceeded our predefined MPPI cut-off level of 10%.

DNA methylation at relevant loci

We assessed the association between long-term exposure to NO_X and degree of methylation for 74 CpG islands in cis with our five noteworthy genes. Methylation data was available for a subset (466 out of 550) subjects. Results from this analysis are presented in Table S3.2. Methylation of two CpG islands (hypomethylation cg03793937 upstream of *MTMR2*; q-value = 0.14, and hypermethylation cg06988775 downstream of *AHCYL2*; q-value = 0.14) was associated (BH-FDR<0.2) with long-term exposure to NO_X in the Italian

Table 3.4: Correlation between expression levels of notheworthy probes and concentrations of inflammatory markers previously reported to be associated to NO_X in the same study population.

		Italian cohort				Swedish cohort			
Noteworthy probes	Gene Name	IL.2	IL.8	IL.10	TNF-α	IL.2	IL.8	IL.10	TNF-α
Empirical candidate genes									
A_23_P252075	*AHCYL2*[a]	-0.18	-0.05	-0.04	-0.04	0.007	0.018	0.057	-0.004
A_24_P406830	*MTMR2*	0.40	0.03	0.00	0.18	-0.082	0.030	-0.044	-0.067
A_32_P156373	Unknown	-0.10	0.09	0.09	-0.01	0.055	-0.015	0.063	-0.007
A_32_P175313	Unknown	0.28	0.16	0.15	0.20	0.001	0.071	0.041	-0.021
A_32_P44961	*LARP1B*	-0.05	0.07	0.06	-0.08	0.002	-0.004	-0.010	-0.067
A_32_P61298	Unknown	0.00	-0.06	-0.13	-0.13	-0.017	0.000	-0.007	-0.068
A_23_P217280	*NOX1*	0.053	-0.12	-0.18	-0.11	-0.018	0.058	0.014	-0.019
A_23_P89380	*IL-8*	0.070	0.126	0.113	0.073	0.015	0.093	0.013	0.049

[a] Gene abbreviations are listed in Supplementary Table S3.1

cohort.

Correlation between noteworthy probes and inflammatory markers

The Pearson correlation coefficients between concentration of four inflammatory markers from our previous study (Mostafavi et al., 2015) (i.e. IL-2, IL-8, IL-10, and TNF-α) and expression levels of the eight noteworthy probes from the agnostic and candidate gene approaches are presented in Table 3.4 for Swedish and Italian cohorts, separately.
The overall median of correlation coefficients and their interquartile range for Italian and Swedish cohorts are ρ_p = -0.04 (-0.05, 0.11) and ρ_p = -0.01 (-0.02, 0.03), respectively. The highest correlation was observed between IL-2 and two of the noteworthy probes A_24_P406830 (*MTMR2*) (ρ_p = 0.4) and A_32_P175313 (ρ_p =0.28) in the Italian cohort.

Gene Set Enrichment Analysis (GSEA)

We conducted GSEA using the results for the Italian cohort. When we compared our results with the set of genes negatively associated to cigarette smoking status, we observed enrichment (p-value < 0.01) of the overlapping genes among the genes that were most strongly downregulated due to exposure to air pollution in our analysis (Figure S3.1(A)). We did not observe enrichment when we compared our upregulated genes to the upregulated smoking genes (Figure S3.1(B)) or the air pollution gene set (Figure S3.2).

Discussion

Our study provides some evidence that subtle changes in gene expression are associated with long-term exposure to air pollution as measured by NO_X in a cohort of adult individuals. We identified seven noteworthy probes in the analysis of the Italian cohort,

and one noteworthy probe in the analysis of the Swedish (and the combined Italian-Swedish) cohort. Of these A_23_P252075 (*AHCYL2*) and A_32_P44961 (*LARP1B*) achieved BH_FDR<0.2 in all sensitivity analyses. Gene set enrichment analyses indicated that our down-regulated genes overlapped to a certain extent with genes for which smoking-induced gene expression differences were previously published.

All noteworthy probes we identified in our agnostic analysis are novel findings. Targeted analyses provided support for a potential association with *NOX1* and *IL-8* and air pollution, whereas for some genes (e.g., *Nrf2*-mediated genes *HMOX1*, *NQO1*, and *SOD2* (Wittkopp et al., 2016)), our point estimates suggested associations in the same direction as previously reported, but these associations were imprecise.

To date, few reports of associations between air pollution and gene expression have been replicated. We attribute the lack of replication to the small study populations that have been used, the relatively low exposure levels that individuals in these cohorts experienced, and the lack of adjustment for multiple testing in most studies (Huang et al., 2011; Pettit et al., 2012; Wittkopp et al., 2016; Pettit et al., 2012). Although our study population is still modest in size, the 95 individuals in the Italian cohort were among the highest exposed within Europe (Cyrys et al., 2012).

In studies such as ours, the risk of false-positive findings is high due to the large number of tests conducted compared with the number of observations available. For the univariate analysis, we reduced this risk by controlling the false discovery rate and by using evidence from the literature as an informal prior in our analysis. A limitation of univariate analysis is that it cannot take the correlation structure among genes into account, further increasing the risk of false-positive findings (Chadeau-Hyam et al., 2013). We therefore applied two additional approaches (Elastic-Net and GUESS) that are capable of capturing the correlation among genes. Application of these approaches reduces the risk of false-positive findings, but at the cost of reduced sensitivity (Agier et al., 2015). Correspondingly, in the present study, agnostic univariate analysis identified 6 noteworthy genes that were associated with NO_X, Elastic-Net regression (FWER 0.2) selected four of these whereas GUESS did not select any. We view the three approaches as complementary, but as the primary goal of our current analysis was to discover potential new gene expression targets of ambient air pollution, we preferred high sensitivity over a lower risk of false-positive findings.

Heterogeneity in results across cohorts

As hypothesized, we observed stronger signals in the Italian cohort than in the Swedish cohort. This is likely due to the different level of exposure in the two cohorts. Although the exposure assessment strategy in both cohorts was the same, absolute exposure levels and the exposure contrast in the Swedish cohort were low (median 6.65 $\mu g/m^3$, SD 5.8),

compared with the Italian cohort (median 94.21 $\mu g/m^3$, SD 43.0).

Biological role of noteworthy genes

Using the Gene Expression Omnibus (Barrett et al., 2013), we identified gene annotations for five of our noteworthy probes (*AHCYL2*, *MTMR2*, *LARP1B*, *IL-8*, and *NOX1*). Functional analysis of these genes using the NIH-DAVID bioinformatics resources (Huang et al., 2009) yielded no evidence for functional enrichment in any biological pathway likely due to the limited number of probes found in this study. Biological role of noteworthy genes based on literature review are presented in Table S3.3.

Gene Set Enrichment Analysis

Using our univariate and variable selection approaches, we focused on a small set of genes that showed the largest perturbation. However, following this approach, we might have missed signals that did not meet our threshold for noteworthiness because the perturbation in gene expression was modest relative to the noise inherent to the microarray technology (Subramanian et al., 2005). Gene set enrichment analysis overcomes this limitation by using information (ranking according to the strength of the association with exposure to air pollution) from multiple genes rather than assessing the genes one by one (Subramanian et al., 2005). The observation that there was enrichment of smoking-associated genes among the genes that were negatively associated to exposure to NO_X in our study provides some indication that the perturbation of the transcriptome by exposure to air pollution we observed was a true finding, rather than a false positive. Furthermore, this observation points towards a shared biological pathway of the effects of cigarette smoke and air pollution on the transcriptome which is of interest due to overlap between health outcomes that have been related to tobacco smoking and air pollution (Wang et al., 2015).

DNA methylation at relevant loci

For two genes (*MTMR2* and *AHCYL2*), we observed an effect of long-term exposure to NO_X on expression level as well as on methylation status of two CpG islands in cis with these genes. The effects of long-term exposure to NO_X on methylation of the two genes were in the same direction as what we observed for gene expression (downregulation for *MTMR2* and upregulation for *AHCYL2*), which does not confirm to the often observed inverse correlation between methylation in promoter regions and gene expression due to a gene silencing effect of methylation. In addition, as we tested 74 CPG islands, there is a likelihood for false positives and we are therefore cautious in interpreting these results as a cross-OMICS signal of air pollution.

Correlation between noteworthy probes and inflammatory markers

We found positive and relatively high correlation between immune marker IL-2 and two noteworthy probes (A_24_P406830 (*MTMR2*), A_32_P175313) in the Italian cohort but not in the Swedish cohort. We did not identify further information regarding a potential shared pathway between the genes and the immune marker. The fact that we observed a correlation in the Italian cohort, but not in the Swedish cohort does provide some indication for a potential role of NO_X in inducing this correlation in the Italian cohort by affecting both probes; A_24_P406830 (*MTMR2*), A_32_P175313 and the concentration of IL-2.
A limitation of our study is that we have included probes without knowing where they bind to the genome (non-annotated probes). Although results for non-annotated probes are less informative than the probes that have been annotated, they do provide some indication of general perturbation of the transcriptome by air pollution. To assess the impact of this decision, we removed the 6848 unannotated probes from the analyses. These analyses did not identify any newly associated probes.

Conclusion

In summary, our study provides some evidence for subtle changes in the transcriptome in relation to long-term exposure to NO_X. Some of these changes are consistent with transcriptome perturbations that have been observed among tobacco smokers. Our results contribute to the further elucidation of the pathways through which long-term exposure to air pollution induces adverse health effects.

Supplemental material

Appendix 1: Variable selection methods

Elastic Net

Elastic Net is a penalized version of the generalized linear regression model where parameter estimates achieved by using a combination of two penalties (so-called Ridge and Lasso penalties) (Zou and Hastie, 2005). Therefore, this approach dose have both merits of Ridge regression, which provides a stable coefficient estimates by shrinking coefficients from correlated predictors toward zero and Lasso regression, which provides a sparse selection of coefficient by shrinking coefficients of the least informative predictor to zero value (Friedman et al., 2010).
We used cross-validation to determine the optimal values for the penalty parameters such that the optimal values yielded the minimum mean-squared error of prediction from repeated 10-fold cross validation (Zou and Hastie, 2005).
To control the number of falsely selected predictors by Elastic Net, we applied the stability selection technique proposed by Meinshausen & Bhlmann (2010). Stability selection involves repeatedly applying Elastic Net to random subsamples of the data and choosing those predictors that are selected most frequently across the subsamples (Meinshausen and Bühlmann, 2010). We accounted for multiple testing by setting family-wise error rate (FWER) to 5%. We set the threshold of selection probability (probability of selecting a predictor by algorithm; π) to 0.6.

Graphical Unit Evolutionary Stochastic Search (GUESS)

Guess is a computationally optimized Bayesian variable selection approach that uses an advanced stochastic search Markov Chain Monte Carlo (MCMC) algorithm (Bottolo et al., 2013). This approach combines tempered multiple chains run together with genetic algorithms that can be ensure both improved mixing of the sampler and exchange of information across chains (Bottolo et al., 2013).
The best models are selected based on *posterior model probability*. The models with highest visited frequency are the best models. The best probes are selected based on Marginal Posterior Probability of Inclusion (MPPI), which is the probability that a predictor is included in a selected model. We ran 5 different chains in GUESS for 90k iterations and discarded the first 30k iterations as burn-in. Expected and standard deviation of the model size were set to 3 and 5, respectively.
As our dataset has a high dimension in gene expression probes and only one air pollution exposure variable, we included the probes as predictor variables and exposure to NO_X as response variable in both Elastic Net and GUESS. To properly adjust for potential

confounding and nuisance factors, we de-noised the data by a univariate linear mixed model before applying the variable selection methods as suggested by Xing et al. (2011) and Bottolo et al. (2013).

Appendix 2: Gene Set Enrichment Analysis (GSEA)

Methods

Our implementation of GSEA involved the following steps.

- Cluster the 29662 probes in our current analysis based on direction of their association (positive or negative) with exposure to NO_X.
- Rank the probes in each cluster based on the strength of the association (t-score).
- Gene annotation of all probes. In case multiple probes were associated to the same gene the probe with the highest t-score was retained in the analysis.
- Assess the overlap between annotated genes from our analyses and the gene-sets related to cigarette smoke and to "air pollution and CVD".
- In each cluster assess whether the distribution of overlapping genes among all annotated genes from our analysis is random, or whether this overlap occurs primarily among our most up- or downregulated genes (i.e. enrichment) using GSEA software (Subramanian et al., 2005).

Results

Figures S3.1 and S3.2 illustrate the outcome of GSEA. The intensity of the blue and red color bars represents the absolute value of the t-score from univariate analysis (Blue: decreased expression; red: increased expression). Each vertical line represents a gene that was included in our analyses and in the gene-set to which we compare our results (an overlapping gene). A qualitative indication for enrichment of the overlapping genes can be acquired by assessing the (lack of) symmetry of the graph, where asymmetry points towards enrichment. GSEA software provides a quantitative indication for enrichment based on the enrichment score which reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes. The enrichment score is calculated by walking down the ranked list of genes, increasing a running-sum statistic (the heigth

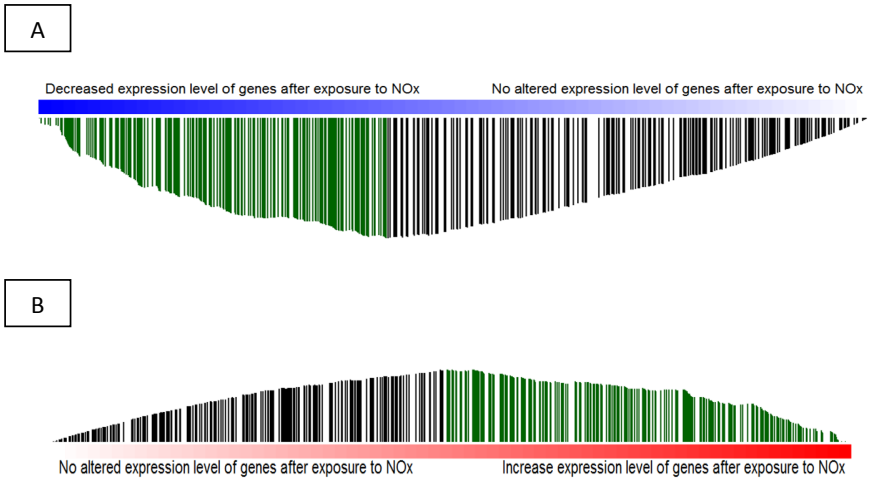

Figure S3.1: A) Distribution of genes underexpressed in association with smoking status vs. down-regulated genes ranked according to their t-statistics in association with NO_X. B) Distribution of genes overexpressed in association with cigarette smoke exposure vs. up-regulated genes ranked according to their t-statistics in association with NO_X. The bottom color bar represents the alteration in level of gene expression after exposure to NO_X (Blue: decreased in expression level, red: increase in expression level). Each vertical line represent one of the genes whose expression is associated with cigarette smoke exposure. The height of the bar represents the running GSEA enrichment score. Green vertical lines represent the leading edge genes responsible for the core enrichment of each gene set.

of the edges) when a gene is in the gene set and decreasing it when it is not. The magnitude of the increment depends on the t-statistic. The enrichment score is the maximum deviation from zero encountered in walking the list. All genes that have been assessed before the maximum deviation is reached are part of the core enrichment set (the green edges). GSEA software uses the Kolmogorov-Smirnov statistic to calculate a p-value for the enrichment score (Subramanian et al., 2005).

Appendix 3

All results of univariate analysis has been uploaded as supplementary file to Zenodo.org under DOI10.5281/zenodo.50661.

Table S3.1: Name and information of candidate genes (both empirical and biological) based on our selection criteria found in the literature.

Gene Symbol	Gene Name	Probes[a]	References
Empirical candidate genes[b]			
NOS2A	Nitric oxide synthase 2	A_23_P502464	(Bind et al., 2014a; Pettit et al., 2012)
SRGAP2	SLIT-ROBO Rho GTPase activating protein 2	A_32_P50123 A_24_P357869 A_23_P200829	(Chuang et al., 2010; Van Leeuwen et al., 2006)
GSTM1	Glutathione S-transferase	A_23_P115407	(Ji and Hershey, 2012; Romieu et al., 2010; Madrigano et al., 2011)
GSTP1	Glutathione S-transferase pi 1	A_23_P202658	(Romieu et al., 2010; Madrigano et al., 2011)
HMOX1	Hemeoxygenase (decycling) 1	A_23_P120883	(Bind et al., 2014a; Peretz et al., 2008)
IL-6	Interleukin 6	A_23_P71037	(Bind et al., 2014a; Adam et al., 2014)
Biological candidate genes[c]			
HMOX1[f]	Hemeoxygenase (decycling) 1	A_23_P120883	
IL-6g	Interleukin 6	A_23_P71037	(Wittkopp et al., 2016)
GCLM	Glutamate-cysteine ligase, modifier subunit	A_23_P103996 A_32_P177953	
CAT	Catalase	A_23_P105138	
KLF2	Kruppel-like factor 2	A_23_P119196 A_24_P151305	
ATF4	Activating transcription factor 4	A_23_P120933 A_23_P120941	
SELP	Selectin P (antigen CD62)	A_23_P137697	
SOD1	Superoxide dismutase 1 [Cu-Zn]	A_23_P154840	
CYP1A1	Cytochrome P450, subfamily A1	A_23_P163402	
GSTP1	Glutathione S-transferase pi	A_23_P202658	
TXNRD1	Thioredoxin reductase 1	A_23_P204581	
CYP1B1	Cytochrome P450, subfamily B1	A_23_P209625	
AHR Aryl	hydrocarbon receptor	A_23_P215566	
NOX1	NADPH oxidase 1	A_23_P217280	
GPX1	Glutathione peroxidase 1	A_23_P250671	
GCLC	Glutamate-cysteine ligase, catalytic subunit	A_23_P352879	
NFE2L2	Nuclear factor (erythroid-derived 2)-like 2	A_23_P5761 A_24_P936444	
ATF6	Activating transcription factor 6	A_23_P62907	
CXCL1	Chemokine (C-X-C motif) ligand 1	A_23_P7144	
IL1B	Interleukin-1beta	A_23_P79518	
IL8	Interleukin-8	A_23_P89380	
CCL2	Chemokine (C-C motif) ligand 2	A_23_P89431	
IL6R	Interleukin 6 receptor	A_24_P379413	
PTGS2	Prostaglandin-endoperoxide synthase 2	A_24_P77008	
SOD2	Superoxide dismutase 2 [Mn], mitochondrial	A_24_P935819	
HSPA8	Heat shock 70kda protein 8	A_32_P13728	
DUSP1	Dual specificity phosphatase 1		
F3	Tissue factor (Thromboplastin, coagulation factor III)		
MPO	Myeloperoxidase		
NQO1	NAD(P)H dehydrogenase, quinone 1		
PON1	Paraoxonase 1		
TNF	Tumor necrosis factor		
TNFRSF1B	Tumor necrosis factor receptor superfamily, member 1B		
XBP1	X-box binding protein 1		
XDH	Xanthine dehydrogenase		

Note: [a] Corresponding probes for the genes from the literature, [b] Empirical candidate genes were selected based on results of previous epidemiological direction in at least two studies.
[c] Biological candidate genes were selected from published studies of gene expression-pollutant associations by Wittkopp et al. (2016). These genes were involved in 10 biological pathways relevant to air pollution exposure responses. The pathways are as follows: coagulation, Klf2-mediated immune response, NF-B signaling, acute phase response , Nrf2-mediated oxidative stress response, endoplasmic reticulum stress (UPR), glutathione metabolism, phase I and phase II metabolism, endogenous reactive oxygen species (ROS) production, and cytokine signaling)

Table S3.2: Associations between long-term exposure to NO_X and CpG island of methylation level of noteworthy genes (q-value <0.2) in the Italian cohort (probes that we identified previously are shown in bold)

Probe id	Gene Name	β^a	SE^b	p-value	q-value^c
cg00419298	*AHCYL2*	6.9E-04	1.3E-03	6.1E-01	0.90
cg00574819	*LARP1B*	-1.5E-03	3.1E-03	6.3E-01	0.90
cg00783759	*LARP1B*	6.3E-04	7.4E-04	3.9E-01	0.82
cg01360773	*LARP1B*	-1.8E-03	4.2E-03	6.7E-01	0.90
cg01933405	*AHCYL2*	9.3E-03	7.3E-03	2.1E-01	0.69
cg02448222	*AHCYL2*	6.0E-03	2.4E-03	1.7E-02	0.29
cg02491457	*AHCYL2*	4.2E-04	4.3E-03	9.2E-01	0.96
cg03043827	*LARP1B*	1.3E-03	1.7E-03	4.6E-01	0.85
cg03493639	*MTMR2*	NA	NA	NA	NA
cg03793937	***MTMR2***	**-6.7E-03**	**2.0E-03**	**1.9E-03**	**0.14**
cg03822873	*AHCYL2*	5.8E-03	7.2E-03	4.2E-01	0.82
cg04323313	*MTMR2*	2.3E-03	2.1E-03	2.9E-01	0.72
cg04392234	*CXCL8*	-9.8E-03	5.5E-03	8.0E-02	0.60
cg04453364	*MTMR2*	-1.3E-02	4.6E-03	2.0E-02	0.29
cg04760493	*LARP1B*	4.0E-03	9.3E-03	6.7E-01	0.90
cg05788526	*MTMR2*	3.0E-03	2.3E-03	2.2E-01	0.69
cg06156768	*LARP1B*	8.1E-04	1.5E-03	5.8E-01	0.90
cg06474646	*AHCYL2*	-1.2E-04	1.2E-03	9.2E-01	0.96
cg06850158	*CXCL8*	-4.0E-03	6.1E-03	5.2E-01	0.89
cg06988775	***AHCYL2***	**3.3E-03**	**1.1E-03**	**3.8E-03**	**0.14**
cg07259151	*AHCYL2*	-3.8E-03	3.9E-03	3.3E-01	0.75
cg08412316	*MTMR2*	-2.1E-03	7.4E-03	7.9E-01	0.93
cg08738403	*AHCYL2*	2.7E-03	3.2E-03	4.0E-01	0.82
cg09424566	*AHCYL2*	-9.2E-03	4.1E-03	2.8E-02	0.33
cg09801523	*LARP1B*	1.2E-03	1.6E-03	4.3E-01	0.82
cg10640379	*LARP1B*	1.5E-03	1.4E-03	2.9E-01	0.72
cg11594622	*NOX1*	-3.8E-03	9.0E-03	6.8E-01	0.90
cg11814672	*CXCL8*	-1.7E-03	2.0E-03	4.3E-01	0.82
cg11961845	*AHCYL2*	1.1E-03	6.6E-03	8.7E-01	0.96
cg12621141	*NOX1*	1.7E-04	5.9E-03	9.8E-01	0.99
cg13077762	*LARP1B*	1.6E-03	1.4E-03	2.8E-01	0.72
cg13365375	*AHCYL2*	1.4E-03	1.0E-03	1.7E-01	0.68
cg13504856	*AHCYL2*	-4.4E-03	2.5E-03	8.3E-02	0.60
cg14022202	*MTMR2*	5.5E-03	2.3E-03	1.9E-02	0.29
cg14256977	*MTMR2*	-2.4E-03	2.2E-03	2.9E-01	0.72
cg14703024	*MTMR2*	-3.0E-03	2.4E-03	2.1E-01	0.69
cg14755690	*AHCYL2*	-1.7E-03	4.1E-03	6.9E-01	0.90
cg14972814	*MTMR2*	-1.9E-03	7.4E-03	7.9E-01	0.93
cg15488143	*AHCYL2*	-6.8E-04	4.0E-03	8.7E-01	0.96
cg15664152	*AHCYL2*	6.5E-03	6.0E-03	2.8E-01	0.72
cg15927257	*AHCYL2*	1.3E-03	1.1E-03	2.7E-01	0.72
cg16166368	*AHCYL2*	3.4E-03	3.2E-03	3.0E-01	0.72
cg16468729	*CXCL8*	1.6E-03	5.6E-03	7.7E-01	0.93
cg16584911	*AHCYL2*	9.5E-04	1.3E-03	4.8E-01	0.86
cg16584945	*LARP1B*	-2.3E-03	1.4E-03	1.6E-01	0.68
cg16986973	*MTMR2*	2.0E-03	1.3E-03	1.2E-01	0.68
cg17483792	*AHCYL2*	-2.4E-03	2.2E-03	2.9E-01	0.72
cg17539479	*AHCYL2*	-1.2E-03	1.6E-03	4.9E-01	0.86
cg17712092	*LARP1B*	-7.1E-03	4.9E-03	1.6E-01	0.68
cg17827670	*AHCYL2*	2.5E-03	4.6E-03	5.9E-01	0.90
cg17920195	*LARP1B*	-8.2E-03	4.7E-03	9.0E-02	0.60

[a] Effect estimates. [b] SE, standard error of β. [c] False Discovery Rate correction for p-value

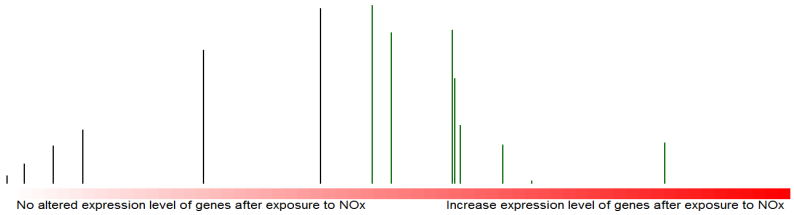

Figure S3.2: Distribution of genes involved in biological pathway that cause cardiovascular disease vs. up-regulated genes ranked according to their t-statistics in association with NO_X. The bottom color bar represents the increase expression level of genes after exposure to NO_X. Each vertical line represent one of the genes whose expression is assochiated with air pollutant-induced cardio vascular disease. The height of the bar represents the running GSEA enrichment score. Green vertical lines represent the leading edge genes responsible for the core enrichment of each gene set.

(continued) Associations between long-term exposure to NO_X and CpG island of methylation level of noteworthy genes (q value <0.2) in the Italian cohort (probes that we identified previously are shown in bold)

Probe id	Gene Name	β	SE	p-value	q-value
cg18007641	*CXCL8*	-2.3E-03	7.9E-03	7.7E-01	0.93
cg18467841	*LARP1B*	-6.9E-04	3.7E-03	8.5E-01	0.96
cg18980650	*NOX1*	-6.8E-03	4.5E-03	1.6E-01	0.68
cg19541976	*MTMR2*	2.2E-03	3.4E-03	5.4E-01	0.89
cg19910937	*LARP1B*	1.4E-03	5.1E-03	7.9E-01	0.93
cg20013533	*AHCYL2*	-8.1E-03	1.0E-02	4.3E-01	0.82
cg20195272	*MTMR2*	-3.5E-04	3.7E-03	9.2E-01	0.96
cg20734154	*AHCYL2*	1.5E-03	2.6E-03	5.9E-01	0.90
cg20915333	*AHCYL2*	-3.7E-03	2.4E-03	1.3E-01	0.68
cg21483539	*LARP1B*	-6.1E-03	4.3E-03	1.6E-01	0.68
cg21871803	*AHCYL2*	-8.6E-04	2.8E-03	7.6E-01	0.93
cg21931372	*AHCYL2*	3.1E-03	2.4E-03	2.1E-01	0.69
cg22584802	*AHCYL2*	2.5E-03	5.7E-03	6.6E-01	0.90
cg23099587	*CXCL8*	4.1E-03	6.6E-03	5.3E-01	0.89
cg23963595	*MTMR2*	-1.9E-02	8.3E-03	3.9E-02	0.36
cg24087613	*LARP1B*	2.4E-03	4.5E-03	6.0E-01	0.90
cg25314817	*MTMR2*	-5.1E-03	6.0E-03	4.0E-01	0.82
cg25969460	*AHCYL2*	-9.5E-03	4.2E-03	3.2E-02	0.33
cg26388009	*NOX1*	1.9E-03	7.3E-03	8.0E-01	0.93
cg26618618	*MTMR2*	2.2E-03	4.8E-03	6.6E-01	0.90
cg26838532	*LARP1B*	-4.2E-03	3.1E-03	1.8E-01	0.69
cg27067621	*NOX1*	1.3E-03	9.4E-03	8.9E-01	0.96
cg27284840	*AHCYL2*	1.4E-05	5.2E-04	9.8E-01	0.99

Table S3.3: biological role of noteworthy genes based on literature review.

Gene Symbol	Gene Name	Biological function
AHCYL2	Adenosylhomocysteinase like 2	Encodes a protein involved in the synthesis of L-homocysteine from S-adenosyl-L-homocysteine. L-homocysteine biosynthesis, which is an amino-acid involved in protein biosynthesis (Consortium et al., 2014; Amberger et al., 2009).
MTMR2	MYOTUBULARIN-RELATED PROTEIN 2	Encodes a protein involved in phosphatase activity, crucial for the proper functioning of peripheral nerves (Consortium et al., 2014; Amberger et al., 2009).
LARP1B	La ribonucleoprotein domain family member 1B	Encodes a protein involved in RNA binding and mitophagy in response to mitochondrial depolarization (Consortium et al., 2014; Amberger et al., 2009)
NOX1	NADPH OXIDASE HOMOLOG 1	Encodes a protein that mediates proton currents of resting phagocytes and other tissues. It participates in the regulation of cellular pH and is blocked by zinc (Consortium et al., 2014; Amberger et al., 2009).
IL-8	Interleukin-8	Encodes a protein which is a chemotactic factor that attracts neutrophils, basophils, and T-cells, but not monocytes. It is also involved in neutrophil activation. It is released from several cell types in response to an inflammatory stimulus (Consortium et al., 2014; Amberger et al., 2009).

CHAPTER 4

ACUTE CHANGES IN DNA METHYLATION IN RELATION TO 24HR PERSONAL AIR POLLUTION EXPOSURE MEASUREMENTS: A PANEL STUDY IN FOUR EUROPEAN COUNTRIES

This chapter is published as: Mostafavi N, Vermeulen R, Ghantous A, Hoek G, Probst-Hensch N, Herceg Z, Tarallo S, Naccarati A, Kleinjans J.C.S, Imboden M, Jeong A, Morley D, Amaral A.F.S, van Nunen E, Gulliver J, Chadeau-Hyam M, Vineis P, and Vlaanderen J. Acute changes in DNA methylation in relation to 24hr personal air pollution exposure measurements: A panel study in four European countries. *Environment International*, 2018 Jul 24; 120:11-21. doi: 10.1016/j.envint.2018.07.026.

One of the potential mechanisms linking air pollution to health effects is through changes in DNA-methylation, which so far has mainly been analyzed globally or at candidate sites. We investigated the association of personal and ambient air pollution exposure measures with genome-wide DNA-methylation changes.

We collected repeated 24-hour personal and ambient exposure measurements of particulate matter ($PM_{2.5}$), $PM_{2.5}$ absorbance, and ultrafine particles (UFP) and peripheral blood samples from a panel of 157 healthy non-smoking adults living in four European countries. We applied univariate mixed-effects model to investigate the association between air pollution and genome-wide DNA methylation perturbations at single CpG (cytosine-guanine dinucleotide) sites and in Differentially Methylated Regions (DMRs). Subsequently, we explored the association of air pollution-induced methylation alterations with gene expression and serum immune marker levels measured in the same subjects.

Personal exposure to $PM_{2.5}$ was associated with methylation changes at 13 CpG sites and 69 DMRs. Two of the 13 identified CpG sites (mapped to genes *KNDC1* and *FAM50B*) were located within these DMRs. In addition, 42 DMRs were associated with personal $PM_{2.5}$ absorbance exposure, 16 DMRs with personal exposure to UFP, 4 DMRs with ambient exposure to $PM_{2.5}$, 16 DMRs with ambient $PM_{2.5}$ absorbance exposure, and 15 DMRs with ambient UFP exposure. Correlation between methylation levels at identified CpG sites and gene expression and immune markers was generally moderate.

This study provides evidence for an association between 24-hour exposure to air pollution and DNA-methylation at single sites and regional clusters of CpGs. Analysis of differentially methylated regions provides a promising avenue to further explore the subtle impact of environmental exposures on DNA-methylation.

Introduction

Air pollution is now the world's largest single environmental health risk, leading to approximately 9 million deaths in 2015 alone (16% of all deaths worldwide)(Landrigan et al., 2017). The death burden relates mostly to the risk of cardiovascular and respiratory diseases (Brunekreef and Holgate, 2002; Hoek et al., 2013) and to lung cancer (Loomis et al., 2013). Oxidative damage, inflammation, and endothelial dysfunction have been suggested as potential underlying mechanisms (Demetriou et al., 2012; Pettit et al., 2012), though the identity and the role of potentially involved biological pathways are far from clear. It has been proposed that systemic effects of air pollution might be detected through assessing variation in epigenetics which can provide further mechanistic insights in air-pollution-health associations (Baccarelli et al., 2009; Herceg et al., 2017; Nawrot and Adcock, 2009). The most studied and best understood epigenetic modification is DNA methylation, the covalent addition of a methyl group to a cytosine primarily in the context of a cytosine-guanine dinucleotide (CpG) (Bellavia et al., 2013; De Prins et al., 2013).
Early studies have provided indications for an impact of air pollution on peripheral blood global methylation using repetitive elements in long (LINE-1) and short (ALU) interspersed nuclear element (Baccarelli et al., 2009; Bellavia et al., 2013; De Prins et al., 2013). Furthermore, sequence-specific analyses have yielded evidence for differential methylation in regions of genes such as inducible nitric oxide synthase (iNOS), tissue factor intercellular adhesion molecule 1, toll-like receptor 2, interferon-y, and interleukin-6 (Bind et al., 2014a; Madrigano et al., 2011; Tarantini et al., 2009; Wang et al., 2016).
In recent years, several cohort studies have conducted peripheral blood methylome-wide association studies of air pollution using the Illumina 450 BeadChip platform (Gruzieva et al., 2017; Panni et al., 2016; Plusquin et al., 2017). The KORA F3, F4 cohort (N=2,300 individuals), the Normative Aging Cohort (N= 657 individuals), the Multi-Ethnic Study of Atherosclerosis (MESA) (N=1,207 individuals), and the EPIC study (N= 613 individuals) identified several sets of CpG sites at which levels were associated with exposure to air pollution (Chi et al., 2016; Panni et al., 2016; Plusquin et al., 2017). In addition, a meta-analysis of multiple European and North American studies identified a set of three CpGs for which methylation levels in children were related to air pollution exposure (NO_2) of the mother during pregnancy (Gruzieva et al., 2017). To estimate exposure to air pollution, these methylome-wide studies used either spatially resolved models that predict long-term (1 year) average levels (MESA, EPIC), or temporally resolved (2-day, 7-day, and 28-day trailing averages) estimates based on citywide monitoring data (KORA). The relation between 24-hour personal measurements of air pollution and peripheral blood methylome-wide changes is currently unknown.
Our study contributes to the existing literature by providing insight into the relation between 24-hour personal and ambient exposure measurements of particulate matter (PM) smaller than 2.5 μm ($PM_{2.5}$), $PM_{2.5}$ absorbance (as a proxy of black carbon and soot),

ultrafine-particulates (UFP) and genome-wide changes in the peripheral blood methylome. For comparison purposes, our study also incorporates modeled long-term average estimates of the same pollutants.
Associations with air pollutants were identified through genome-wide univariate screening of the peripheral-blood methylome and by analyses of differentially methylated regions (DMRs). PM from different pyrogenic sources may have toxicological similarities, we, therefore, assessed the overlap between our results and a set of CpG sites previously reported to be associated with cigarette smoke. In addition, to gain a better understanding of the functional role of differential methylation at air pollutant related CpG sites, we investigated their association with expression data of closely located genes and a set of 13 immune markers measured in the same individuals.

Material and methods

Study population

Our study, which is part of the EXPOsOMICS project (Vineis et al., 2017), was conducted in four European countries (Italy, Netherlands, Switzerland, and United Kingdom) from December 2013 to February 2015. We aimed to recruit 40 individuals from each country of whom half lived on busy roads (road with >10K vehicles per day; house on ground/first floor) and the other half on quiet roads (at least 100m away from busy roads) in order to increase contrast in air pollutant levels. 157 healthy never-smoking adults between 50 and 70 years old, with no history of a pulmonary or cardiovascular disease, diabetes, or other acute or chronic health conditions participated in our study (Table 4.1). Participants performed three personal exposure monitoring (PEM) sessions in different seasons (summer, winter, and spring/autumn) spread over one year year (on average 28 days (P25-P75): (20-37) between repeated sample collection). During each PEM session, air pollution measurements and blood samples were collected for each subject. Subjects performed their own daily routine and filled in a questionnaire on food intake and time-activity. Blood samples from the first two PEM sessions were analyzed. Ethics approval for each country was obtained from local authorized review boards. All subjects signed an informed consent before participation.

Exposure assessment

During each PEM session, participants carried a backpack containing air pollution sensors to measure 24-hour personal air pollution exposure and a belt to measure locations (GPS) and an accelerometer. Simultaneously, with the personal measurements, 24-hour air pollution was measured outdoor at the subject's home address.

$PM_{2.5}$

For the personal measurements, 24-hour $PM_{2.5}$ was sampled on a $2\mu m$ pore size Teflon filter, using a BGI 400 pump unit and GK 2.05 SH (BGI Inc., Waltham MA, USA) cyclone. Changes in filter weight (pre and post-weighted) measured in a central lab were used to assess $PM_{2.5}$ concentration. Ambient $PM_{2.5}$ was collected via installing the same devices as used for personal sampling, directly outside of the subject's home in the garden or balcony. Filter reflectance was measured to assess $PM_{2.5}$ absorbance concentration (soot levels) in both personal and home outdoor monitoring. Details on analytical procedures have been published previously (Eeftens et al., 2012a).

Ultrafine particles (UFP)

Ultrafine particles (UFP) were continuously monitored with a MiniDiSC (Testo AG, Lenzkirch, Germany). The MiniDiSC operated at the flow of 1000 ml/min measuring particles from 10-300nm at 1-second intervals. We used the term UFP to refer to the median of particle number counts from the MiniDiSC over 24-hour PEM session. Additionally, in The Netherlands and Switzerland ambient UFP was collected over each 24-hour PEM session at the subject's home facade using MiniDiSCs. Details on methods and quality control have been published previously (Nunen et al., 2017).

Average exposure and modeled long-term exposure

In addition to the main personal exposure metrics, reflecting exposure to air pollution in the 24-hour before blood collection, we calculated the average of the 24-hour personal air pollution measurements across the three PEM sessions as a proxy for long-term average personal exposure. We also used estimated annual average ambient concentrations of air pollution for each individual using the land-use-regression (LUR) prediction models of the ESCAPE study (European study of cohorts for air pollution effects) and EXPOsOMICS (Eeftens et al., 2012a; Nunen et al., 2017) as a proxy for long-term modeled ambient exposure. Further details are provided in the supplemental materials S1.

Physical activity

During each PEM session, physical activity was continuously measured with an accelerometer (Actigraph GT3X+, Pensacola FL, USA), which was attached to the participant's belt. We estimated energy expenditure in the metabolic equivalent of task (MET) from the raw acceleration data to express participant's physical activity over 24-hour before blood collection (Ainsworth et al., 2000).

Biological sample collection

Biological samples were collected in the morning (9:00 - 11:00) after each PEM session. Collected blood samples were processed, aliquoted, and transferred to $-80^{\circ}C$ within 2 hours.

DNA methylation measurements

Genome-wide DNA methylation analyses were performed on peripheral blood samples (buffy coats) using the Illumina Infinium HumanMethylation450 BeadChip (HM450K), with optimized robotic pipelines (Ghantous et al., 2014; Morris and Beck, 2015). The HM450K array measures DNA methylation at 485,512 cytosine positions across the human genome (CpG). Samples were randomized within each country. All laboratory procedures were carried out at the Epigenetics Group, International Agency for Research on Cancer (IARC, Lyon, France) according to manufacturers' protocols. CpG sites with more than 10% missing methylation values were replaced by the average methylation level of that particular CpG site. After quality control exclusions, 459,333 CpGs was selected for further analyses.

The percentage of methylation of a given cytosine (β-value) corresponds to the ratio of the methylated signal over the sum of the methylated and unmethylated signals. The M-value was calculated as the logit transformation of the β-value.The M-value was used in statistical analysis because it is less heteroscedastic and performs better in differential methylation analysis (Du et al., 2010). For each CpG, extreme values were considered as outliers if they fell out of the interquartile range (IQR) (2.5th percentile $3 \times$ IQR, 97.5th percentile $+ 3 \times$ IQR) interval and were removed from further analysis. The maximum percentage of outlier samples was 3.7% in less than 10% of methylation markers.

DNA methylation sites were mapped to genes based on data provided by Illumina (Bibikova et al., 2011) (for the individual probes) and on the February 2009 human reference sequence (GRCh37) using the UCSC genomes browser (Meyer et al., 2012; Pedersen et al., 2013) (both for the individual probes and the DMRs). Functional analyses of mapped genes that were identified in either the univariate analysis or DMRs to be associated with air pollution were assessed using NIH-DAVID bioinformatics resources (Huang et al., 2009).

Gene expression

Total RNA was isolated from stabilized blood specimens (400 μl of whole blood and 1600 μl of RNA later) using RiboPureTM-Blood (Ambion), according to the manufacturers instructions. RNA from 301 samples was hybridized on Agilent $8 \times 60K$ Whole Human Genome microarrays for mRNA. Raw data on pixel intensities were extracted using Agilent Feature Extraction Software. After quality control exclusion, a total of 23,557 probes

were available for the subset (245 of 301) samples. Further details are provided in the supplemental materials S2.

Immune markers

A panel of 23 immune markers was measured in serum samples of all subjects using an $R\&D$ Systems (Abingdon, UK) Luminex screening assay. The panel included interleukins (IL) 1β, IL-4, IL-5, IL-6, IL-8, IL-10, IL-13, IL-17, IL-25, tumor necrosis factor alpha (TNF-α), eotaxin, IL-1 receptor antagonist (IL-1ra), CXC chemokine ligand 10 (CXCL10), epidermal growth factor (EGF), fibroblast growth factor beta (FGF-β), granulocyte colony-stimulating factor (G-CSF), melanoma growth stimulatory activity/growth-related oncogene (GRO), chemokine (C-C motif) ligand 2 (CCL2), C-C motif chemokine 22 (CCL22), macrophage inflammatory protein-1 beta (MIP-1β), vascular endothelial growth factor (VEGF), Myeloperoxidase (MPO), and periostin. In addition, C-reactive protein (CRP) was assessed using $R\&D$ System Solid Phase ELISA. The panel of immune markers was a priori selected based on an informal review of the literature on air pollution, asthma, CVD, colon cancer, and lung cancer. This panel was applied in several studies part of the EXPOsOMICS project (Vineis et al., 2017). After quality control assessment, 13 immune markers remained for further analysis. Immune markers were log-transformed as the distributions were skewed. Further details are provided in the supplemental material S3.1.

Data analysis

Univariate mixed-effects model

We used linear mixed effect models setting a random intercept (i) for each subject, capturing the correlation among measurements within the same subject ($N = 2$ observation per subject), (ii) for each microtiter plate ($N = 37$) and (iii) for the position of the sample on the chip ($N = 12$), capturing technically-induced variation potentially biasing the DNA methylation measurements. To control for potential confounding, models were all adjusted for sex, age, body mass index (BMI) (kg/m^2), education (primary, secondary school, university), country (Italy, Netherlands, Switzerland, and United Kingdom), season and physical activity expressed in MET-values. In addition, temperature and relative humidity as a proxies of meteorological conditions were included in all models.
A number of sensitivity analyses were conducted. First, we ran a minimally adjusted model (only age, sex, and BMI included as covariates) and a model in which country was excluded from the primary set of covariates. To explore the possible influence of country, we ran our original model on each country separately and on the pooled set of data leaving data from one country out one at the time. We explored the stability of the methylation

levels over time by using the random effects part of the mixed model to estimate intra-class correlation coefficients (ICC; the proportion of the total variance in the CpG sites that was due to variation between subjects, and not due to variation between plates, position on the plates, or error). We also assessed the association between DNA methylation and long-term average personal exposures as well as long-term modeled exposure to air pollution in two separate models with the same set of covariate as were considered in our original model. Missing values in covariate and exposure data were imputed using multivariate imputation by chained equations (MICE) in R (Buuren and Groothuis-Oudshoorn, 2011). No covariate had more than 4% missing data (except for MET; 13%). In exposure data, 6.3% of personal $PM_{2.5}$ and 5.6% of ambient $PM_{2.5}$ values were missing. Since the proportion of missing data was low, only one imputation iteration was used in association analysis.

White blood cell composition was estimated based on the DNA methylation data, using the Houseman algorithm (Houseman et al., 2012). Potential confounding by white blood cell composition was evaluated by assessing the association between white blood cell composition and the air pollutants using the linear mixed-models as described above.

Multiple testing adjustments on the resulting p-values were performed using false discovery rate method of Benjamini-Hochberg (BH-FDR) (Benjamini and Hochberg, 1995) at the level of 0.05. Univariate mixed-effects models and imputation were conducted using R version 3.0.2 (packages: $lme4$, $mice$ (De Boeck et al., 2011; Buuren and Groothuis-Oudshoorn, 2011)).

Differentially methylated regions (DMRs)

We identified DMRs using the $comb-p$ package in Python (version 2.7.13) (Pedersen et al., 2012). P-values for each probe generated by the univariate linear mixed-effect models (ordered by chromosomal location) were used as input. We set a window size of 300 bases and restricted DMRs to those that included at least 2 probes. A Stouffer-Liptak-Kechris (SLK) and Šidák corrected P-value <0.05 was used as the statistical cut-off for significance. This resulted in regions of adjacent probes with low P-values that stand up to genome-wide correction. Regions were annotated to nearest gene and CpG island from human genome version "hg19" with CruzDB software package (Meyer et al., 2012; Pedersen et al., 2013). A further description of these reference sets is provided in supplemental material S4.

Overlap with CpG sites associated with cigarette smoking

We assessed the distribution of the CpG sites associated with personal $PM_{2.5}$ exposure among a reference set of CpG sites associated with "exposure to cigarette smoke" (Joehanes et al., 2016) using the gene set enrichment methodology (Mostafavi et al., 2017; Sub-

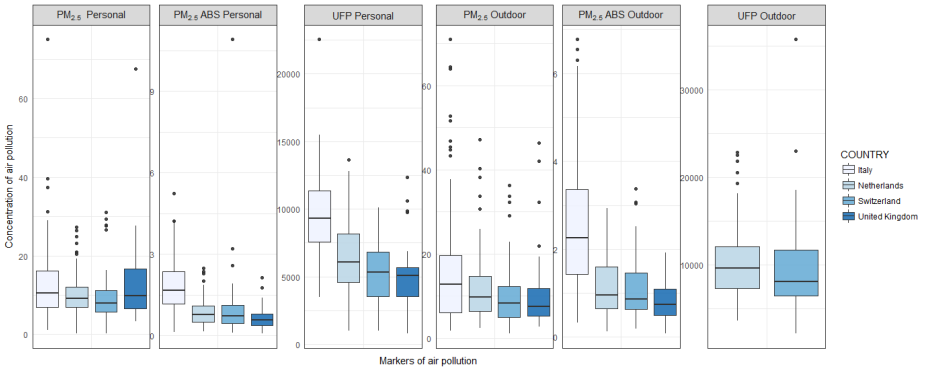

Figure 4.1: Box plots of air pollution concentrations by country. Each panel shows one air pollution marker; personal $PM_{2.5}$; $\mu g/m^3$, personal $PM_{2.5}$ absorbance; m^{-1}, personal UFP; $particles/cm^3$, ambient $PM_{2.5}$; $\mu g/m^3$, ambient $PM_{2.5}$ absorbance; m^{-1}, and ambient UFP; $particles/cm^3$. Each box represents one country (Italy, Netherlands, Switzerland, and United Kingdom). Horizontal lines correspond to medians, and boxes to the 25th75th percentiles; whiskers extend to data within the interquartile range times 1.5, and data beyond this are plotted as dots.

ramanian et al., 2005). A P-value <0.05 was used as the statistical cut-off for enrichment. A further description of these reference sets is provided in the supplemental materials S5.1.

Impact of air pollution on immune markers

We assessed the association between air pollution measurements and immune markers (n=13) using univariate mixed-effects models. We included random intercepts for each subject which captures the correlation among measurements within the same subject and a random intercepts for each microtiter plate (N=9) which captures the nuisance variation generated in the assessment of the immune markers in the model. The same set of covariates as we defined in our primary univariate methylation models were used to control for potential confounding variables. Moreover, the same sensitivity analysis as we described in the methylation analysis was performed to assess the stability of our findings to variation in the confounder models. A further description of these analyses is provided in the supplemental materials S3.2.

Functional readout of the impact of differential methylation at the identified CpG sites

We assessed interrelationships between methylation levels at significant CpG sites identified in the methylation analysis (both univariate and DMRs analysis for each air pollutants separately; 415 CpG sites), expression of mapped genes (n= 58 transcripts), and

Table 4.1: Characteristics of the study participants from four European countries.

Characteristics	Italy[a]	Netherlands[a]	Switzerland[a]	United Kingdom[a]	Pooled[c]
Session[b] (N samples)					
1	43	41	48	23	155
2	42	40	44	20	146
Sex (N individuals)					
Female	22	34	25	15	96
Male	21	7	23	8	59
Education (N individuals)					
Secondary school	30	7	4	11	25
University	13	34	44	12	103
Age (years; median (P_{25}-P_{75}))	60 (56-63)	62 (56-68)	60 (53-68)	63 (57-65)	61 (55-66)
BMI (Kg/m^2; median (P_{25}-P_{75}))	24.8 (22.5-26.6)	24.6 (22.6-27.1)	25.1 (21.6-26.9)	26.8 (24.3-29.3)	25 (22.5-27.9)
Physical activity [d] (Mets; SD)	1.65 (0.17)	1.66 (0.17)	1.45 (0.14)	1.65 (0.19)	1.6 (0.18)
Season (N samples)					
1: spring (21/3-20/6)	34	18	35	15	102
2: summer (21/6-20/9)	16	31	18	16	81
3: autumn (21/9-20/12)	0 32	7	12	51	
4: winter (21/12-20/3)	35	0	32	0	67

[a] In Italy individuals lived in Turin, in the Netherlands individuals lived in Utrecht and Amsterdam, in Switzerland lived in Basel, and in the United Kingdom lived in Norwich.

[b] Personal exposure monitoring and blood collection were performed twice for each individual. Session 1 and 2 were separated by a couple of months (in average 28 days (P25-P75): (20-37) between repeated sample collection); [c] Our study population consisted of 157 non-smoking, healthy adults. Of which 144 indivduals had two measurements and 13 individuals had one measurement. In total, 301 samples were analyzed and included in the statistical analyses; [d] Physical activity was measured using the accelerometer (Actigraph GT3X+) and expressed in the 'The Metabolic Equivalent of tasks' (METs).

the immune markers (n=13) incorporated in this study using the $XMWAS$ package (Uppal et al., 2017). We set the minimum pairwise correlation threshold to 0.4 and used a canonical Partial Least Squares regression (10 components) to generate pair-wise similarity matrices (González et al., 2012). We used a clustering approach to address the paired measurements in this analysis. Connections with a correlation $<$0.4 were not shown in the figure. The correlation structure between the three platforms was displayed using the Cytoscape software environment (Shannon et al., 2003).

Results

Baseline characteristics of the study participants are presented in Table 4.1. Our study population consisted of 157 non-smoking, healthy adults for whom 301 blood samples and air pollutant measurements were collected across two sampling campaigns (144 participants with two measurements and 13 participants with one measurement). About 61% of participants were female and more than 65% of participants had a university or college degree. The population had a median (P_{25}-P_{75}) age of 61 (55-66) years.

In Figure 4.1, we show the distribution of 24-hour average personal and ambient air pollution measurements by country. Highest air pollutant levels were measured in Italy, followed by The Netherlands. Lower levels were measured in the United Kingdom and in Switzerland (Figure 4.1).

We observed a relatively high correlation between measured ambient and personal air

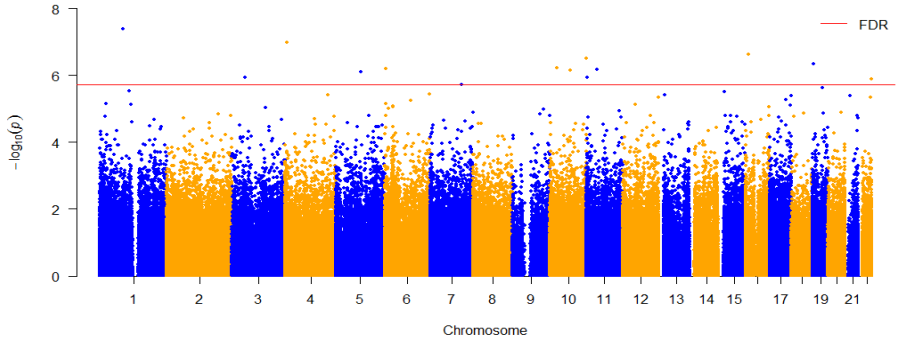

Figure 4.2: Manhattan plot showing fixed-effect p-values of association between personal $PM_{2.5}$ measurements and DNA methylation for fully adjusted model. Each dot corresponds to the CpG methylation site. Horizontal lines correspond to adjusted p-value (FDR) at 0.05 level. 13 CpG sites above the line were considered statistically significant.

pollution concentration measurements, especially for $PM_{2.5}$ absorbance (r = 0.74). Correlations between short-term measured air pollution concentrations and modeled long-term air pollution concentrations were low. Correlations between UFP and the other air pollutants were weak to moderate (r = 0.18; Figure S4.1).
We did not observe any significant association between the air pollutants and estimated white blood cell type composition (Table S4.2); therefore, white blood cell type composition was not used in further analyses.

Univariate analyses

We identified 13 CpG sites (Mapped to 14 genes) at which methylation levels were significantly associated (FDR < 0.05) with personal exposure to $PM_{2.5}$ (Figure 4.2). For ten sites methylation increased with increasing exposure to $PM_{2.5}$, while for the remaining sites methylation decreased (Table 4.2). Between subject intraclass correlation for the 13 CpG sites was generally low (median 0.13), indicating considerable within individual variation. No clear correlation patterns in methylation levels were observed between the 13 CpG sites (Figure S4.2). No significant associations were identified with any of the other personal or ambient air pollution measurements (Table S4.3). However, coefficients for the 13 CpG sites associated with personal $PM_{2.5}$ in the models for the other air pollution markers were largely comparable reaching often nominal statistical significance at the $p<$ 0.05 level, especially for personal $PM_{2.5}$ absorbance and ambient $PM_{2.5}$ (Table S4.4).

Sensitivity analyses indicated that the 13 associations found for $PM_{2.5}$ were robust to the set of confounders considered (Figure 4.3A). Stratified analyses using data from

Table 4.2: CpG sites significantly (at FDR 5%) associated with personal measurements of $PM_{2.5}$.

Probe ID (CpG)	β^a	P-value	FDR	Chromosome	Location	Mapped Gene[b] (region)	Relation to CpG Island	Nearby genes[c] (distance to TSS)
cg26692818	0.23	4.1E-08	0.02	1	87108066	CLCA3P (gene body)	OpenSea	CLCA3P;LOC105378828
cg02556634	0.06	1.0E-07	0.02	4	6449000	PPP2R2C (gene body)	OpenSea	PPP2R2C
cg03873392	-0.13	2.4E-07	0.03	16	10801987		OpenSea	TEKT5
cg07669973	-0.12	3.0E-07	0.03	10	135029365	KNDC1 (gene body)	Island	KNDC1
cg02508204	0.20	6.8E-07	0.04	11	39367436		OpenSea	LINC01493
cg19850855	0.20	7.8E-07	0.04	5	93333655	FAM172A (gene body)	OpenSea	FAM172A
cg20673255	0.14	4.7E-07	0.04	19	5787465	DUS3L (gene body)	N_Shore	DUS3L
cg23468453	0.23	7.1E-07	0.04	10	73906532	ASCC1 (gene body)	OpenSea	ASCC1
cg26559703	0.24	6.1E-07	0.04	10	26853625	APBB1IP (gene body)	N_Shelf	APBB1IP
cg01905633	0.06	6.3E-07	0.04	6	3849391	FAM50B (TSS1500)	Island	FAM50B
cg03408122	0.13	1.3E-06	0.04	22	50903314	SBF1 (gene body)	Island	SBF1
cg05404940	-0.11	1.2E-06	0.04	11	1446911	BRSK2 (gene body)	OpenSea	BRSK2
cg20693615	0.23	1.2E-06	0.04	3	50234688	GNAT1 (3'UTR)	S_Shelf	GNAT1

[a] The coefficient estimate from a model with M-values as the outcome.

[b] Annotations provided by Illumina (Bibikova et al., 2011).

[c] Annotation to nearest gene from human genome version hg19 with *CruzDB* package (Meyer et al., 2012; Pedersen et al., 2013).

each country separately yielded consistent effect size estimates in sign and magnitude (Figure 4.3B). However, estimated strength of association were generally weakened upon stratification and revealed some heterogeneity across countries. Overall, of the 13 CpG sites identified in the pooled analysis, all were significantly associated to personal $PM_{2.5}$ in the Dutch population, 9 in the Basel population, 4 in the Turin population, and one in the Norwich population. The weakening of the strength of the associations can partially be attributed to the smaller sample size due to stratification. To account for this we ran a further sensitivity analysis excluding data from each country one-by-one. Results confirmed consistent effect size estimates with some evidence of stronger effects in The Netherlands for some of the CpGs (Figure 4.3C).

Nominally significant ($p<0.05$) effect estimates from a model that included the long-term average personal $PM_{2.5}$ exposure were correlated ($r = 0.77$; Figure S4.3) to the nominally significant effect estimates from our main model based on 24-hr personal measurements for $PM_{2.5}$, though none of the associations identified in this analysis were genome-wide significant. No correlation was observed with nominally effect estimates from a model that included long-term modeled exposure to $PM_{2.5}$ ($r = 0.02$; Figure S4.3) and no genome-wide significant associations were identified.

Differentially methylated regions (DMRs)

We identified 162 DMRs (SLK-Šidák P-value <0.05) of which 69 DMRs (containing 74 genes and 404 CpG sites) were significantly associated with personal exposure to $PM_{2.5}$ (Table 4.3), 42 DMRs (42 genes and 266 CpG sites) with personal exposure to $PM_{2.5}$ absorbance, 16 DMRs (21 genes and 109 CpG sites) with personal exposure to UFP, 4 DMRs (4 genes and 12 CpG sites) with ambient exposure to $PM_{2.5}$, 16 DMRs with (19 genes and

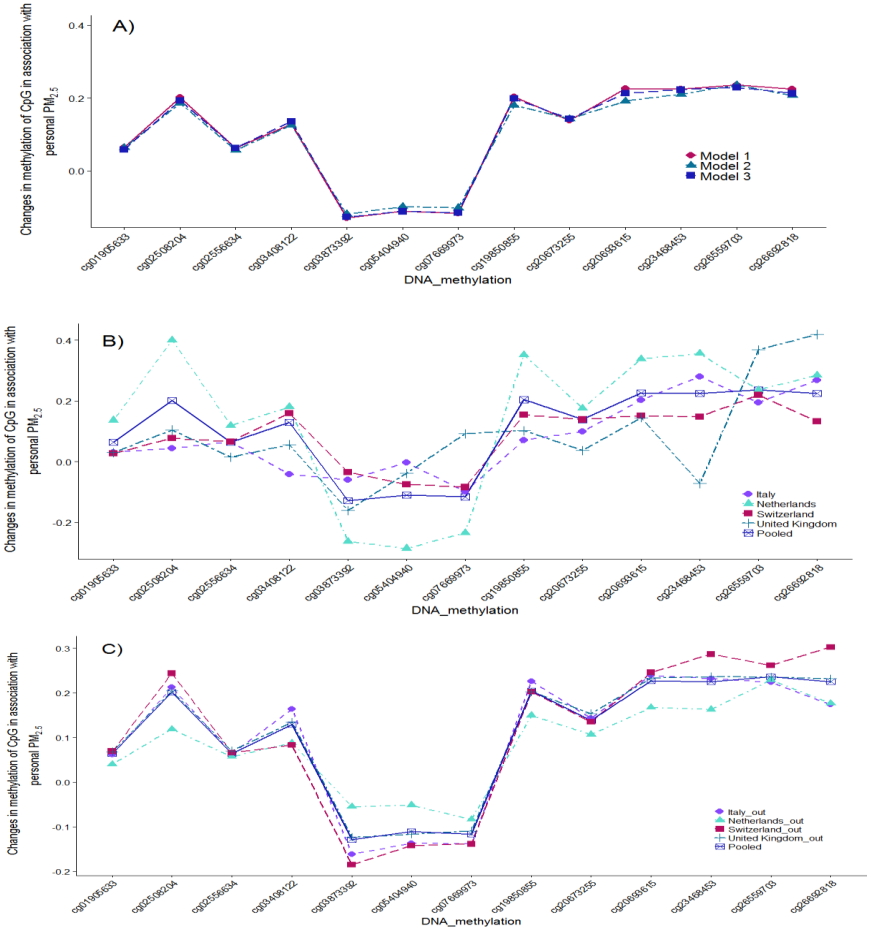

Figure 4.3: (A) Association between personal $PM_{2.5}$ and methylation of 13 identified CpG sites for a different set of confounder adjustment including Model 1 (full model) adjusted for sex, age, BMI, education, season, physical activity, temperature, relative humidity, and country. Model 2 adjusted for sex, age, BMI, and country. Model 3 adjusted for all variables of Model 1 except for country. (B) Association between personal $PM_{2.5}$ and methylation of 13 identified CpG sites in pooled population and stratified by country. (C) Association between personal $PM_{2.5}$ and methylation of 13 identified CpG sites in pooled population and after leaving out countries one-by-one. In Italy individuals lived in Turin, in the Netherlands individuals lived in Utrecht and Amsterdam, in Switzerland lived in Basel, and in the United Kingdom lived in Norwich.

Table 4.3: DMRs associated with personal measurements of $PM_{2.5}$ at an SLK-Sidak P-value < 0.05.

DMR	chromosome	Start	End	No. of probes	SLK-Sidak P value	Mapped Gene [a]	Relation to CpG Island
1	1	26231347	26231585	2	9.1E-03	STMN1	shore
2	1	167408509	167409199	7	5.6E-03	CD247	island
3	1	240656217	240656668	5	4.4E-03	MIR1273E;GREM2	island
4	1	153599479	153600157	8	1.8E-02	S100A13	
5	1	92414221	92414911	8	3.1E-03	BRDT	island
6	1	244952269	244952395	3	3.2E-02	COX20	
7	1	154474344	154474801	5	2.7E-02	SHE;TDRD10	island
8	1	153234037	153234388	4	1.3E-02	LOR	
9	10	135029294	135029462	3	4.5E-02	KNDC1	island
10	10	14051636	14052029	6	9.6E-03	FRMD4A	
11	10	70321554	70321960	6	2.1E-03	TET1	shore
12	11	1413145	1413316	3	2.2E-02	BRSK2	shore
13	11	1456891	1457347	4	7.0E-03	BRSK2	
14	11	1463541	1463663	4	5.8E-04	BRSK2	shore
15	11	19736150	19736334	5	1.6E-03	LOC100126784;NAV2	shore
16	11	1474588	1474842	3	3.4E-02	BRSK2	shore
17	12	95945120	95945385	4	9.2E-03	USP44	shore
18	12	33205702	33205863	3	1.8E-02	PKP2	
19	12	4918848	4919231	5	2.3E-02	KCNA6	island
20	12	71552188	71552570	4	7.3E-03	TSPAN8	
21	12	46319990	46320111	4	3.1E-02	SCAF11	
22	12	47219626	47220093	11	1.5E-02	SLC38A4	
23	12	6649733	6649995	4	9.3E-03	IFFO1	island
24	15	83240550	83240792	7	2.5E-02	CPEB1	
25	15	57510284	57510576	3	1.0E-02	TCF12	
26	16	85551478	85551749	3	2.1E-02	GSE1	island
27	16	84691498	84691851	3	2.8E-02	KLHL36	island
28	16	89408076	89408568	6	3.0E-02	ANKRD11	island
29	16	67686832	67687120	3	2.0E-03	CARMIL2	island
30	17	33759484	33760250	10	7.9E-06	SLFN12	
31	18	74799250	74799573	4	1.1E-03	MBP	island
32	19	45448959	45449302	5	4.7E-02	APOC4-APOC2	
33	19	2888943	2889257	3	3.0E-03	ZNF556	island
34	19	55660514	55660626	5	2.3E-02	TNNT1	
35	19	57742112	57742445	9	1.1E-06	AURKC	island
36	19	23941207	23941768	7	6.2E-04	ZNF681	island
37	19	11649462	11649702	5	2.2E-02	CNN1	
38	19	45975694	45976196	5	4.2E-02	FOSB	island
39	19	38281047	38281560	5	9.4E-04	ZNF573	island
40	2	239983744	239984106	4	5.6E-03	HDAC4	
41	2	207506480	207507164	9	3.0E-02	FAM237A	island
42	2	38892846	38893252	6	4.8E-04	GALM	
43	2	20870812	20871402	6	6.2E-07	GDF7	island
44	2	64868515	64868711	2	1.9E-03	SERTAD2	
45	2	242763794	242763983	2	8.7E-03	NEU4	shore
46	2	236506413	236506614	2	2.3E-02	AGAP1	island
47	2	20870087	20870363	3	1.1E-04	GDF7	island
48	20	26190328	26190355	3	1.9E-02	MIR663AHG	island
49	22	38092643	38093208	11	4.7E-03	TRIOBP	
50	3	39321710	39322104	3	3.6E-02	CX3CR1	
51	3	116163694	116164243	6	6.1E-04	LSAMP	
52	4	74847646	74848017	8	1.2E-04	PF4	island
53	4	7657070	7657709	5	6.7E-03	SORCS2	
54	4	174429263	174429542	6	4.0E-02	SCRG1	shore
55	4	46995203	46995744	7	1.9E-03	GABRA4	island
56	5	157079312	157079521	5	2.1E-02	SOX30	island
57	5	140719008	140719303	4	3.1E-03	PCDHGA2;PCDHGA1	island
58	5	1962311	1962555	3	9.1E-04	CTD-2194D22.4	
59	6	31938984	31939547	12	5.5E-04	DXO;STK19	shore
60	6	3849190	3849703	21	7.0E-04	FAM50B	island
61	6	30038791	30039901	39	3.2E-11	RNF39	island
62	6	529098	529342	3	1.5E-02	EXOC2	
63	6	126080132	126080724	4	3.9E-04	HEY2	island
64	7	158905317	158905537	3	1.9E-02	VIPR2	island
65	7	56515510	56516256	11	4.7E-02	LOC650226	island
66	8	599963	600556	6	2.8E-06	ERICH1	island
67	8	82644373	82644769	8	3.3E-02	CHMP4C	island
68	8	26722496	26722966	5	2.1E-02	ADRA1A	island
69	9	136149908	136150033	3	3.1E-03	ABO	island

[a] DMRs were annotated to nearest gene from human genome version hg19 with *CruzDB* package (Meyer et al., 2012; Pedersen et al., 2013).

142 CpG sites) with ambient $PM_{2.5}$ absorbance exposure, and 15 DMRs (20 genes and 125 CpG sites) with ambient exposure to UFP (Table S4.5 and supplementary excel-file Tables E1, E2). Overlap of annotated genes across significant DMRs was observed between personal $PM_{2.5}$ and personal $PM_{2.5}$ absorbance (n=11), and between personal $PM_{2.5}$ absorbance and ambient $PM_{2.5}$ absorbance (n=10) while for the other combinations overlap was less pronounced (Table S4.5). Two of the significant CpGs identified in our univariate analysis (mapped to *KNDC1* and *FAM50B*) were located within the DMRs that were associated with $PM_{2.5}$.

Overlap with CpG sites associated with cigarette smoking

Cigarette smoke is a source of particulate exposure with some similar characteristics to ambient air pollution (but contributing to a much higher exposure). To assess the overlap between our results and a set of CpG sites previously reported to be associated to air pollution (Joehanes et al., 2016), we conducted an approach comparable to gene set enrichment (Subramanian et al., 2005). Hypermethylated CpG sites that were associated with cigarette smoking were significantly enriched (p-value < 0.01; Figure S4.4A) among the most strongly hypermethylated CpG sites in our analysis for $PM_{2.5}$. A similar enrichment was observed for smoking-related hypomethylated CpG sites in our analysis ($p < 0.01$; Figure S4.4B). Further description of these results is provided in the supplemental materials S5.2.

Impact of air pollution on immune markers

We identified a significant association between ambient exposure to $PM_{2.5}$ absorbance and serum concentration of CCL22 (β = 0.11; FDR=0.03; Figure S4.5A/B). The other pollutants yielded no significant association with any of the immune markers. Some indications of an association (FDR < 0.02) were observed between personal exposure measurements of UFP and serum concentration of two additional immune markers [CXCL10 (β =0.16; FDR=0.15) and G.CSF (β = 0.32; FDR =0.15)], which are nevertheless noteworthy as comparable effects were observed for these markers in an independent population (PISCINA) that was part of the EXPOsOMICS project (Figure S4.5A). The observed associations in that study were in the same direction as in our main study and significant (at FDR< 0.2) for G.CSF (β=0.25; FDR=0.17) but not for CXCL10 (β=0.34; FDR= 0.28). The observed associations for CCL22, G.CSF, and CXCL10 were robust in a series of sensitivity analyses (Figure S4.6). Further description of these results is provided in the supplemental materials S3.3.

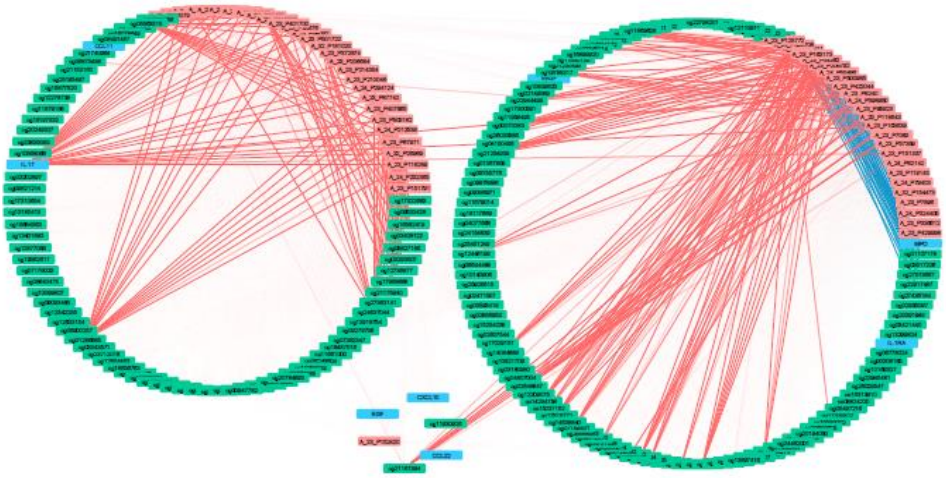

Figure 4.4: Illustration of the interrelations between the three omics platform: Methylation (CpG sites significantly associated with personal $PM_{2.5}$ (both in DMR (69 CpGs) and univariate analysis (13 CpGs)); Green rectangle), immune markers (Blue rectangle) and gene expression (expression of mapped genes; Red rectangle). Three rings show different clusters (Cluster1 contained 134 omics markers and cluster 2 and 3, 100 and 6 omics markers, respectively. The connecting lines show association (correlation) between different markers (with absolutes range between 0.4, 0.73) Positive in Red; negative in Blue). The color intensity and the thickness of the lines both indicate the strength of the association (thicker lines indicate a correlation higher than 0.6).

Functional readout of the impact of differential methylation at the identified CpG sites

Figure 4.4 shows the interrelations between CpG sites significantly associated to personal $PM_{2.5}$ (both in DMR (69 CpGs) and univariate analysis (13 CpGs)), expression of mapped genes, and the immune markers based on a similarity matrix generated using partial least square regression (González et al., 2012). Though modest in absolute correlation (absolute range of correlation was between 0.4 and 0.73), Figure 4.4 illustrates a high degree of interrelatedness between the selected set of markers: 240 markers including 177 CpG sites (mapped to 56 unique genes), 55 expression of mapped genes (annotated to 41 unique genes), and 8 immune markers. We identified two clusters of considerably correlated markers and a very small weakly interrelated third cluster, which were primarily driven by the direction of the correlation. Cluster one and two contained 134 and 100 markers (nodes), respectively. Further information about clusters and interrelatedness is provided in Excel-file Table E4.

Discussion

Our study provides evidence that measurements of personal exposure to particulate air pollution, collected in a panel of healthy participants from four European countries, were associated with DNA methylation changes in specific regions of the genome. Correlation between methylation at these CpG sites and a set of immune markers measured in peripheral blood was generally moderate, as was the correlation with the expression of genes mapped to the identified CpG sites. Several immune markers were independently associated with some of the air pollution measurements. Our results do not replicate findings from previous studies of short-term air pollution and blood DNA methylation (Bellavia et al., 2013; Panni et al., 2016; Wang et al., 2017), though no consistent patterns have been observed in these studies. As we observed relatively low ICCs for the CpG sites that were associated to the $PM_{2.5}$ measurements, our study points towards (partly reversible) short-term variation of DNA methylation in peripheral blood in response to daily variation in $PM_{2.5}$ levels. Due to the ubiquitous presence of air pollution in our environment, studying the temporal behavior of the DNA methylome in response to exposure to air pollution in observational studies is a challenge.

A handful of experimental studies among both animals and humans do indicate a short-term response to exposure to air pollution. In one experimental study, demethylation of the interleukin-2 gene of lymph node T cells was observed in vivo 20 minutes after antigen presentation (Bruniquel and Schwartz, 2003). Furthermore, in a bioassay conducted by (Ding et al., 2017) methylation of LINE-1, iNOS, p16CDKN2A, and APC, in both lung tissue as well as in peripheral blood, changed after four hours of exposure to $PM_{2.5}$. Bellavia et al. (2013) observed global DNA demethylation after 130 minutes of exposure to fine concentrated ambient particles. Chen et al. (2016) observed significant reduction of both global and site-specific methylation after 48 hours of exposure to $PM_{2.5}$. Jiang et al. (2014) reported significant changes in DNA methylation 6 hours after exposure to diesel exhaust.

In addition to experimental evidence, several observational studies identified significant associations between short-term changes in methylation and exposure to air pollution. Panni et al. (2016) observed demethylation to be associated with average air pollution levels in the two days before blood sampling (though a stronger signal was observed with longer trailing averages, while Wang et al. (2017) observed demethylation within 12 hours of exposure to $PM_{2.5}$.

Combined, results from our and other studies suggest that the impact of air pollution on the blood methylome is likely subtle and requires high-quality exposure assessment and considerable statistical power to be studied.

Strengths of our study include the application of a common study protocol and standardized operating procedures applied across four European countries. DNA methylation, gene expression, and immune markers were each analyzed in a single laboratory and

blinded duplicate samples were included for quality assurance. State-of-the-art personal exposure measurements of air pollution were conducted to reduce the impact of measurement error and we repeated measurements within a time period of a year to cover seasonal variation. High-quality exposure assessment is resource intensive. Our study was, therefore, necessarily modest in the number of individuals that could be included in the study, which is a limitation. Another limitation of our study was the fact that we were unable to identify an independent dataset in which we could replicate the methylation findings of this study. However, we did assess the robustness of the identified signals within our dataset as much as possible.

In addition to univariate regression analysis, we applied two alternative approaches (gene set enrichment and DMR) that maximized the information we could derive from our study. Both approaches overcome statistical limitations in the assessment of perturbations by combining information from multiple CpG sites rather than assessing sites one by one. Enrichment of smoking-related CpG sites among the CpG sites most strongly associated with air pollution points towards a shared biologic pathway of the effects of cigarette smoke and air pollution on the epigenome. This is of interest due to overlap between health outcomes that have been related to tobacco smoking and air pollution (Wang et al., 2015). We recently discovered a similar overlap in an analysis of another air pollutant (NO_X) and the peripheral blood transcriptome(Mostafavi et al., 2017). We focused on DMRs in our analysis because CpG sites mostly function in groups to regulate gene expression, rather than independently (Breton et al., 2017). Similar to gene set enrichment, our DMR analysis combined information from multiple CpG sites rather than assessing sites one by one and therefore also increased the statistical power to detect subtle perturbations in methylation patterns, leading to a much richer signal in our DMR analysis than in our univariate regression. Two CpG sites that were independently significantly associated to $PM_{2.5}$ were located within a DMR, which increased the credibility of these two associations.

Acknowledging that the 450K array chip was designed to interrogate the methylation status of CpG sites in proximity to genes, a potential approach to gain insight into the biological role of DMRs and single CpG sites associated with air pollution involves mapping them to nearby genes and conducting functional enrichment analysis on these genes (Wright et al., 2016). Functional analysis of the identified CpG sites at the mapped genes using the NIH-DAVID bioinformatics resources (Huang et al., 2009) did not provide any insight into potentially affected biologic pathways, likely due to the limited number of CpG sites found in this study. Further functional interpretation of the identified CpG sites requires moving towards regulatory enrichment analysis, incorporating not only mapping of nearby genes, but also acknowledging trans-acting effects of DNA methylation on gene expression (Wright et al., 2016). In this study, we observe some evidence for such complex interactions between DNA methylation, gene expression, and immune markers by showing the correlations between these markers. The existence of three correlation

clusters could point toward shared biological function, though further analyses are required to confirm this interpretation.

24-hour personal exposure measurements of $PM_{2.5}$ were more strongly associated with DNA-methylation than long-term modeled exposure estimates of $PM_{2.5}$. Plusquin et al. (2017) using a much larger study among healthy adults, reported a similar absence of significant association of DNA methylation in response to long-term $PM_{2.5}$, NO_X, NO_2 (Plusquin et al., 2017). These observations could be an indication that acute changes in methylation due to exposure to air pollution identified in this study did not reflect a long-term association between air pollution and DNA methylation. However, long-term modeled exposure estimates could also be much more affected by measurement error which would have reduced the statistical power to detect significant associations.

24-hour personal exposure measurements of $PM_{2.5}$ were also more strongly associated with DNA methylation than ambient measurements of $PM_{2.5}$ and measurements of the other air pollutants. Measurements of ambient $PM_{2.5}$ and $PM_{2.5}$ absorbance were likely more strongly affected by measurement error. The weaker effect for personal $PM_{2.5}$ absorbance is surprising as $PM_{2.5}$ absorbance has been reported to be a very good marker for exposure to traffic-related particles (Cyrys et al., 2003). Direct comparison of the results for $PM_{2.5}$ with those for personal UFP measurements is complex as these pollutants not only differ in particle size but also in the way they were assessed (mass weight versus counts). Differences in observed associations might, therefore, be due to measurement error but might also have a biological interpretation related to particle size. Further work is required to clarify this issue.

The optimal way to summarize continuous UFP measurements in an aggregated metric is still an ongoing discussion (Nunen et al., 2017). Exposure to UFP appears to be primarily determined by peaks and therefore summarizing exposure into a mean or median might not optimally reflect the impact of 24-hour exposure to UFP on the human system (Nunen et al., 2017). The collected continuous UFP measurements provide opportunities for further explorations of sensitive time windows and the contribution of outdoor versus indoor sources to the total UFP exposure received (Nunen et al., 2017).

In this study, we assessed exposure in the 24-hour before blood draw. Limited information is available on the time-scale on which environmental factors are expected to have an impact on changes in DNA methylation. Therefore, the time-period over which exposure should be assessed to maximize the power to assess potential associations is not known. In a study by (Wang et al., 2016, 2017), personal monitoring of $PM_{2.5}$ was conducted 3 days before scheduled blood draw using a microPEM device. Based on this data different time lags were calculated, of which the 0-24h time lag was most closely related to differential methylation. This result provides some support for the time window included in our study.

Even though we did not observe any associations between exposure to air pollution and estimated cell type composition, we repeated the main analysis for $PM_{2.5}$ correcting for

the estimated cell type composition by including seven cell types (”Monocytes”, ”B”, ”CD4T”, ”NK”, ”CD8T”, ”Eosinophils”, ”Neutrophils”). Results did not change substantially (Table S4.6) and were strongly correlated (r = 0.92) with the effect estimates from our original model.

Our study also provides evidence for perturbation of serum concentration of three immune markers (CCL22, CXCL10, and G.CSF) in association with air pollution (CCL22 with ambient $PM_{2.5}$ absorbance; CXCL10 and G.CSF with personal UFP). So far, these immune markers have not been reported in association with air pollution markers in other studies. However, previously published reports on the association between short-term air pollutants and systemic inflammation in human study populations were inconsistent (Hassanvand et al., 2017; Larsson et al., 2013; Steenhof et al., 2013).

Conclusion

In conclusion, this study provide evidence for an association between 24-hour exposure to air pollution, in particular measurements of $PM_{2.5}$, and DNA methylation both at single CpG sites and DMRs. Analysis of DMRs provides a promising avenue to further explore the subtle impact of environmental exposures on DNA methylation.

Supplementary material

S1. Long-term modeled exposure to air pollution

We generated estimated annual average ambient concentrations of air pollution for each individual using the land-use-regression (LUR) prediction models as a proxy for long-term modeled ambient exposure. To estimate $PM_{2.5}$, we used models which provided in the ESCAPE study (European study of cohorts for air pollution effects) (Eeftens et al., 2012b) and to estimate UFP we used models which provided in EXPOsOMICS study (Nunen et al., 2017). LUR models were developed and evaluated for each study area based on repeated exposure measurements obtained from monitoring sites. In each area, 160240 sites were monitored to develop LUR models by supervised stepwise selection of Geographic Information System (GIS)-derived predictor variables (e.g., traffic intensity, population, and land-use).

S2. Gene expression assessment

The normalization of the gene expression was performed using the Bioconductor (github.com/BiGCAT- M/arrayQC_Module), followed by a selection of genes with less than 30% flagged bad spots and imputation by k-nearest neighbors (k-NN; k-value). The final number of probes was 24,893 transcripts. The study design includes two quality control samples per batch of microarray hybridization. These quality control samples were identical biological samples that were distributed across batches, and enable correction of batch-related influences on gene expression levels. The pre-processing of quality control samples was performed together with the study samples except for the imputation step which was omitted in the case of quality control samples. A ratio was calculated by subtracting (since expression values were log2 transformed) the median expression of the quality control samples of one day from the median expression of all quality control samples of the study. The study samples were corrected by the addition of these ratios to the samples belonging to the same batch. All laboratory procedures were carried out at the Maastricht University according to manufacturer's protocols.
We identified 148 probe-level of transcripts that were in annotated to the genes (n=152) which identified in methylation analysis significantly associated with different air pollution markers (both in univariate analysis and in DMR analysis). Of these transcripts, we had availability of 58 transcripts in our gene expression data.

S3. Impact of air pollution on immune markers

S3.1. Immune markers assessment

Quality control samples were run in duplicate with the study samples in each batch. All laboratory procedures were carried out at the institute for risk assessment science Utrecht University according to manufacturer's protocols.
For 11 immune markers (interleukin IL-1β, IL-4, IL-5, IL-6, IL-10, IL-13, IL-25, TNF-α, MIP-1β, GRO, and FGF-β) serum concentration was below the limit of quantification (LOQ) in more than 70% of the samples. Therefore, they were excluded from further analyses. The remaining markers were imputed using a maximum likelihood estimation procedure (Lubin et al., 2004). To allow for a plate to plate variation, we imputed based on each plate-specific LOQ and included plate as a predictor variable in the imputation model. The maximum percentage of imputed samples was 35% (G-CSF) and 27% (eotaxin), while other markers had less than 4% imputed values.

S3.2. Impact of air pollution on immune markers (Methods)

We applied the univariate mixed-effects models to investigate the association between immune markers and exposure to personal/ambient levels of air pollution. We included random intercepts for each subject to capture the correlation among measurements within the same subject and for microtiter plate (N=9) to capture nuisance variation generated in the assessment of the immune markers in the model. The same set of covariates as we defined in our primary univariate methylation models were used to control for potential confounding variables.
Moreover, to test replication of our findings from the association between air pollution and immune markers, we used an independent study (PISCINA) which is part of EXPOsOMICS project that used similar methods. In this study, 33 young healthy non-smoker adults (aged 18-40) have been enrolled (Vlaanderen et al., 2017). Personal exposure to air pollution ($PM_{2.5}$, $PM_{2.5}$ absorbance, and UFP) was measured 24-hour before blood sample collection following the same protocol as our study. Samples were collected once and blood samples were analyzed for immune markers (DNA methylation or gene expression data was not available in this study) in the same lab of EXPOsOMICS study. Baseline characteristics of participants in this study are presented in Table S4.1. We used the same univariate mixed-effect model as described above however excluding random intercepts for each subject.
In addition, a number of sensitivity analysis were performed to assess how sensitive our findings were to variations in the confounder model. First, We ran a minimally adjusted model (only age and sex, BMI included as covariates), and a model in which country was excluded from the primary set of covariates. Then, to explore whether all cities contribute to the parameter estimation similarly, we repeated our primary model while leave one

country out at the time as well as in the stratified data by country.

S3.3. Impact of air pollution on immune markers (Results)

Sensitivity analyses of the association between air pollution and immune markers indicated robust results to the set of confounders considered (Figure S4.6). In the minimally adjusted model, we observed the association between CCL22 and ambient level of $PM_{2.5}$ absorbance in the same direction as it was in the primary model, but it was no longer significant (0.07; FDR = 0.2). The significant association (0.13; FDR=0.01) between CCL22 and ambient level of $PM_{2.5}$ absorbance identified in our primary model was retained in the model in which country was excluded as covariates in the primary model. We show the effect size and direction of the association between three significant signals (CCL22, G.CSF, CXCL10) and different air pollution markers for main and different sensitivity analyses (Figure S4.6A). Results in the stratified model by countries were similar, though with larger standard errors, compared to the results of the analysis in pooled population. Association between $PM_{2.5}$ absorbance and CCL22 was borderline significant (0.25; FDR=0.05) in Switzerland with larger effect size and standard errors than in pooled population (Figure S4.6B). In general, we did not observe a large deviation of effect estimates of different air pollution markers when excluding data from each country one-by-one in comparison to the overall effect estimates in pooled population. When individuals from Italy were excluded from the analysis, a significant association between CCL22 and $PM_{2.5}$ absorbance (0.17; FDR= 0.001) that was identified in pooled population was retained (Figure S4.6C).

S4. Differentially methylated regions (DMR)

We identified DMRs from linear mixed-effect model P-values by applying $comb-p$ (Pedersen et al., 2012), with a window size of 300 bases and containing a minimum of 2 probes. The identification of DMRs is described in detailed inPedersen et al. (2012). Briefly, $comb-p$ takes as input unadjusted P-values for each probe from our primary univariate analysis and calculates an adjusted P-value for each probe that accounts for the local correlation and the P-values of neighboring probes within the set window (300 bases). Probe-level P values are then adjusted by using the Benjamini-Hochberg methodology, resulting in multiple testing-corrected P values for individual probes that are influenced by their neighboring CpG motifs. Next, comb-p finds regions or DMRs and calculates P-values for the regions based on correlation (SLK P value). Finally, the region P values are adjusted for multiple testing by using the Sidak correction based on the size of the region and number of possible regions of that size, such that larger regions undergo less stringent correction because there are fewer possible large regions (SLK-Sidak P value). DMRs were annotated with $CruzDB$ tool (Pedersen et al., 2013). Statistical analyses were performed using Python version 2.7.13.

S5. Overlap with CpG sites associated with cigarette smoking

S5.1. Methods

We assessed whether the patterns of CpG sites associated with air pollution in our data has similarities to pattern associated with cigarette smoking using the gene set enrichment methodology (GSEA software) explained by Subramanian et al. (2005). Our implementation of this method involved three steps. First, all CpGs in our univariate analysis based on direction of their association (positive or negative) with personal exposure to $PM_{2.5}$ were clustered. Second, CpGs were ranked within each (i) positive- and ii) negative-associations) clusters, based on the strength of the association (t-score). Last, the overlap between CpGs from our analyses and the CpG-sets related to "cigarette smoke" was assessed. In each cluster (positive and negative associated CpGs), we assessed whether the distribution of overlapping CpGs among all CpGs from our analysis is random, or whether this overlap occurs primarily among our most hypo- or hypermethylated CpGs. We considered 2,623 CpG sites in peripheral blood that were associated with smoking status in a study by Joehanes et al. (2016) as the reference set. Additionally, we used 1253 CpGs in 67 genes from different biological pathways relevant to air pollution responses based on a literature search for air pollution associated genes (Gruzieva et al., 2017; Mostafavi et al., 2017).

S5.2. Results

Figure S4.4 illustrates the outcome of enrichment analysis. The intensity of the blue and red color bars represents the absolute value of the t-score from univariate analysis (Blue: hypomethylation; red: hypermethylation). Each vertical line represents the overlap between CpGs included in our analysis and the CpG-set which we compared our results with. A qualitative indication for enrichment of the overlapping CpGs can be acquired by assessing the (lack of) symmetry of the graph, where asymmetry points towards enrichment. GSEA software provides a quantitative indication for enrichment based on the enrichment score which reflects the degree to which a CpG-set is overrepresented at the top or bottom of a ranked list of CpGs. The enrichment score is calculated by walking down the ranked list of CpGs, increasing a running-sum statistic (the heigth of the edges) when a CpG is in the CpG-set and decreasing it when it is not. The magnitude of the increment depends on the t-statistic. The enrichment score is the maximum deviation from zero encountered in walking the list. All CpGs that have been assessed before the maximum deviation is reached is part of the core enrichment set (the green edges). GSEA software uses the Kolmogorov-Smirnov statistic to calculate a p-value for the enrichment score.

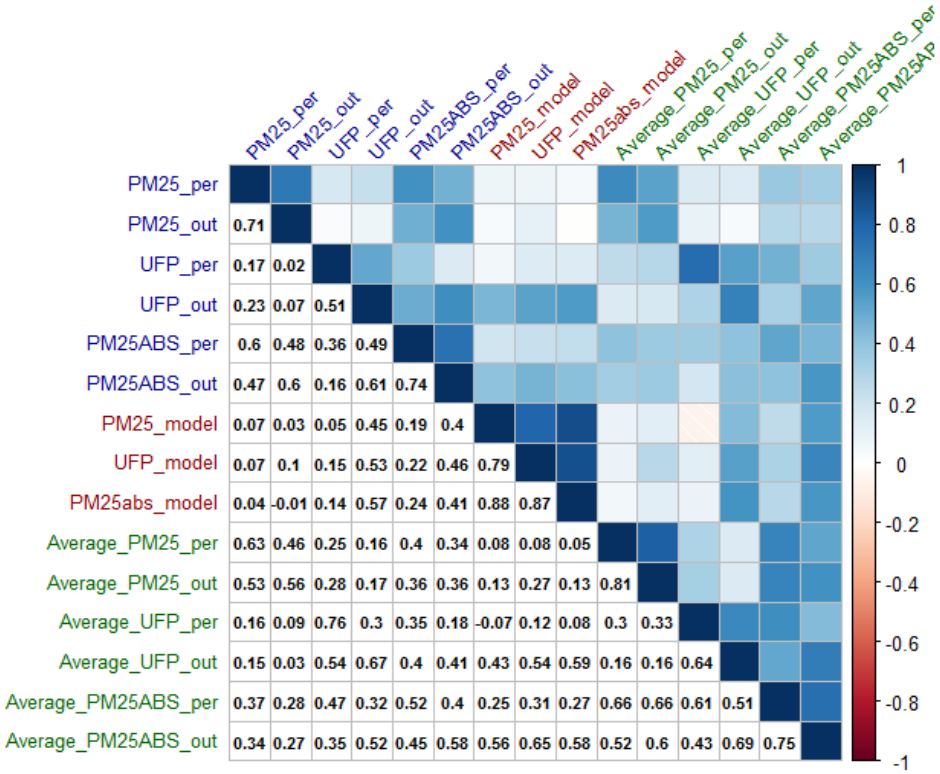

Figure S4.1: Correlation coefficients between air pollution measures. Three types of measures are included: 24-h measured, average of three 24-h measurements, long-term modeled. The color intensity of boxes indicates the magnitude of the correlation. Blue (red) color indicates the positive (negative) correlation. Numbers show the magnitude of correlation. Abbreviations are as follows: PM25_per (personal $PM_{2.5}$), PM25abs_per (personal $PM_{2.5}$ absorbance), UFP_per (personal UFP), PM25_out (ambient $PM_{2.5}$), and PM25abs_out (ambient $PM_{2.5}$ absorbance).

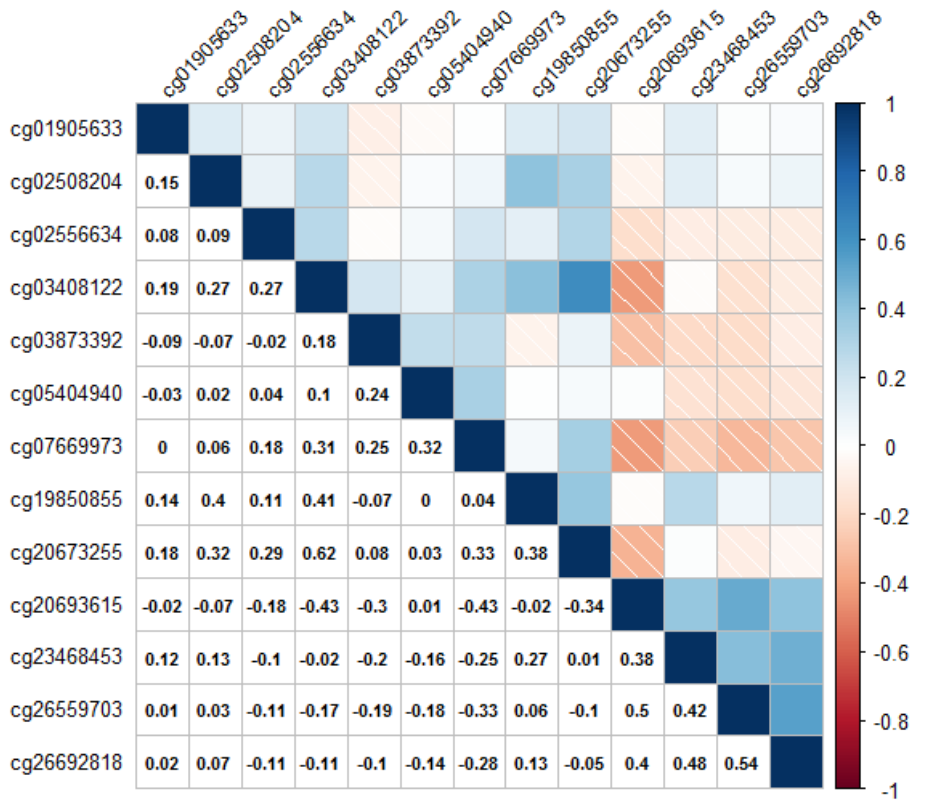

Figure S4.2: Pearson correlation coefficients between methylation at significant CpG sites identified in association with personal $PM_{2.5}$ exposure. The color intensity of boxes indicates the magnitude of the correlation. Blue (red) color indicates the positive (negative) correlation. Numbers show the magnitude of correlation.

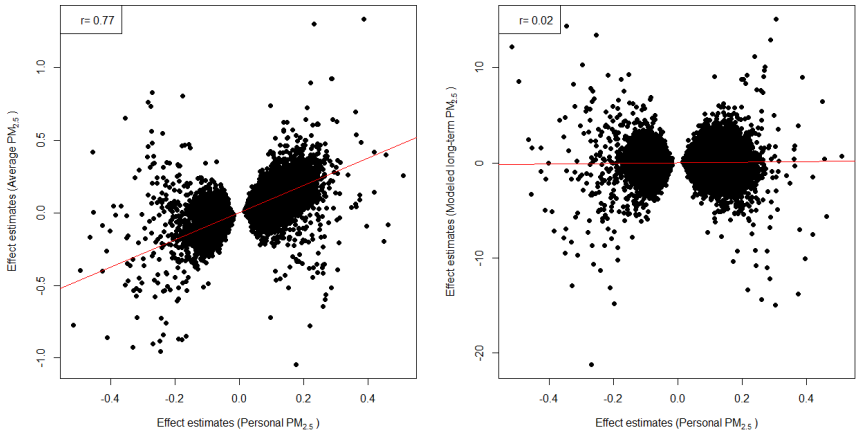

Figure S4.3: Scatterplot of significant effect estimates of personal $PM_{2.5}$ (at P-value < 0.05; 40420 CpGs) on methylation of CpG sites and corresponding effect estimates of Modeled and long-term $PM_{2.5}$. Regression line is indicated by Red color in each graph.

Table S4.1: Characteristics of the study participants in the Piscina study.

Characteristic	PISCINA[a]
Sex (N individual)	
Female	15
Male	18
Education (N individual)	
Secondary school	19
University	14
Age (years; median (P_{25}-P_{75}))	23 (21, 27)
BMI (Kg/m^2; median (P_{25}-P_{75}))	23 (21.5,25.62)
Physical activity [b] (Mets; SD)	1.45 (0.13)
Season (N samples)	
1: spring (21/3-20/6)	8
2: summer (21/6-20/9)	8
3: autumn (21/9-20/12)	17

[a] A small-scale study from Spain (the PISCINA study) served as an independent replication study for the immune markers data (DNA-methylation data not available).

[b] Physical activity was measured using the accelerometer (Actigraph GT3X+) and expressed in the 'The Metabolic Equivalent of tasks' (Mets).

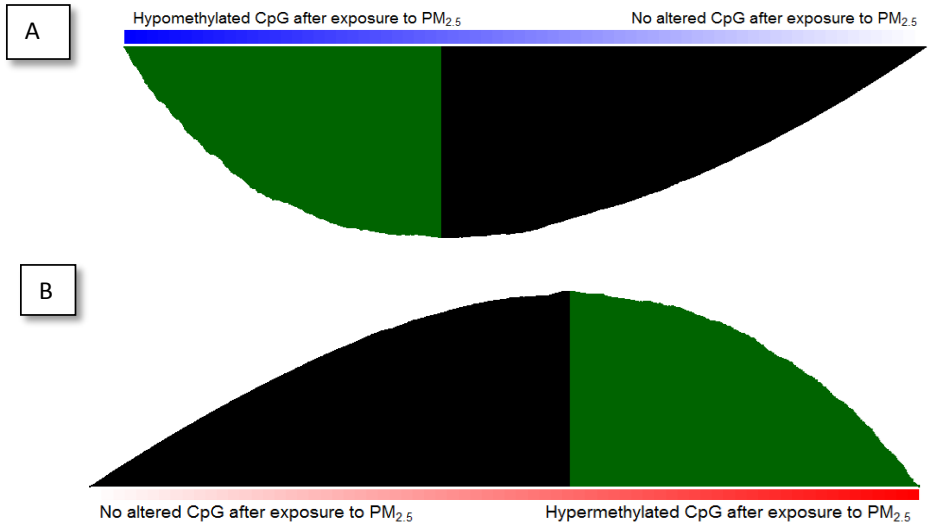

Figure S4.4: A) Distribution of hypo-methylated CpG sites in association (p-value < 0.01) with cigarette smoke vs. hypo-methylated CpGs ranked according to their t-statistics in association with personal $PM_{2.5}$. B) Distribution of hyper-methylated CpGs in association (p-value <0.01) with smoking exposure vs. hyper-methylated CpGs ranked according to their t-statistics in association with personal $PM_{2.5}$. The bottom color bar represents the alteration in methylation after exposure to $PM_{2.5}$ (Blue: hypo-methylation, red: hyper-methylation). Each vertical line represents one of the CpGs which is associated with cigarette smoke exposure. The height of this bar represents the running enrichment score. Green vertical lines represent the leading edge CpG responsible for the core enrichment of each CpG set.

Table S4.2: Effect of personal air pollution on blood cell type composition.

	$PM_{2.5}$[a]			$PM_{2.5}$ absorbance			UFP[b]		
Cell-type	β[c]	P-value	FDR[d]	β	P-value	FDR	β	P-value	FDR
Monocytes	0.003	0.30	0.75	0.003	0.25	0.89	0.001	0.73	0.91
B	-0.001	0.45	0.75	-0.001	0.55	0.89	0.002	0.52	0.91
CD4T	0.003	0.57	0.75	0.006	0.36	0.89	0.001	0.91	0.91
NK	0.006	0.10	0.70	0.001	0.81	0.89	0.004	0.49	0.91
CD8T	-0.003	0.67	0.75	-0.01	0.42	0.89	-0.005	0.66	0.91
Eosinophils	0.0002	0.75	0.75	0.0001	0.89	0.89	-0.0003	0.84	0.91
Neutrophils	-0.004	0.64	0.75	-0.002	0.85	0.89	-0.01	0.74	0.91

[a] Personal exposure to particulate matter (PM)

[b] Personal exposure to Ultra-fine particle

[c] Effect estimate per unit changes of the exposure

[d] False discovery rate correction for p-value

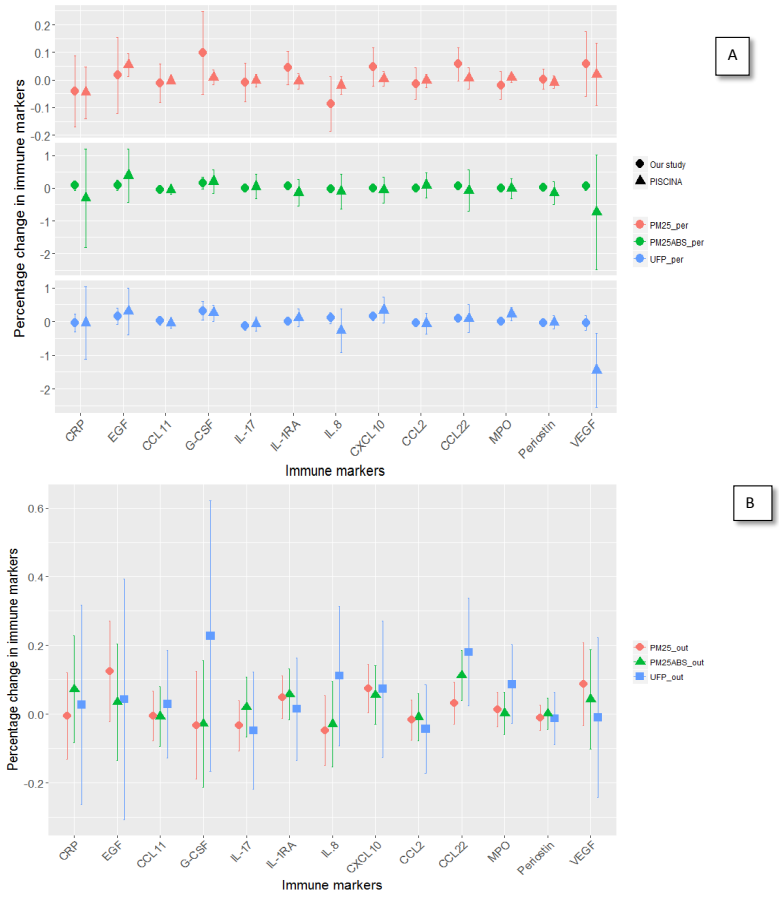

Figure S4.5: (A) Association between immune markers and personal air pollution measurements in our study population and in the PISCINA. (B) Association between immune markers and ambient air pollution measurements in our study population. Abbreviations: PM25_per (personal $PM_{2.5}$), PM25abs_per (personal $PM_{2.5}$ absorbance), UFP_per (personal UFP), PM25_out (ambient $PM_{2.5}$), and PM25abs_out (ambient $PM_{2.5}$ absorbance).

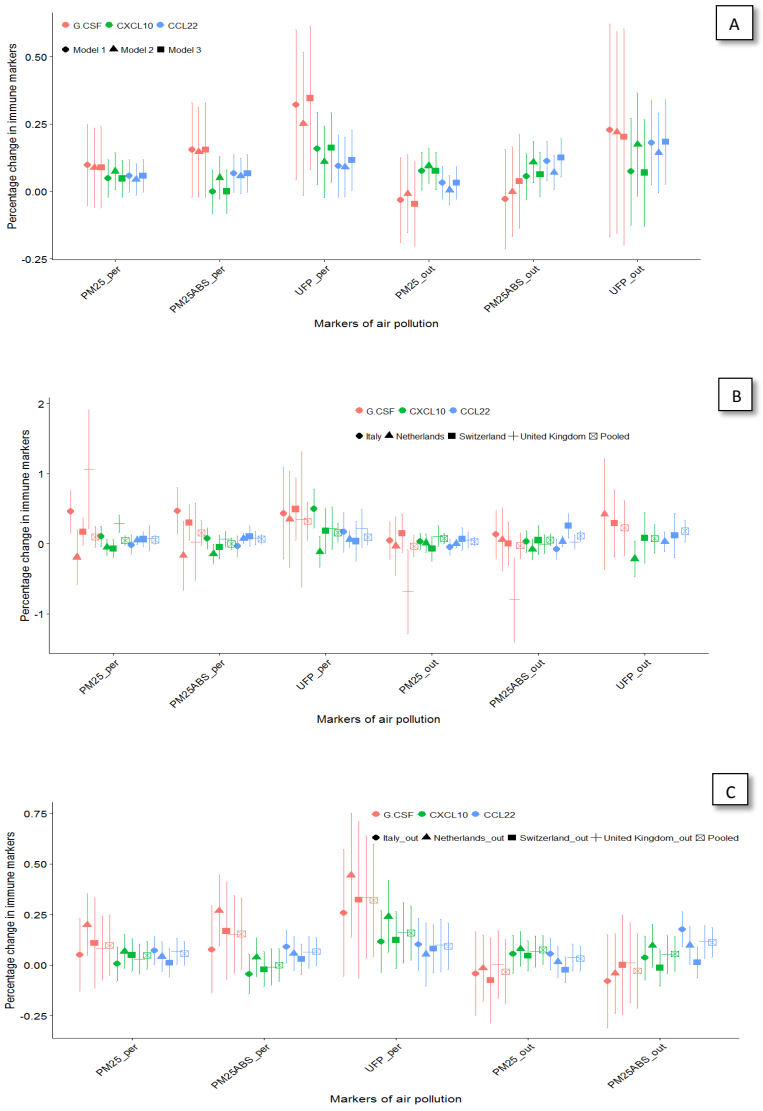

Figure S4.6: (A) Association between significant immune markers and different air pollution markers for the different set of confounder adjustment including Model 1 (full model) adjusted for sex, age, BMI, education, season, physical activity, temperature, relative humidity, and country. Model 2 adjusted for sex, age, BMI, and country. Model 3 adjusted for all variables of Model 1 except for country. (B) Association between significant immune markers and different air pollution markers in the pooled population and stratified by country. (C) Association between significant immune markers and different air pollution markers in the pooled population and after leaving one of the countries out each time. Abbreviations:PM25_per (personal $PM_{2.5}$), PM25abs_per (personal $PM_{2.5}$ absorbance), UFP_per (personal UFP), PM25_out (ambient $PM_{2.5}$), and PM25abs_out (ambient $PM_{2.5}$ absorbance).

Table S4.3: Results for 20 top-ranked (based on FDR) associations between CpGs and each of the air pollution markers ($PM_{2.5}$, $PM_{2.5}$ absorbance, UFP) from the univariate analysis of methylation-air pollution

Personal exposure to air pollution								
$PM_{2.5}$			$PM_{2.5}$ absorbance			UFP		
Probe ID	β^a	FDR[b]	Probe ID	β	FDR	Probe ID	β	FDR
cg26692818	0.23	0.02	cg09074113	-0.14	0.17	cg25938287	-0.42	0.78
cg02556634	0.06	0.02	cg18593317	-0.17	0.17	cg00000029	0.05	1.00
cg03873392	-0.13	0.03	cg24860812	-0.27	0.17	cg00000108	-0.07	1.00
cg07669973	-0.12	0.03	cg11051221	-0.29	0.24	cg00000109	0.00	1.00
cg02508204	0.20	0.04	cg10541674	0.14	0.28	cg00000165	0.11	1.00
cg19850855	0.20	0.04	cg26006214	-0.17	0.28	cg00000236	0.07	1.00
cg20673255	0.14	0.04	cg14355748	0.26	0.87	cg00000289	0.01	1.00
cg23468453	0.23	0.04	cg00000029	0.03	1.00	cg00000292	0.09	1.00
cg26559703	0.24	0.04	cg00000108	-0.01	1.00	cg00000363	-0.01	1.00
cg01905633	0.06	0.04	cg00000109	-0.03	1.00	cg00000622	-0.02	1.00
cg03408122	0.13	0.04	cg00000165	0.04	1.00	cg00000658	0.03	1.00
cg05404940	-0.11	0.04	cg00000236	0.00	1.00	cg00000714	-0.04	1.00
cg20693615	0.23	0.04	cg00000289	-0.06	1.00	cg00000721	0.11	1.00
cg13051197	0.24	0.06	cg00000292	0.08	1.00	cg00000734	-0.01	1.00
cg14218481	-0.17	0.07	cg00000363	0.03	1.00	cg00000769	-0.06	1.00
cg04817870	0.12	0.08	cg00000622	0.01	1.00	cg00000807	0.01	1.00
cg06804625	-0.11	0.08	cg00000658	0.01	1.00	cg00000884	0.12	1.00
cg12166520	0.18	0.08	cg00000714	0.00	1.00	cg00000905	-0.06	1.00
cg12961010	0.18	0.08	cg00000721	0.05	1.00	cg00000924	-0.01	1.00
cg15420468	-0.08	0.08	cg00000734	-0.01	1.00	cg00000948	0.01	1.00
Ambient exposure to air pollution								
$PM_{2.5}$			$PM_{2.5}$ absorbance			UFP		
Probe ID	β^a	FDR[b]	Probe ID	β	FDR	Probe ID	β	FDR
cg08086751	0.15	0.28	cg15782852	0.21	0.87	cg02180579	-0.75	0.51
cg22039287	0.19	0.28	cg00000029	-0.01	1.00	cg10435092	0.36	0.51
cg12462224	0.13	0.64	cg00000108	-0.01	1.00	cg13451937	-0.55	0.60
cg12721603	0.15	0.72	cg00000109	-0.01	1.00	cg19117501	-0.55	0.60
cg00000029	0.02	1.00	cg00000165	0.02	1.00	cg26922690	0.36	0.60
cg00000108	-0.03	1.00	cg00000236	-0.01	1.00	cg07079489	-0.42	0.61
cg00000109	-0.01	1.00	cg00000289	-0.05	1.00	cg00000029	0.00	1.00
cg00000165	0.03	1.00	cg00000292	0.04	1.00	cg00000108	-0.12	1.00
cg00000236	0.04	1.00	cg00000363	0.03	1.00	cg00000109	-0.16	1.00
cg00000289	-0.03	1.00	cg00000622	-0.04	1.00	cg00000165	-0.04	1.00
cg00000292	0.02	1.00	cg00000658	0.04	1.00	cg00000236	0.06	1.00
cg00000363	0.04	1.00	cg00000714	0.01	1.00	cg00000289	-0.04	1.00
cg00000622	-0.01	1.00	cg00000721	-0.05	1.00	cg00000292	0.06	1.00
cg00000658	0.03	1.00	cg00000734	-0.01	1.00	cg00000363	0.01	1.00
cg00000714	0.00	1.00	cg00000769	0.01	1.00	cg00000622	-0.03	1.00
cg00000721	-0.04	1.00	cg00000807	0.01	1.00	cg00000658	0.04	1.00
cg00000734	-0.01	1.00	cg00000884	0.03	1.00	cg00000714	0.01	1.00
cg00000769	0.02	1.00	cg00000905	-0.01	1.00	cg00000721	0.05	1.00
cg00000807	-0.02	1.00	cg00000924	-0.01	1.00	cg00000734	-0.04	1.00
cg00000884	0.04	1.00	cg00000948	0.10	1.00	cg00000769	0.01	1.00

[a] Effect estimate per unit changes of the exposure

[b] False discovery rate correction for p-value

Table S4.4: Effect estimates of personal and ambient exposure to $PM_{2.5}$, $PM_{2.5}$ absorbance, UFP for CpG sites significantly associated with personal $PM_{2.5}$ (at FDR 5%) in univariate analysis

	Personal exposure to air pollution								
	$PM_{2.5}$			$PM_{2.5}$ absorbance			UFP		
Probe ID	β	P-value	FDR	β	P-value	FDR	β	P-value	FDR
cg26692818	0.23	4.1E-08	0.02	0.12	0.03	1.00	0.06	0.43	1.00
cg02556634	0.06	1.0E-07	0.02	0.03	0.06	1.00	0.01	0.68	1.00
cg03873392	-0.13	2.4E-07	0.03	-0.10	0.001	1.00	-0.02	0.74	1.00
cg07669973	-0.12	3.0E-07	0.03	-0.08	0.004	1.00	-0.05	0.26	1.00
cg02508204	0.20	6.8E-07	0.04	0.09	0.06	1.00	-0.08	0.33	1.00
cg19850855	0.20	7.8E-07	0.04	0.14	0.01	1.00	-0.05	0.54	1.00
cg20673255	0.14	4.7E-07	0.04	0.05	0.18	1.00	-0.02	0.66	1.00
cg23468453	0.23	7.1E-07	0.04	0.14	0.01	1.00	0.06	0.48	1.00
cg26559703	0.24	6.1E-07	0.04	0.17	0.00	1.00	-0.08	0.38	1.00
cg01905633	0.06	6.3E-07	0.04	0.03	0.08	1.00	-0.04	0.15	1.00
cg03408122	0.13	1.3E-06	0.04	0.05	0.16	1.00	0.04	0.46	1.00
cg05404940	-0.11	1.2E-06	0.04	-0.10	0.0004	1.00	-0.08	0.10	1.00
cg20693615	0.23	1.2E-06	0.04	0.18	0.002	1.00	0.06	0.49	1.00
	Ambient exposure to air pollution								
	$PM_{2.5}$			$PM_{2.5}$ absorbance			UFP		
Probe ID	β	P-value	FDR	β	P-value	FDR	β	P-value	FDR
cg26692818	0.07	0.16	1.00	0.10	0.07	1.00	-0.001	0.99	1.00
cg02556634	0.03	0.05	1.00	0.01	0.35	1.00	-0.01	0.68	1.00
cg03873392	-0.01	0.72	1.00	-0.03	0.31	1.00	-0.02	0.83	1.00
cg07669973	-0.06	0.01	1.00	-0.05	0.08	1.00	-0.04	0.53	1.00
cg02508204	0.09	0.04	1.00	0.07	0.20	1.00	-0.05	0.59	1.00
cg19850855	0.10	0.02	1.00	0.06	0.26	1.00	-0.04	0.72	1.00
cg20673255	0.05	0.08	1.00	0.02	0.63	1.00	-0.02	0.80	1.00
cg23468453	0.07	0.19	1.00	0.04	0.48	1.00	0.12	0.30	1.00
cg26559703	0.05	0.35	1.00	0.03	0.65	1.00	-0.12	0.30	1.00
cg01905633	0.004	0.75	1.00	-0.003	0.85	1.00	-0.04	0.22	1.00
cg03408122	0.02	0.56	1.00	-0.001	0.97	1.00	0.08	0.31	1.00
cg05404940	-0.05	0.05	1.00	-0.04	0.13	1.00	-0.07	0.32	1.00
cg20693615	0.08	0.13	1.00	0.05	0.41	1.00	-0.03	0.78	1.00

Table S4.5: Number of DMRs significantly associated with different air pollutants and overlap in significant DMRs for each bivariate combination of pollutants

			Overlap between annotated genes across significant DMRs					
	Sig DMRs	No. of probe	$PM_{2.5}$_per	$PM_{2.5}$ABS_per	UFP_per	$PM_{2.5}$_out	$PM_{2.5}$ABS_out	UFP_out
$PM_{2.5}$_per	69	404	74	11	0	0	2	0
$PM_{2.5}$ABS_per	42	266	-	42	1	0	10	4
UFP_per	16	109	-	-	21	0	2	4
$PM_{2.5}$_out	4	12	-	-	-	4	0	0
$PM_{2.5}$ABS_out	16	142	-	-	-	-	19	6
UFP_out	15	125	-	-	-	-	-	20

Abbreviations are: $PM_{2.5}$-per (personal $PM_{2.5}$), $PM_{2.5}$ ABS_per (personal $PM_{2.5}$ absorbance), UFP_per (personal UFP), $PM_{2.5}$_out (ambient $PM_{2.5}$), and $PM_{2.5}$ABS_out (ambient $PM_{2.5}$ absorbance)

Table S4.6: Results for 20 top-ranked (based on P-value) associations between CpGs and personal $PM_{2.5}$ air pollution from the univariate analysis of methylation-air pollution when adjusting for white blood cell composition and when blood cell was not adjusted

Ajusted for white blood cell				Not adjusted for white blood cell			
Probe ID	β	P-value	FDR	Probe ID	β	P-value	FDR
cg02556634	0.06	2.92E-07	0.06	cg26692818	0.23	4E-08	0.02
cg06804625	-0.10	4.12E-07	0.06	cg02556634	0.06	1E-07	0.02
cg20971998	-0.12	5.06E-07	0.06	cg03873392	-0.13	2.4E-07	0.03
cg00304884	-0.19	7.16E-07	0.06	cg07669973	-0.12	3E-07	0.03
cg26692818	0.20	8.30E-07	0.06	cg20673255	0.14	4.7E-07	0.04
cg07669973	-0.11	1.10E-06	0.07	cg26559703	0.24	6.1E-07	0.04
cg25649826	-0.09	1.54E-06	0.08	cg01905633	0.06	6.3E-07	0.04
cg05404940	-0.10	1.63E-06	0.08	cg02508204	0.2	6.8E-07	0.04
cg13464448	0.09	1.66E-06	0.08	cg23468453	0.23	7.1E-07	0.04
cg20673255	0.13	2.04E-06	0.09	cg19850855	0.2	7.8E-07	0.04
cg03873392	-0.11	3.48E-06	0.12	cg20693615	0.23	1.17E-06	0.04
cg14218481	-0.15	3.61E-06	0.12	cg05404940	-0.11	1.19E-06	0.04
cg21743649	-0.26	3.80E-06	0.12	cg03408122	0.13	1.27E-06	0.04
cg20693615	0.22	4.13E-06	0.13	cg13051197	0.24	1.91E-06	0.06
cg01905633	0.06	4.44E-06	0.13	cg14218481	-0.17	2.39E-06	0.07
cg04726374	-0.10	5.55E-06	0.15	cg06804625	-0.11	2.99E-06	0.08
cg00355656	-0.08	5.99E-06	0.15	cg19195724	0.22	3.18E-06	0.08
cg26559703	0.21	6.54E-06	0.16	cg15420468	-0.08	3.62E-06	0.08
cg19850855	0.17	6.90E-06	0.16	cg10690439	0.2	3.79E-06	0.08
cg02508204	0.18	7.73E-06	0.16	cg22906397	0.23	3.86E-06	0.08

CHAPTER 5

THE MEDIATING EFFECT OF IMMUNE MARKERS ON THE ASSOCIATION BETWEEN AMBIENT AIR POLLUTION AND ADULT-ONSET ASTHMA

This chapter is submitted as: Mostafavi N, Jeong A, Vlaanderen J, Imboden M, Vineis P, Jarvis D, Kogevinas M, Probst-Hensch N, Vermeulen R . The mediating effect of immune markers on the association between ambient air pollution and adult-onset asthma. *Scientific Reports*, 2018.

LONG-TERM exposure to ambient air pollution has been associated with asthma, yet the molecular mechanisms involved are not entirely understood. It has been postulated that inflammatory markers may play an important role in this association. We aim to investigate to what extent a set of immune markers mediate the association between air pollution and adult-onset asthma using case-control study nested in the prospective SAPALDIA cohort.

A panel of 13 immune markers in peripheral blood samples was collected from 140 adult cases (asthma diagnosis after the age of 16) and 199 controls. Biannual mean exposure to nitrogen oxide (NO_2), and ultra-fine particulates (UFP) was estimated based on land use regression models. Annual mean exposure to PM_{10} and $PM_{2.5}$ was estimated from the Swiss PolluMap dispersion model. We tested associations between air pollutants and inflammatory markers and adult-onset asthma using mixed-effects (logistic) regression models, adjusted for confounding variables. In order to evaluate a possible mediating effect of the full set of immune markers as a proxy of immune-modulation, we modelled the relationship between asthma and air pollution with a partial least square path (PLS-PM) model.

We observed a significant association of IL-1RA [OR 1.37; 95% CI (1.09, 1.73) per IQR (224.84 pg/mL) increase in natural-logarithm of IL-1RA] with adult-onset asthma. Univariate mixed-effect models did not yield any significant association between air pollution and inflammatory markers. Results of the PLS-PM analyses indicated that the combination of air pollutants (mainly driven by UFP) was associated with risk of adult-onset asthma (path coefficient= 0.5 (odds ratio= 1.7), P-value= 0.002) and with a cluster of immune-marker changes (i.e., immune-modulation) (path coefficient= 0.14, P-value=0.05). In turn Immune-modulation was associated with risk of adult-onset asthma (path coefficient (β) =0.62 (odds ratio= 1.9), P-value=0.0004). Mediation analyses indicated that 15% of the effect of air pollution on risk of adult-onset asthma was mediated through immune modulation (path coefficient (β) = 0.09, p-value= 0.09). This effect appeared to be stronger for allergic asthma (22%) and overweight subjects (27%).

Our results indicated that the immune system is important in asthma development and provides supportive evidence for a mediating effect of the immune system in the association between air pollution (in particular UFP) and adult-onset asthma.

Introduction

Asthma is a chronic inflammatory disease of the airways which has a large impact on quality of life (Cai et al., 2017). Current research suggests a complex etiology for asthma (Wu et al., 2018), and there is emerging evidence that air pollution is one of the important environmental factors that may be involved in the pathogenesis of the disease (Gilmour et al., 2006). However, previous studies are inconsistent with respect to the role of air pollution in the development of asthma in adults (Guarnieri and Balmes, 2014; Jacquemin et al., 2015).

The mechanism by which air pollution may lead to asthma is an active research area (Ji et al., 2016). Exposure to air pollutants induces oxidative stress and activates inflammatory pathways (Kelly, 2003; Mostafavi et al., 2015). These pathways have also been implicated in asthma development (Wouters et al., 2009). Interleukins (IL), tumor necrosis factor (TNF), and pro-inflammatory cytokines have been suggested as key molecules involved in recognition of air pollutants and the subsequent inflammatory response (Vawda et al., 2013).

Several studies have reported on associations between long-term exposure to air pollution and chronic changes in immune markers (Hajat et al., 2015; Mostafavi et al., 2015; Pope III et al., 2016). In the ESCAPE study, living close to busy traffic was associated with increased C-reactive protein concentrations (Lanki et al., 2015). In another study long-term exposure to NO_x was reported to be associated with decreased levels of circulating IL-8, IL-10, IL-2, and TNF-α concentrations (Mostafavi et al., 2015). Human airway cells or bronchial epithelial cells have been shown to release anti-inflammatory cytokines and chemokines such as IL-10 or IL-8 upon incubation of ambient air particles (Alfaro-Moreno et al., 2009; Huang et al., 2003). Altogether, both observational and experimental studies have provided supportive evidence for the interaction between air pollutants and the immune-system in particular inflammatory pathways.

As air pollution has been linked to inflammatory responses and as these responses have been linked to adult asthma it would follow that the observed effect of air pollution on asthma is potentially mediated by the immune system. There is however a dearth of studies that has investigated this. As immune markers resembling inflammatory responses are pleiotropic it could be argued that such analyses should not be done on single immune markers but on a combined measure of the immune response. We therefore used partial least square path modeling (PLS-PM) to identify the relationship between adult-onset asthma and air pollutants and the possible mediating effect of immune markers simultaneously. PLS-PM enables us to combine information from multiple air pollution and immune markers by using a latent structure pattern for both the air pollutants (i.e. *air pollution*) and immune-markers (i.e. *immune-modulation*). We considered long-term exposure to multiple air pollution markers (particulate matter (PM) smaller than 2.5 μm ($PM_{2.5}$), and smaller than 10 μm (PM_{10}), nitrogen dioxide (NO_2), ultra-fine particulates

(UFP) [based on two metrics: particle number concentration (PNC) and lung deposited surface area (LDSA)]) and a panel of 13 immune markers in peripheral blood samples collected from 140 adult cases (asthma diagnosis after the age of 16) and 199 controls using a nested-case control design. Previous analysis in this cohort has shown an association between air pollutants and adult-onset asthma, in particular UFP (Jeong et al., 2018).

Materials and methods

Study population

SAPALDIA is a population-based study that was initiated in 1991 to investigate the long-term effect of air pollution on respiratory health in the Swiss adult population (1860 years, N = 9651). As previously described in detail (Ackermann-Liebrich et al., 2005; Martin et al., 1997), the study comprised eight distinct rural and urban areas covering the environmental and geographic diversity of Switzerland (Basel, Geneva, Davos, Aarau, Payerne, Montana, Wald, and Lugano). A subset of baseline participants completed two follow-up assessments in 2002-2003 (SAPALDIA2, n = 8047) and in 2010-2011 (SAPALDIA3, n = 6088). In both follow-up assessments blood was drawn and stored at $-80^{\circ}C$ in the SAPALDIA biobank. The present study included individuals selected among participants of SAPALDIA3. Information on age, sex and level of education were obtained from self-reported questionnaire data provided at study examinations. Smokers were excluded so that all cases and controls had not smoked for at least one year before blood was drawn at SAPALDIA2.
This study complies with the Declaration of Helsinki principles and all participants signed an informed consent. The study protocol of SAPALDIA was approved by the Swiss Academy of Medical Sciences and the regional committees for each study center.

Asthma definition

Asthma cases (N=203) were defined as participants who reported asthma in SAPALDIA3. In this study we limited our analyses to adult-onset asthma, which refers to participants who reported age of asthma onset 16 years or later (140 out of 203 cases). Controls (N=199) were participants who never reported the following since SAPALDIA1: self-reported asthma; physician-diagnosed asthma; asthma attack in the last 12 months; current asthma medication; wheezing without cold; three or more asthma-related symptoms in the last 12 months (symptoms considered: breathless while wheezing; woken up with a feeling of chest tightness; attack of shortness of breath after exercise; attack of shortness of breath while at rest; woken by attack of shortness of breath).
Participants were classified as atopic if they developed a wheal to one or more of eight common inhalant allergens (cat, timothy grass, parietaria, birch, house-dust mite, Al-

ternaria tenuis, Cladosporium herbarum, and dog) tested at baseline through a skin-prick test.

Exposure assessment

Biennial mean exposure to UFP was estimated based on multi-area land use regression (LUR) models derived from SAPALDIA specific-measurement campaigns conducted in 2011-2012 (Eeftens et al., 2016). Particle number concentration (PNC), lung deposited surface area (LDSA) was used as UFP metric and covered 4 out of 8 SAPALDIA study areas (covering 76 cases, 113 controls). NO_2 exposure was estimated by a newly developed European LUR model derived from measurements in 2010 (de Hoogh et al., 2016). We also obtained annual mean estimates of $PM_{2.5}$ and PM_{10} in 2010 from the PolluMap dispersion models developed by the Swiss agency for environment, forest, and landscape (Liu et al., 2007).

Assessment of immune markers

Serum levels of a panel of 23 immune markers were measured in blood samples collected from all subjects ($R\&D$ Systems, Abingdon, UK) (Vlaanderen et al., 2017). Luminex screening assay was performed according to the protocol described by the manufacturer. The panel included interleukin (IL) 1β, IL-4, IL-5, IL-6, IL-8, IL-10, IL-13, IL-17, IL-25, tumor necrosis factor alpha (TNF-α), C-C motif chemokine 11 (CCL11), IL1 receptor antagonist (IL1ra), CXC chemokine ligand 10 (CXCL10), epidermal growth factor (EGF), fibroblast growth factor beta (FGF-β), granulocyte colony-stimulating factor (G-CSF), melanoma growth stimulatory activity/growth-related oncogene (GRO), chemokine (C-C motif) ligand 2 (CCL2), C-C motif chemokine 22 (CCL22), macrophage inflammatory protein-1 beta (MIP-1β), vascular endothelial growth factor (VEGF), Myeloperoxidase (MPO), and periostin. Additionally, C-reactive protein (CRP) was assessed using $R\&D$ System Solid Phase Swedish ELISA. Quality control samples were run in duplicate together with the study samples in each batch.
For 10 immune markers (interleukin IL-1β, IL-4, IL-5, IL-6, IL-10, IL-13, IL-25, TNF-α, GRO, and FGF-β) the serum concentrations were below the limit of quantification (LOQ) in more than 70% of the samples. Therefore, they were excluded from analyses. The remaining markers were imputed using a maximum likelihood estimation procedure (Lubin et al., 2004).To allow for plate to plate variation in LOQ we imputed based on each plate-specific LOQ and included the plate as a predictor variable in the imputation model. The maximum percentage of imputed samples was 35% (G-CSF) and 27% (eotaxin), while other markers had less than 4% imputed values. Immune markers were natural log-transformed to normalize the distribution of the outcome and stabilize the variance.

Statistical analysis

Univariate analysis

The heterogeneity of demographic and lifestyle variables between asthma cases and controls were compared using chi-square test. We performed random-effects logistic regressions for all air pollution markers and all immune markers to assess the association with adult-onset asthma. Linear mixed-effects regression analysis was carried out to evaluate the relationship between immune markers and long-term exposure to air pollutants.
All analyses were adjusted for sex, age, body mass index (BMI) (kg/m^2) [only for air pollution and asthma analysis], education (primary, high school, university) as fixed-effects to control for potential confounding effects. To capture heterogeneity between the study areas, city (Basel, Wald, Davos, Lugano, Montana, Payerne, Aarau, and Geneva) was included as a random-intercept in all analysis.
In the immune marker analyses, we included a random-intercept for microtiter plate to capture nuisance variation generated in the assessment of the immune markers. Additionally, we adjusted for bench time, fasting time, Fourier-transformed venipuncture time point assuming one or two periods per day, and their multiplicative interaction terms with fasting time as potential confounders in immune marker analysis. As BMI could be on the causal path from air pollution to immune-modulation and asthma we did not adjust for BMI in the main analyses of immune markers. The set of confounders that were considered for each analysis in the current study is consistent with our previous publication (Jeong et al., 2018) using the same case-control dataset.
Statistical analyses were conducted using R version 3.4.0 (packages: *glmer* (De Boeck et al., 2011) and *lme4* (De Boeck et al., 2011). Multiple testing adjustments on the resulting p-values were performed using the false discovery rate method of Benjamini-Hochberg (BH-FDR) (Benjamini and Hochberg, 1995) at the level of 0.2.

Sensitivity analysis

To assess how sensitive our findings were to variations in the confounder model, we conducted a set of additional analysis. We ran a minimally adjusted model (only age and sex included as covariates), a model in which study area was excluded from the primary set of covariates, and a model in which we added season, and physical activity (insufficiently active and active) as covariates. For immune marker analysis, we ran a model in which BMI was included as a covariate and a model in which we excluded Fourier-transformed venipuncture time point from the list of covariates. In analysis of air pollution and asthma we ran a model in which BMI was not adjusted.
We examined effect modification of the associations by BMI (dichotomized into Overweight [BMI $\geq$ 25 kg/m^2; N=164], or Normal weight [18$\geq$ BMI $<$ 25; N=170]; skin prick test defined as allergic (positive test [78 cases]) or Non-allergic (negative test [62 cases]) in

a further sensitivity analysis, as differences in susceptibility to air pollution by BMI and difference in reaction to air pollution by atopic status has been suggested (Adam et al., 2015; Schikowski et al., 2013). Few individuals (N = 5) in our study were underweight (BMI $< 18\ kg/m^2$) and were therefore excluded from the stratified analysis by BMI.

Mediating effect of immune markers in association between air pollution and asthma

We analyzed the potential mediating effect of immune markers on the association between air pollution and adult-onset asthma by partial least-squares path modeling (PLS-PM) using the R package plspm (Sanchez, 2013). The technique relies on taking into account all relationships among unobserved or latent variables (LV), each measured by several observed indicators usually defined as observed or manifest variables (MV). PLS-PM models are formally defined by two sets of linear equations: structural (Inner) model and measurement (Outer) model. The structural model indicates the relationships among the latent variables *"immune-modulation"*, *"air pollution"*, and *"asthma"*, which are inferred from the observed immune markers, air pollutant markers, and adult-onset asthma in this study, respectively.
The measurement (outer) model indicates the relationship between the latent variables and their corresponding manifest variables. In our study, the manifest variables for the latent variable of air pollution are based on the six air pollution markers and observed values for the latent variable of *immune-modulation* are based on 13 immune markers. The algorithm includes two steps, the first step is an iterative estimation of latent construct scores, and the second step is the final estimation of coefficients of the PLS-PM. To estimate the coefficients of the PLS-PM in the final step, we analyzed the association between *asthma* and latent variables of *air pollution* and *immune-modulation* using logistic regression. To evaluate the good-ness of fit, we calculated the average coefficient of determination R^2 value for two latent variables in our analysis, *immune-modulation*, and *asthma* (pseudo-Nagelkerke R^2). The detailed methodology and algorithm can be found in previous publications (Petrarca et al., 2017; Tenenhaus et al., 2005). In order to determine the confidence intervals for the path coefficient of air pollution on *immune-modulation*, a bootstrap analysis was carried out using 200 sample data sets (Tenenhaus et al., 2005). Additionally, we tested the statistical significance of the indirect effect of air pollution on asthma (Helm et al., 2010) to establish the mediating effect of *immune-modulation*. PLS-PM does not allow for confounders to be considered in the modeling. To address this issue, we additionally employed a two-stage regression approach. First each exposure block (immune markers and air pollution) were separately regressed on potential confounders, and second, PLS-PM was fit on the obtained residuals.
We also performed sensitivity analyses. We tested the possible mediating effect of immune markers when only using UFP (both PNC and LDSA) to reflect the LV of air pollution. We

Table 5.1: Characteristics of study participants.

Characteristic	Control (N=199)	Cases (N= 140)	Differences P-value[a]
Sex (N sample (%))			
Male	96 (48)	53 (38)	
Female	103 (52)	87 (62)	0.074
BMI (Kg/m^2; median and P_{25}-P_{75})	24 (22.4, 27.2)	26 (23.1, 29.6)	0.002
Age (years; median and P_{25}-P_{75})	57.1 (48.7, 64.8)	59.5 (48.6, 67.9)	0.26
Education (N sample (%))			
Primary school	2 (1)	4 (3)	
Secondary/middle school	126 (63)	87 (62)	
College or university	71 (36)	49 (35)	0.44
Physical activity[b]			
Insufficiently active	40 (20)	38 (27)	
Sufficiently active	156 (80)	101 (72)	0.18
Season (N sample)			
1: spring (21/3-20/6)	52 (26)	33 (24)	
2: summer (21/6-20/9)	58 (29)	39 (28)	
3: autumn (21/9-20/12)	48 (24)	30 (22)	
4: winter (21/12-20/3)	41 (21)	37 (27)	0.63
Area (N sample (%))			
Basel	26 (13)	19 (14)	
Wald	56 (28)	19 (14)	
Davos	21 (11)	13 (9)	
Lugano	21 (11)	27 (19)	
Montana	21 (11)	19 (14)	
Payerne	12 (6)	5 (4)	
Aarau	27 (14)	20 (14)	
Geneva	15 (8)	18 (13)	0.02

Note: [a] P-value for difference was calculated using the chi-squared test for categorical baseline variables and the students t-test for continuous variables, [b] Sufficiently active: either moderate physical activity $\geq$ 150 min/week, vigorous physical activity $\geq$ 60 min/week, or combined duration (duration of moderate physical activity + 2 $\times$ duration of vigorous physical activity) $\geq$ 150 min/week; Insufficiently active: otherwise.

also tested the mediation effect of immune markers in the subset of allergic cases and overweight individuals.

Results

Our study population consists of 339 non-smoking adults (median age 58 years) from 8 areas in Switzerland of whom 140 participants developed asthma after the age of 16 years old (Table 5.1). There were no significant differences between cases and controls for any of the baseline variables investigated except BMI and study area. Median BMI for cases (26 Kg/m^2; [P_{25}-P_{75}]: [23.1, 29.6]) was significantly (p-value =0.002) higher than controls (24 Kg/m^2; [P_{25}-P_{75}]: [22.4, 27.2]). In Figure 5.1, we show the distribution characteristics of different air pollutants by study area. For all air pollutants, substantial variation was observed between (particularly PM_{10}, $PM_{2.5}$) as well as within (particularly NO_2, PNC, and LDSA) areas. UFP exposures (both PNC and LDSA) were only measured and estimated in 4 of the 8 areas (Basel, Geneva, Lugano, and Wald) and were available for a subset of

Table 5.2: Association between air pollution and adult-onset of asthma using random-effect logistic regression analysis

	Adjusted for BMI		Not adjusted for BMI	
Air pollution metric[a]	P-value	OR[b] [95% CI]	P-value	OR[b] [95% CI]
PM_{10}	0.83	1.17 (0.29, 4.68)	0.36	1.62 (0.58, 4.71)
$PM_{2.5}$	0.88	1.13 (0.24, 5.33)	0.41	1.63 (0.51, 5.35)
NO_2	0.37	1.37 (0.69, 2.76)	0.09	1.65 (0.94, 2.94)
PNC	0.001	3.23 (1.62, 6.43)	0.001	3.18 (1.65, 6.40)
LDSA	0.0002	5.22 (2.16, 12.6)	0.0001	5.26 (2.28, 12.74)

Note: models adjusted for age, sex, education level, and study area as random effect.

[a] All air pollution metrics have been natural log-transformed (N=338 for PM_{10}, N=339 for NO_2 and $PM_{2.5}$; N=189 for PNC and LDSA)

[b] Odds ratio (OR) for adult-onset asthma per one unite increase in the natural-logarithm of each air pollutants.

samples (76 cases, 113 controls). Median concentrations of air pollution were generally higher in the cities of Geneva and Lugano and were lower in the more rural and smaller communities Davos (for PM_{10}, $PM_{2.5}$ and NO_2 concentrations) and Wald (for PNC and LDSA). We observed high correlations between air pollutants within the study population (Figure S5.1) especially between PNC and LDSA (r=0.95) and between PM_{10} and $PM_{2.5}$ (r= 0.95). Exposure to UFP was significantly associated with an increased risk of adult-onset asthma as previously reported (Jeong et al., 2018) (Table 5.2). The odds ratio (OR) was 3.23 [95% CI (1.62, 6.43)] for one unit increase in natural log PNC, and 5.22 [95% CI (2.16, 12.6)] for one $\mu m^2/cm^3$ increase in natural-logarithm of LDSA (Table 5.2). These associations remained similar when not correcting for the effect of BMI in the model (Table 5.2). The odds ratio (OR) was 3.18 [95% CI (1.65, 6.40)] for one $particles/cm^3$ increase in natural-logarithm of PNC, and 5.26 [95% CI (2.28, 12.74)] for one $\mu m^2/cm^3$ increase in natural-logarithm of LDSA. No significant associations were identified for any of the other air pollution metrics.

Higher concentrations of IL-1RA, EGF, CCL22, CCL2, CRP, CXCL10, IL-17, and MPO were associated with an increased risk of adult-onset asthma (Figure 5.2, Table S5.1). The strongest association was seen for IL-1RA and adult-onset asthma with a corresponding OR of 1.37 [95% CI (1.09, 1.73); FDR= 0.08] for IQR (224.84 pg/mL) increase in natural-logarithm of IL-1RA. Noteworthy were the results for EGF OR=1.28 [95% CI (0.97, 1.96)] and CCL22 OR=1.20 [95% CI (0.9, 1.60)] which displayed marginally significant effects. Interestingly, CRP OR=1.05 [95% CI (0.92, 1.19)] showed no clear association with adult-onset asthma. (Figure 5.2, Table S5.1). Multiple air pollutants were associated with perturbations in immune markers although none of these associations were statistically significant (Figure 5.3). Higher concentration of IL-1RA and lower concentrations of CXCL10 were observed with all air pollutants. Positive associations between CRP and exposure to $PM_{2.5}$ and LDSA were observed. The association of IL-8 was negative with PNC, NO_2, and LDSA. MIP1 was positively associated with PNC and LDSA and negatively associated

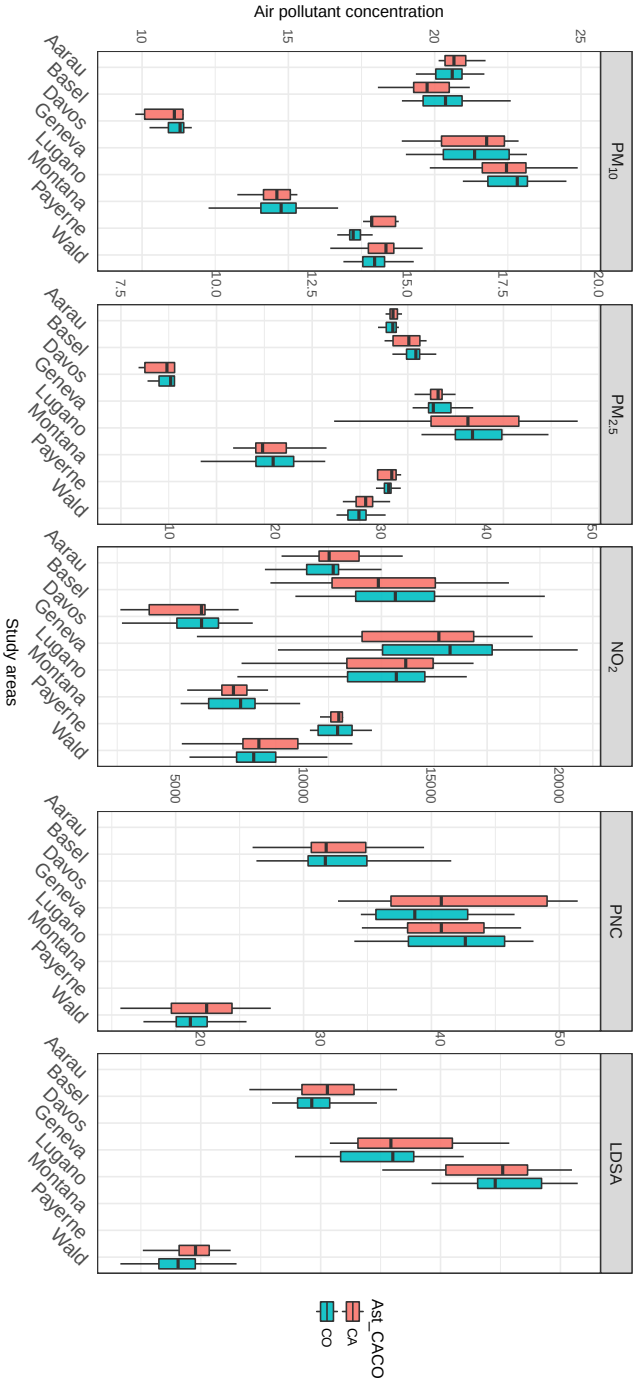

Figure 5.1: Box plots of distribution of air pollution concentrations by case (CA; red) and control (CO; green) per study area. Each panel shows one air pollution marker; PM_{10}, $PM_{2.5}$ ($PM_{2.5}$ and PM_{10} both, estimated from the PolluMap dispersion models; $\mu g/m^3$), NO_2 (estimated from LUR model; $\mu g/m^3$), PNC (particle number concentration; $particles/cm^3$), and LDSA (lung deposited surface area; $\mu m^2/cm^3$). Horizontal lines correspond to medians, and boxes to the 25^{th} - 75^{th} percentiles; whiskers extend to data within the interquartile range times 1.5.

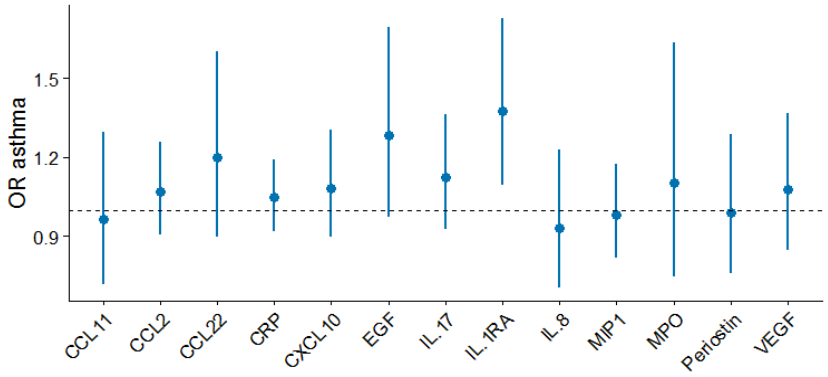

Figure 5.2: Odds ratios (OR) and 95%-confidence intervals for adult-onset asthma per IQR increase in the natural-log of each immune markers. IL-1RA was significantly (p-value = 0.01, FDR= 0.08) associated with risk of adult-onset asthma.

with PM_{10} and $PM_{2.5}$ exposures.

Partial least square path modeling (PLS-PM)

Figure 5.4, shows PLS path diagrams of the latent variables *immune-modulation* and *air pollution* for adult-onset asthma. We conducted PLS-PM on the subset of data where UFP estimates were available (76 cases, 113 controls). In the measurement (outer) model, the loading and the paths revealed a quantitative relationship between the latent and observed variables. All air pollution markers had positive effects on the latent variable *air pollution* and each explained around 86-96% variance of the corresponding latent variable (Figure 5.4, Table S5.2). In the block of *immune-modulation*, IL-1RA, CRP, VEGF, MIP1, and CCL22 explained the highest variability with a positive effect on the corresponding latent variable (each more than 40%; Figure 5.4, Table S5.2). In the structural (inner) model, the relationships between the latent variables were quantified with the standardized path coefficients (β) (Figure 5.4). *Air pollution* had a significant positive effect on *asthma* (path coefficient= 0.49 (OR= 1.6; for one-unit increase in the *air pollution* latent normalized structure), P-value= 0.002; Table 5.3) and on *immune-modulation* (path coefficient= 0.14, P-value=0.05; Table 5.3). The association between *immune-modulation* and *asthma* was positive and significant (path coefficient=0.62 (OR= 1.9; for every one-unit increase in the

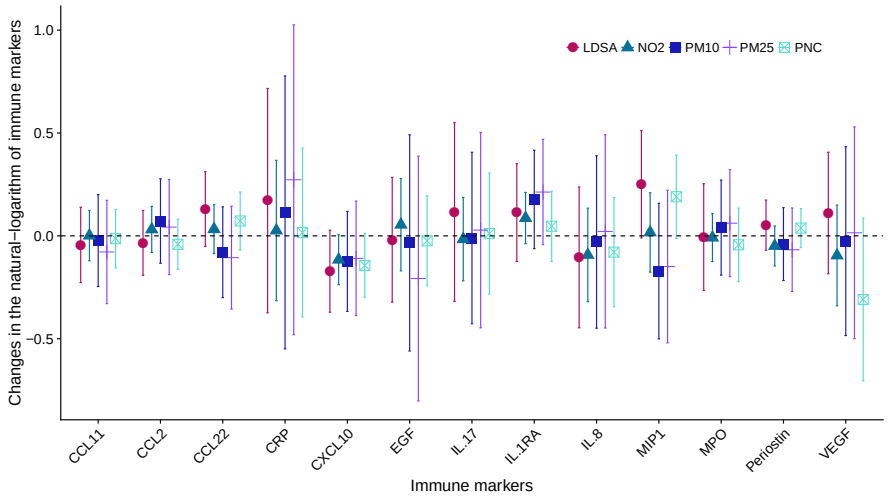

Figure 5.3: Effect estimates and 95%-confidence intervals for changes in the natural-logarithm of immune markers in per unit increase in the natural-logarithm of air pollution markers (N=338 for PM_{10}, N=339 for NO_2 and $PM_{2.5}$; N=189 for PNC and LDSA).

Table 5.3: Partial least square path modeling analysis for the relationships between latent variables

	β^a	P-value	R^2	β^b	95% CI	OR^d
$\beta_{Airpollution->immune-modulation}$	0.14	0.057	0.02	0.08	[-0.27, 0.32]	
$\beta_{Airpollution->Asthma}$	0.49	0.002	0.11^c	0.49	[0.18, 0.82]	1.6
$\beta_{immune-modulation->Asthma}$	0.62	0.0004	0.11^c	0.62	[0.29, 0.98]	1.9

[a] Path Coefficients calculated using data; [b] Path Coefficients calculated using 200 data set in Bootstrap;

[c] For logistic regression we calculated pseudo-Nagelkerke R^2

[d] Odds ratios (OR) for adult-onset asthma per one unit increase in the corresponding latent variable.

immune-modulation), P-value=0.0004; Table 5.3). The overall positive association of *air pollution* with *asthma* can be partialy explained by *immune-modulation* (path coefficient = 0.58). This overal effect can be calculated by summing up the indirect (path coefficients = 0.14 $\times$ 0.62) and direct (path coefficient= 0.49) effect of *air pollution* on *asthma*. The results of the bootstrap sampling indicated that the original coefficient path value and the one obtained from the bootstrap are close (Table 5.3). The average coefficient of determination R^2 value for two latent variables in our analysis, *immune-modulation*, and *asthma* (pseudo-Nagelkerke R^2), was found to be 0.02 and 0.11 (Table 5.3) respectively. This indicated that *immune-modulation* and *air pollution* together explained 11% of variation in *asthma*.
Our results indicated that of the total effect of *air pollution* on asthma 15% (ratio of an indi-

rect effect on total effect) is explained by the indirect effect of *immune-modulation* although this effect was not formally statistically significant (path coefficient= 0.09, Z-statistics= 1.64, P-value= 0.09).

Sensitivity analysis

Sensitivity analysis indicated that the significant associations of LDSA and PNC with adult-onset of asthma were generally robust to the set of confounders considered. The significant association of IL-1RA with adult-onset of asthma that we observed in the main analysis was recovered in all of the sensitivity analysis. The significant effect disappeared when we included BMI in the model as a confounder or in the analysis stratified by BMI status (Figure S5.2B). The association between air pollution and immune markers did not changed when we stratified analysis by BMI or included BMI as a confounder in the model (Figure S5.3). Results of PLS-PM among the overweight subgroup indicated a stronger association between *immune-modulation* and asthma (path coefficient=0.66 (OR= 1.93; for one-unit increase in the *immune-modulation*), P-value=0.01; Table S5.5) as well as between *air pollution* and *immune-modulation*, path coefficient = 0.29, P-value=0.003; Table S5.5), compared to the main analysis. This analysis also showed that a higher proportion (27%) of the total effect of *air pollution* on *asthma* is explained by the indirect effect of *immune-modulation* (path coefficient= 0.19, Z-statistics= 1.92, P-value= 0.05) in comparison with the main analysis (15%).
The magnitude of the risks (ORs) of immune markers associated with adult-onset asthma was often stronger for allergic asthma than non-allergic asthma (Figure S5.2C). In the allergic group, we observed a positive significant association of CCL22 and IL-1RA with adult-onset asthma. The odds ratios were 1.55 [95% CI 1.07, 2.23] for IQR (245.6 pg/mL) increase in natural-logarithm of CCL22, and 1.38 [95% CI 1.06, 1.80] for IQR (220.4 pg/mL) increase in natural-logarithm of IL-1RA. The associations of EGF and Periostin with adult-onset asthma were positive but marginally significant. The odds ratios were 1.36 [95% CI 0.97, 1.90] for IQR (56.8 pg/mL) increase in natural-logarithm of EGF, and 1.34 [95% CI 0.96, 1.88] for IQR (33862.4 pg/mL) increase in natural-logarithm of Periostin. For allergic asthma, PLS-PM also indicated a stronger association between *immune-modulation* and *asthma* (path coefficient=0.82 (OR= 2.27; for one-unit increase in the *immune-modulation*), P-value=0.0003; Table S5.4). This analysis also showed that a higher proportion (22%) of the total effect of *air pollution* on *asthma* is explained by the indirect effect of *immune-modulation* (path coefficient= 0.12, Z-statistics= 1.6, P-value= 0.1) in comparison with the main analysis (15%).

Discussion

We investigated the relationship between exposure to air pollution, immune markers, and adult-onset asthma using data from the Swiss Cohort Study on Air Pollution and Lung and Heart Diseases in Adults (SAPALDIA). Previous studies have investigated the associations of immune markers either with air pollution or with adult-onset asthma separately, thus limiting causal interpretation. The main aim of this study was to elucidate the biological mechanism linking exposure to air pollution to adult-onset asthma and to study if immune-markers mediate this association. We previously published on the positive association between biannual mean exposure to UFP (both PNC and LDSA) with adult-onset asthma (Jeong et al., 2018). In these extended analyses, we observed non-significant positive associations between PM_{10}, $PM_{2.5}$ and NO_2 and adult-onset asthma. Additionally, we observed an increased risk of asthma with elevated concentrations of IL-1RA, and to some extend with EGF, and CCL22. Of these IL-1RA achieved Benjamini-Hochberg (BH-FDR) $<$ 0.2 in main analysis as well as in sensitivity analyses. We observed limited evidence for a univariate association between air pollution and immune markers. IL-1RA was positively associated with all air pollutants and higher concentration of MIP1 and CRP and lower concentration of IL-8 was observed with UFP exposures. Our study further suggested that 15% of the total effect of air pollution on adult-onset asthma is explained by the indirect effect of immune markers. These associations were robust after inclusion of different confounders. A previous review of epidemiological studies related to air pollution and adult-onset asthma found inconsistent evidence in support of a causal relationship and identified the need for large-scale cohort studies and the inclusion of local-scale traffic pollutants like UFP (Jacquemin et al., 2012). Our investigation addressed both of these needs and our results indicated that UFP contributes to increases in adult-onset asthma independent of other potential risk factors. The significant association, which we reported previously, with UFP in our study was in contrast with a recent study that assessed long-term exposure to UFP derived from a city-specific LUR model in Toronto and linked it to health registry data of 1.1 million adult city residents. They reported no significant association of UFP exposure with respiratory disease incidence including adult-onset asthma (Weichenthal et al., 2017). A study of young adult Italians (aged 2044 years) found that each 18.3 $\mu g/m^3$ increase of NO_2 was associated with increased asthma prevalence (OR 1.13, 95% CI 0.98, 1.32), albeit not statistically significantly (De Marco et al., 2002). A national study among 12 177 Australian females reported a null association between 3-year mean annual NO_2 and asthma prevalence (Lazarevic et al., 2015).

Local inflammation together with oxidative stress has been suggested as a potential biological pathway explaining higher susceptibility of individuals with asthma to air pollution exposure (Auerbach and Hernandez, 2012). Growing evidence also indicates that systemic inflammation plays an important role in the pathogenesis of asthma (Wouters et al., 2009). We investigated a relatively wide panel of inflammatory biomarkers to eval-

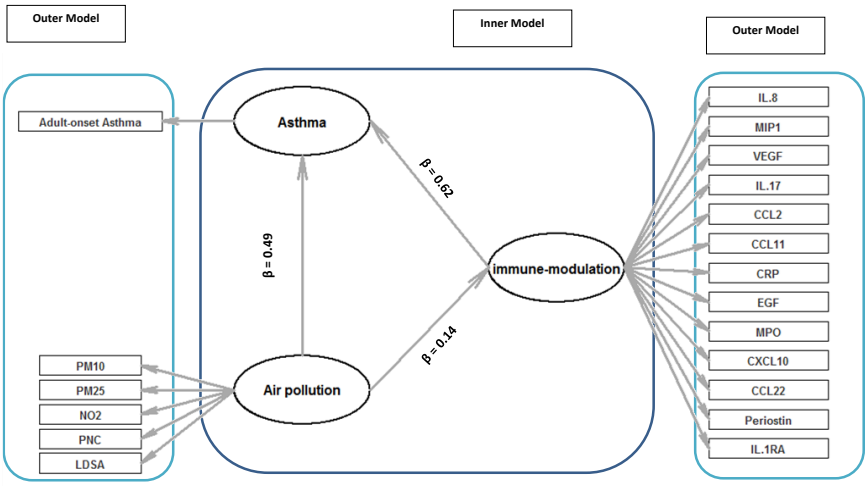

Figure 5.4: Path diagram indicating the conceptual model behind the relations among latent variables and their manifest variables. Rectangles refer to the manifest variables (outer model) and the ellipses refer to the latent variables (Inner model). Arrows show assumed causations among the variables (either latent or manifest), and the direction of the arrow defines the assumed direction of the relation. Path coefficients ($\beta's$) indicate the quantification of the relationship between latent variables. Corresponding ORs for the one unit increase in immune-modulation and air pollution on asthma are 1.9 and 1.6, respectively.

uate the effects of long-term exposure on systemic inflammatory status in adults. Our results are consistent, in terms of direction, with other studies on immune markers measured in adults showing elevated levels of CRP, and MIP1 (Fiorito et al., 2018; Hajat et al., 2015; Lanki et al., 2015) and decreased levels of IL-8 and EGF (Fiorito et al., 2018; Mostafavi et al., 2015) in relation to long-term exposure to air pollution. These markers were not statistically significant in our study. Our study also provides evidence for perturbation of plasma concentration of IL-1RA in association with exposure to air pollution in particular UFP. To date, this immune marker has not been reported in previous studies on immune markers and air pollution. IL-1RA, a pleiotropic cytokine released from various types of cells, including mononuclear cells, macrophages, T-lymphocytes, is one of the most important anti-inflammatory cytokines regulating the progression of the immune response by suppressing the production of a number of pro-inflammatory cytokines, including TNF-α, IL-1β, IL-6, and IL-8 (Chung, 2001; Dobreva et al., 2009).

Our study shows that plasma concentration of IL-1RA, EGF, and CCL22 were elevated in asthma patients as compared to the control group. This is in line with previously re-

ported increases in expression of IL-1RA in asthmatic airway epithelium (Sousa et al., 1997). Moreover, it has been well established that IL-1, of which IL-1RA is the receptor antagonist, is a critical mediator of the inflammatory process in various diseases, including asthma (Whelan et al., 2004). Therefore, it is reasonable to suggest that these immune markers may, at least in part, be responsible for a dysregulated airway inflammatory response in asthma (Whelan et al., 2004).

In addition to univariate regression analysis, we applied PLS path modeling that is designed to test for associations in a network of causal relationships. This approach overcomes statistical limitations in the assessment of perturbations by combining information from multiple variables rather than assessing them one by one. We focused on PLS path modeling in our analysis because we were interested in identifying to what extent the association between air pollution and asthma can be explained by air pollution induced immune modulation. Moreover, PLS path modeling enabled a joint assessment of the full set of immune biomarkers and air pollution exposures and does not suffer from instability in estimates and failure to converge due to multicollinearity. Our study also demonstrated that PLS path modeling is effective in assessing the total effects and relative importance of various pathways, as reflected in the latent variables, on adult-onset asthma. As univariate analysis indicated a stronger association of UFP with asthma, we also tested the mediating effect of immune markers when only using PNC and LDSA in the outer model of PLS-PM Table S5.3). Interestingly, this model indicated a similar mediating effect of UFP (as measured by PNC and LSDA) on adult-onset asthma (15%). Sensitivity analyses using a two-stage regression approach to correct first for confounding factors and then apply a PLS-PM on the residuals lead to essential similar results (not shown).

Epidemiological studies have shown that obesity represents a pro-inflammatory state (Jacobsen et al., 2016) which itself may increase susceptibility to air pollution by increasing the response to inflammatory stimuli. Being overweight has been shown to increase susceptibility to the respiratory effects of exposure to air pollution (Adam et al., 2015; Miller and Peden, 2014; Schikowski et al., 2013). In our analysis, the significant effect of immune markers on asthma weakened when we corrected for the effect of BMI. As BMI may be on the causal path, inclusion of BMI in the model will in that scenario lead to an underestimation of the effect of air pollution on immune markers. We therefore put more emphasis on the non BMI corrected analyses. Interestingly, PLS mediation analysis among the overweight sub group showed a stronger mediation effect of the immune system on the relation between air pollution and asthma (27% versus 15% in the overall analyses). This result may provide additional support to the hypothesis that obesity may indeed increase susceptibility to air pollution.

The strengths of our study include a relatively large and well-characterized study population, individual assessment of exposure to long-term air pollution based on locally optimized models, precise measurements of plasma immune markers concentrations, and evaluation of combined mediating effect of pleiotropic immune biomarkers using PLS-

PM.
Our study was also subject to some (potential) limitations. Even though we investigated a relatively large set of immune markers in relation to air pollution exposure and adult-onset asthma, these 13 immune markers may not fully represent all aspects of the immune system. As such the effect of the immune system on adult-onset asthma may be underestimated as well as its mediating effect of air pollution on asthma. Future large-scale studies should attempt to explore a broader range of inflammatory cytokines to expand our understanding of underlying mechanism and interactions between cytokines/chemokines in the inflammatory response related to air pollution exposure and asthma. It is not straightforward to define incident cases for asthma. Asthma is a complex chronic disease phenotype that develops over a long period of time, can go unnoticed for years if not for decades, and can also disappear as well as resurface. This difficulty inherent to asthma research is complicating the assessment of causality to identified risks such as air pollution (Jeong et al., 2018). We therefore restricted the outcome to adult (after the 16years of age) onset of asthma which is less susceptible to reverse causation bias and exposure misclassification.

Conclusion

This study adds to the growing body of evidence demonstrating inflammatory effects of exposure to long-term air pollution. Furthermore, this study provided supportive evidence for the mediatory effect of the immune system in the association between air pollutants (in particular UFP) and adult-onset asthma.

Supplemental material

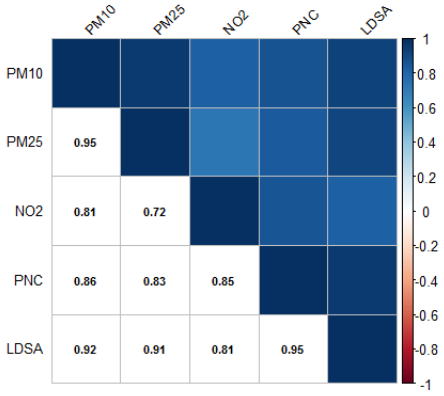

Figure S5.1: Pearson correlation coefficients for air pollution exposures. The color intensity of boxes indicates the magnitude of the correlation. Blue (red) color indicates the positive (negative) correlation. Numbers show the magnitude of correlation. Abbreviations are as follows: PM_{10}, $PM_{2.5}$ (PM_{10}, $PM_{2.5}$ both estimated from the PollouMap dispersion models; $\mu g/m^3$), NO_2 (NO_2, estimated from LUR model; $\mu g/m^3$), PNC (particle number concentration; $particles/cm^3$), and LDSA (lung deposited surface area; $\mu m^2/cm^3$).

Table S5.1: Association between immune markers and adult-onset of asthma (univariate analysis).

Immune markers	OR[a] [95% CI]	P-value	FDR	OR [b] [95% CI]
IL.8	0.98 [0.9, 1.06]	0.61	0.81	0.93 [0.71, 1.23]
MIP1	1.00 [1.00, 1.00]	0.82	0.89	0.98 [0.82. 1.17]
VEGF	1.00 [1.00, 1.01]	0.55	0.81	1.08 [0.85, 1.37]
IL.17	1.03 [0.98, 1.09]	0.23	0.76	1.12 [0.93, 1.36]
CCL2	1.00 [1.00, 1.00]	0.44	0.81	1.07 [0.90, 1.26]
CCL11	1.00 [0.99, 1.01]	0.80	0.89	0.96 [0.72, 1.29]
CRP	1.00 [1.00, 1.00]	0.49	0.81	1.05 [0.92, 1.19]
EGF	1.00 [1.00, 1.01]	0.08	0.51	1.28 [0.97, 1.69]
MPO	1.00 [1.00, 1.00]	0.62	0.81	1.10 [0.75, 1.64]
CXCL10	1.01 [0.99, 1.03]	0.41	0.81	1.08 [0.90, 1.31]
CCL22	1.00 [1.00, 1.00]	0.22	0.76	1.20 [0.90, 1.60]
Periostin	1.00 [1.00, 1.00]	0.93	0.93	0.99 [0.76, 1.29]
IL-1RA	1.001 [1.0004, 1.002]	0.01	0.08	1.37 [1.09, 1.73]

[a] Odds ratio for one pg/mL increase in natural-logarithm of each immune marker;

[b] Odds ratio for IQR (inter quartile range) increase in natural-logarithm of each immune marker;

Models adjusted for sex, age, education, bench time, fasting time, Fourier-transformed venipuncture time point assuming one or two periods per day, and their multiplicative interaction terms with fasting time as fixed effect and study area and plate as a random-effects.

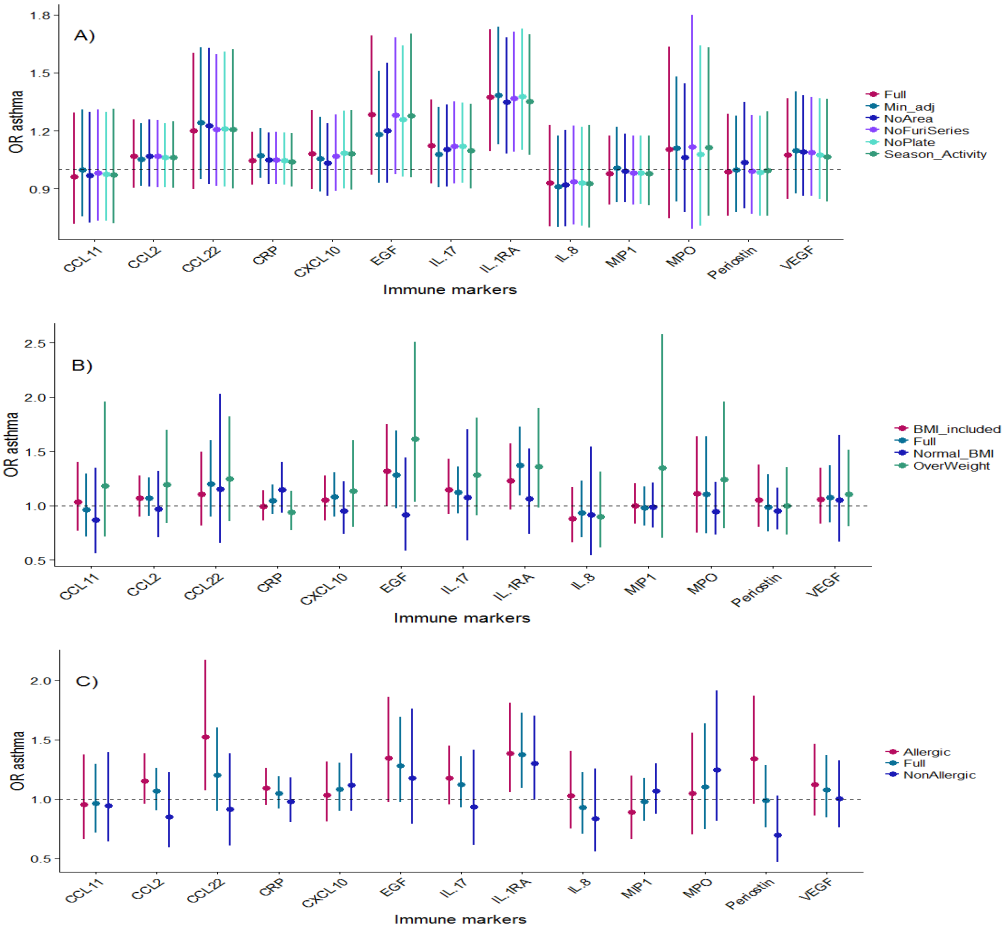

Figure S5.2: (A) Association between immune markers and adult-onset asthma for the different set of confounder adjustment including Full model adjusted for fixed-effects of sex, age, bench time, fasting time, Fourier-transformed venipuncture time point assuming one or two periods per day, and their multiplicative interaction terms with fasting time and random-effects of study area and plate. Min-adj model adjusted for age and sex. NoArea model adjusted for all variables of Full model except for study area. NoFourier-Series model adjusted for all variables of Full model except for Fourier-transformed venipuncture time point assuming one or two periods per day, and their multiplicative interaction terms with fasting time. NoPlate model adjusted for all variables of Full model except for microtiter plate. Season-Activity model adjusted for all variables of Full model and additionally for physical activity and season. (B) Association between immune markers and adult-onset asthma in the pooled population, full model additionally adjusted for BMI and stratified by BMI. (C) Association between immune markers and adult-onset asthma in the pooled population, and stratified by Allergic and none allergic.

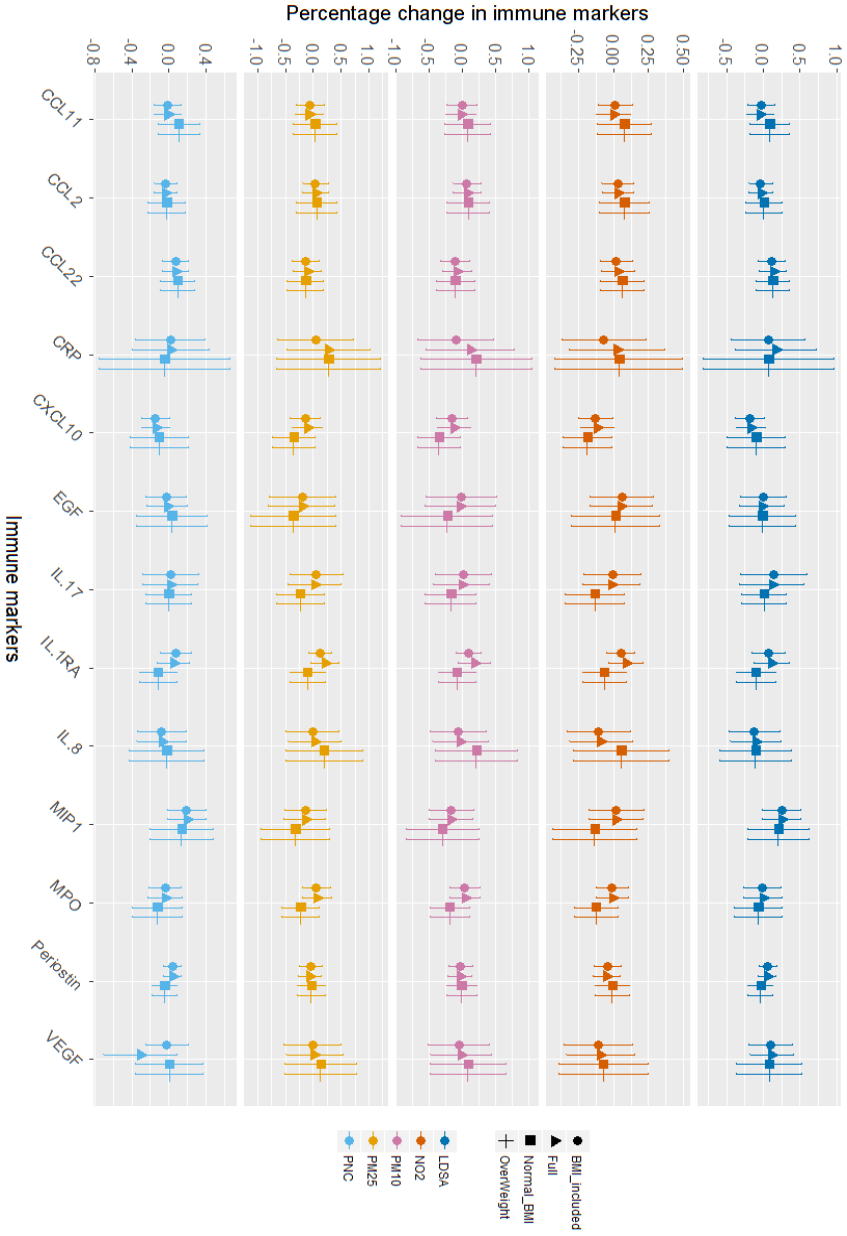

Figure S5.3: Association between immune markers and different air pollution exposures in the pooled population, full model additionally adjusted for BMI and stratified by BMI. Colors show the different air pollution exposure and Shapes show different model.

Table S5.2: Partial least-squared path modeling (PLS-PM) for air pollutants, immune markers and asthma. Loadings of manifest variables on latent variables.

Manifest variable	loadings
	Air pollution
PM_{10}	0.96
PM_{25}	0.93
NO_2	0.86
PNC	0.96
LDSA	0.97
	Immune modulation
IL.8	-0.005
MIP1	0.47
VEGF	0.52
IL.17	0.05
CCL2	-0.05
CCL11	0.11
CRP	0.65
EGF	0.15
MPO	0.33
CXCL10	0.24
CCL22	0.43
Periostin	0.12
IL-1RA	0.77
	Asthma
Adult-onset asthma	1.00

Table S5.3: Partial least square path modeling for the relationships between latent variables (when only using PNC and LDSA as manifest variables of air pollution).

	β^a	P-value	R^2	β^b	95% CI	OR^d
$\beta_{Airpollution->immune-modulation}$	0.15	0.04	0.02	0.08	[-0.26, 0.33]	-
$\beta_{Airpollution->Asthma}$	0.54	0.001	0.11^c	0.53	[0.22, 0.86]	1.7
$\beta_{immune-modulation->Asthma}$	0.62	0.0004	0.11^c	0.62	[0.28, 0.98]	1.9

^{a} Path Coefficients calculated using data; ^{b} Path Coefficients calculated using 200 data set in Bootstrap;

^{c} For logistic regression we calculated pseudo-Nagelkerke R^2

^{d} Odds ratios (OR) for adult-onset asthma per one unit increase in the corresponding latent variable.

Total effect of air pollution on asthma is 0.54+ 0.15$\times$0.62 = 0.63;

Indirect effect of air pollution on asthma = 0.15$\times$0.62 = 0.09 (Z score= 1.76; P-value= 0.07);

Ratio of indirect effect/total (VAF) is 15%.

Table S5.4: Partial least square path modeling analysis for the relationships between latent variables (for Allergic asthma (N= 154)).

	β^a	P-value	R^2	β^b	95% CI	OR[d]
$\beta_{Airpollution->immune-modulation}$	0.15	0.07	0.02	0.08	[-0.30, 0.37]	-
$\beta_{Airpollution->Asthma}$	0.42	0.032	0.13[c]	0.42	[0.04, 0.82]	1.53
$\beta_{immune-modulation->Asthma}$	0.82	0.0003	0.13[c]	0.82	[0.40, 1.29]	2.27

[a] Path Coefficients calculated using data; [b] Path Coefficients calculated using 200 data set in Bootstrap;

[c] For logistic regression we calculated pseudo-Nagelkerke R^2

[d] Odds ratios (OR) for adult-onset asthma per one unit increase in the corresponding latent variable.

Total effect of air pollution on asthma is 0.42+ 0.15×0.82 = 0.54;

Indirect effect of air pollution on asthma = 0.15×0.82 = 0.12 (Z score= 1.6; P-value= 0.1);

Ratio of indirect effect/total is %0.22; Goodness of fit 0.18.

Table S5.5: Partial least square path modeling analysis for the relationships between latent variables among overweight (N=102).

	β^a	P-value	R^2	β^b	95% CI	OR[d]
$\beta_{Airpollution->immune-modulation}$	0.29	0.003	0.1	0.17	[-0.39, 0.48]	-
$\beta_{Airpollution->Asthma}$	0.52	0.02	0.12a	0.52	[0.09, 0.98]	1.7
$\beta_{immune-modulation->Asthma}$	0.66	0.01	0.12a	0.66	[0.18, 1.18]	1.93

[a] Path Coefficients calculated using data; [b] Path Coefficients calculated using 200 data set in Bootstrap;

[c] For logistic regression we calculated pseudo-Nagelkerke R^2

[d] Odds ratios (OR) for adult-onset asthma per one unit increase in the corresponding latent variable.

Total effect of air pollution on asthma is 0.52+ 0.29×0.66 = 0.73;

Indirect effect of air pollution on asthma = 0.29×0.66 = 0.19 (Z score= 1.92; P-value= 0.05);

Ratio of indirect effect/total is %0.27; Goodness of fit 0.2.

CHAPTER 6

GENERAL DISCUSSION

Humans are exposed to many potentially harmful compounds in daily life. Both long- and short-term exposure to (low) levels of these compounds by itself or in mixtures may adversely affect health. Investigating perturbations in the internal exposome, which can be measured via OMICS technologies, can provide information on direct measures of exposure or on physiological perturbations that are the reflection of a certain exposure. Studying the internal exposome can provide biological underpinnings for empirically observed exposure-disease associations.
Recent developments in high-throughput technologies, bio-informatics tools, and sophisticated statistical methods have enabled the analysis of OMICS data in the context of environmental health studies. This thesis addresses the impact of long- and short-term exposure to air pollution as a specific environmental pollutant on three well developed OMICS platforms (i.e. proteomics, transcriptomics, and DNA-methylomics) using mainly untargeted approaches. We aimed to identify perturbations in the OMICS data in order to have a better understanding of the biological process that link air pollution exposure to health effects. In this chapter, the main findings of this dissertation are briefly summarized, followed by a discussion of the challenges encountered and possible future improvements and extensions.

Main findings

We first evaluated the effect of long-term exposure to air pollution (**Chapter 2** and **3**). Annual modeled outdoor concentrations of nitrogen oxides (NO_X) at the study participants baseline home-addresses were available from the European Study of Cohorts for Air Pollution Effects (ESCAPE) project. Biological samples collected from 550 healthy subjects participating in cohorts from Italy and Sweden were available from the Genomics Biomarkers of Environmental Health project (EnviroGenoMarkers).
In **Chapter 2**, perturbation in plasma concentrations of a large panel of inflammatory markers (N=28) was studied. Long-term exposure to NO_X was associated with decreased levels of interleukin (IL)-2, IL-8, IL-10 and tumor necrosis factor-α in Italy, but not in Sweden. Combining data from Italy and Sweden we observed a significant association between long-term exposure to NO_X and decreased levels of circulating IL-8. Effects were stronger in Italy than in Sweden, potentially reflecting the difference in air pollution levels between the two cohorts. NO_X exposure levels were considerably lower in Sweden than in Italy. Median NO_X level was 6.65 μ/m^3 in Sweden and 94.2 μ/m^3 in Italy.
In **Chapter 3**, we studied alterations in genome-wide gene expression associated with long-term average exposure to NO_X. We applied several statistical methods to reduce the risk of false positive findings in the high dimensional data analyses. For the univariate analysis, we reduced this risk by controlling the false discovery rate (FDR) and by using evidence from the literature as informal priors in our analysis. Furthermore, we applied

two additional variable selection approaches (Elastic-Net and GUESS). Agnostic univariate analysis identified six noteworthy genes that were associated with NO_X, Elastic-Net regression selected four of these whereas GUESS did not select any. Functional analysis of these genes using the NIH-DAVID bioinformatics resource yielded no evidence for functional enrichment in any biologic pathway. This is likely due to the limited number of probes found in this study. However, genes hypothesized to be downregulated due to cigarette smoking were enriched among the most strongly downregulated genes from our study. On a global level, the observed changes in the transcriptome may indicate similarities between air pollution- and smoking-induced changes in the transcriptome. These results indicate that by combining evidence bases from different pollutants we can elucidate biological phenomenon that otherwise would have gone unnoticed.

In **Chapter 4**, we addressed the acute health effects of exposures to different air pollution markers. We collected repeated 24-hour personal and ambient exposure measurements of particulate matter ($PM_{2.5}$), $PM_{2.5}$ absorbance, and ultrafine particles (UFP) as well as peripheral blood samples from a panel of 157 healthy non-smoking adults living in four European countries. We investigated the association of personal and ambient air pollution exposure measures with genome-wide DNA-methylation perturbations at single CpG (cytosine-guanine dinucleotide) sites. We identified 13 DNA methylation sites that were associated with personal exposure to $PM_{2.5}$. Functional analysis of the identified CpG sites at the mapped genes did not provide insight on potentially affected biologic pathways. This is likely due to the limited number of CpG sites found in this study. We extended the analyses by looking at differentially methylated regions (DMRs), where we identified several associations between both air pollutants and DMRs. The impact of air pollution on the blood methylome is likely subtle. Its characterization therefore requires high-quality exposure assessment and considerable statistical power. Techniques such as DMRs may increase the power of a study by reducing the overall dimension of the data by combining inference of nearby CpGs together. However, biological interpretation of both DMRs and CpGs remains challenging. Cross-OMIC analyses relating methylation changes to functional changes in RNA, protein, and metabolite expression may aid in this regard. Datasets that allow such analyses are still limited but can be expected to grow (e.g. BIOS (https://omics-explorer.bbmri.nl/)) in the near future, allowing for more functional interpretations of methylation changes due to environmental exposures.

In **Chapter 5**, the interplay between air pollution, immune markers, and adult-onset asthma was investigated following the Meet-in-the-Middle concept. We considered long-term exposure to multiple air pollution markers (particulate matter smaller than 2.5 μm ($PM_{2.5}$), and smaller than 10 μm (PM_{10}), nitrogen dioxide (NO_2), ultra-fine particulates (UFP) [based on two metrics: particle number concentration (PNC) and lung deposited surface area (LDSA)]) and a panel of 13 immune markers in peripheral blood samples collected from 140 adult cases (asthma diagnosis after the age of 16) and 199 controls using a nested-case control design. The main aim of this study was to elucidate the biologi-

cal mechanism linking air pollution exposure to adult-onset asthma and to study if any immune-markers mediate this association. We used a novel method, partial least square path modeling (PLS-PM), which allows study of latent structures as risk factors and mediators. The combination of air pollutants was associated with risk of adult-onset asthma (odds ratio (OR) = 1.7, P-value= 0.002) and with a cluster of immune-marker changes (path coefficient= 0.14, P-value=0.05). In turn immune-marker changes was associated with risk of adult-onset asthma (OR = 1.9, P-value=0.0004). Our study indicated that 15% of the effect of air pollution on risk of adult-onset asthma was explained through the immune system (path coefficient = 0.09, p-value= 0.09). This effect was mostly driven by UFP while the mediation was through a complex interplay between different immune-markers.

Promises of OMIC marker in air pollution-health studies

Our studies and others conducted in recent years have provided initial clues for the impact of air pollution on various OMICS markers. In the introduction of this thesis I listed three potential contributions of OMICS markers to air pollution health research: generating new hypotheses, identifying pollutant specific pathways, and detecting biomarkers of early disease (meet-in-the-middle). In this section, I reflect on the early contributions of the OMICS markers to the field of air pollution research, as part of the wider field of using OMICs to study the internal exposome, based on the studies included in this thesis and the literature.

- Generating new hypotheses on air pollution induced health effects
  The studies described in this thesis provided some evidence that several OMICS markers were associated with short- or long-term exposure to air pollution. Recent studies of air pollution and OMICS markers by others reported similar results, albeit overlap between specific findings is generally limited. For example, in the KORA F3, F4 (N=2,300) and Normative Aging Cohorts (N= 657), several sets of CpG sites were associated to (7 and 28 day) trailing averages of $PM_{2.5}$ monitoring station measurements (Panni et al., 2016). In the MESA ($PM_{2.5}$; N=1,207), EPIC (NO_2 and NO_X; N= 613), and LIFELINES (NO_2; N= 1,017) studies, several sets of CpG sites were identified to be associated with long-term modeled average exposure air pollutants (Chi et al., 2016; de FC Lichtenfels et al., 2018; Plusquin et al., 2017). A combined cohort of 400 Swiss and 400 Italian adults with long-term air pollution exposure estimates reported enrichment of metabolic pathways for linoleate metabolism and fatty acid activation for $PM_{2.5}$ exposure, linoleate, glycerophospholipid and glycosphingolipid metabolism for UFP exposure, and carnitine shuttle and pyrimidine metabolism for NO_2 exposure (Jeong et al., 2018). Additionally, Wittkopp et al. (2016) tested whether gene expression levels were associated with low level air pollution exposures (average PM2.5 1012 $\mu g/m^3$) in a cohort of elderly

subjects. Candidate genes were selected from published studies of gene expression linked to pollutants. The author found positive associations of traffic-related pollutants (including $PM_{0.25-2.5}$ PAH and/or $PM_{0.2}$ PAH, and NO_X) with the expression of several genes, in particular of Nrf2-mediated genes, indicating the involvement of oxidative stress pathways (Wittkopp et al., 2016).
True to the nature of observational evidence one should be careful in translating each of these findings into a hypothesis of the potential impact of air pollution on human biology. TThe process of generating new hypotheses from empirical associations found in observational studies typically involve both an assessment of the consistency across multiple studies and some mechanistic evidence for the plausibility of the observed associations. The air pollution-OMICS literature to date is characterized by a general lack of replication and consistency of identified markers across the studies. This is potentially driven by differences in exposure and study populations, as well as by false positive findings in different studies. At the same time, number of findings per study are often modest, precluding the effective use of pathway exploration tools such as enrichment analysis that could shed light on the coherence of findings using existing mechanistic models.
A new generation of air pollution OMICS studies may overcome these limitations by significantly increasing statistical power by establishing large consortia in which (international) data can be pooled (further discussed below). Findings from the first generation of air pollution OMICS studies, including ours, can contribute in these analyses by providing insight into potential air pollution targets on the different OMICS platforms, thereby further improving statistical power. We provided an example of such an approach in chapter 3, where we used a list of markers previously associated to air pollution to conduct targeted analysis in addition to agnostic screening. In the future more advanced solutions may be implemented, for example using Bayesian statistics.
While the studies described in my thesis were primarily purely agnostic in nature (uninformed screening for statistical associations) further steps towards generating new hypotheses and supporting existing hypotheses can be made by incorporating biological knowledge from other domains such as systems biology and mechanistic studies into the statistical models, resulting in 'informed screening'. DMR is a simple example of such an approach, as it uses the location of CpG sites on the chromosomes to cluster findings. Incorporating more complex relations between individual OMICs markers, for example using information on biological pathways, can further extend such informed screening approaches.

- Identifying pollutant specific pathway
  Air pollution is a complex mixture of gases and particles. Many epidemiological studies have demonstrated that associations with health outcomes could differ be-

tween the different components of this mixture. This is in line with the fact that different components of air pollution might pose different health risks depending on particle size, reactive surface area, and chemical composition. Currently, there is a paucity of knowledge about whether this is the result of differential measurement error between these components, or whether they actually operate via pollutant specific pathways. I evaluated results described in Chapter 4 and in other studies to see to which extent we observed evidence for the hypothesis that the pollutants in the mixture in fact operate via pollutant specific pathways. Such information would be valuable in developing preventive measures to reduce the impact of air pollution on the burden of disease because different measures may have different effects on various components in the mixture understanding better the drivers of the disease and underlying biology also provide insights in the etiology of disease.

In Chapter 4, we studied 3 specific pollutants: $PM_{2.5}$, absorbance of $PM_{2.5}$, and UFP. Associated numbers of DMRs varied between these pollutants. There were overlaps between the annotated genes across significant DMRs between personal $PM_{2.5}$ and personal $PM_{2.5}$ absorbance, and between personal $PM_{2.5}$ absorbance and ambient $PM_{2.5}$ absorbance while for other combinations overlap was less clear. Considering the relatively high correlation that often exists in exposure measurements of these pollutants (in our study varying from 0.48 to 0.74), such overlap is not surprising and highlights the difficulty of isolating pollutant specific pathways in an observational epidemiological study.

Various other studies reported distinct epigenetic alterations in response to different components of air pollution, including gases and particles in different size fractions. An EWAS study of the European Prospective Investigation into Cancer and Nutrition (EPIC) on exposure to different particle sizes (PM10; coarse PM, subset of PM_{10} that is larger than 2.5 m; and $PM_{2.5}$), soot (absorbance of the $PM_{2.5}$ filter), NO_X, and NO_2 described only limited similarity among the CpG sites that are differentially methylated in response to each exposure (Plusquin et al., 2017). The Oxford street study, with an experimental crossover design, revealed different sets of miRNAs associated with each of NO_2, UFP, $PM_{2.5}$, PM_{10}, and black carbon exposure (Krauskopf et al., 2018).

The previously noted lack in overlap between the results of most air pollution OMICS studies complicates the interpretation of the findings. While pollutant specific pathways have been reported, some of the observed differences might be due to factors other than biological differences. Further insight into this issue could perhaps be gained by focusing on the elemental composition of the particulates that are part of the air pollution mixture. Examples of such approaches include the Normative Aging Study, in which differential DNA methylation with $PM_{2.5}$ species (Fe, Ni, and V) was reported (Dai et al., 2017). Similarly in a targeted study on repair genes and components of PM_{10} showed significant CpG sites associated with ace-

naphthene, indeno [1, 2, 3-cd] pyrene, and pyrene (Alvarado-Cruz et al., 2017). However, as already mentioned above differences between studies and analyses on specific components of air pollution may in general be hampered by statistical power of these studies. It is therefore likely that part of the heterogeneity in results between components is driven by false positives. Larger studies will be imperative to move the field further towards identifying pollutant specific pathways.

- Detecting biomarkers of early disease (Meet-in-the-Middle)
  The Meet-in-the-Middle strategy, in which one searches for intermediate biomarkers that are linked to both exposure and disease, might provide insights into biological plausibility that can strengthen causal inference. This approach involves three steps: an investigation into the association between exposure and disease, an assessment of the relationship between exposure and biomarkers of exposure and the intermediate biomarkers, and an assessment of the relationship between the disease outcome and the intermediate biomarkers. Chapter 5 in this thesis is an example of the Meet-in-the-Middle strategy. I sought to illustrate the mediating effect of immune markers in the association between air pollution and adult onset asthma using a case-control study nested in the prospective SAPALDIA cohort. Using group level mediation analysis (PLS-PM) I observed that 15% of the effect of UFP on adult-onset asthma could be explained by the measured immune modulation. As such the results in Chapter 5 point towards the possibility that the measured immune markers in that study might play a role in the biological pathway that connects exposure to UFP to adult-onset asthma. The results are tentative because a mediation effect of the individual markers could not be detected. If confirmed by future studies, these findings would point towards a role for these immune markers as biomarkers of early disease for asthma.
  Another example of a Meet-in-the-Middle approach focused on air pollution is a recent study of DNA methylation/immune marker and cardiovascular disease (CDV) (Fiorito et al., 2018). The authors observed that altered DNA methylation of genes in the "ROS/Glutathione/Cytotoxic granules" and "Cytokine signaling" pathways was associated both with long-term exposure estimates of NO_2 and $PM_{2.5}$ as well as CVD risk (Fiorito et al., 2018). A pathway was suggested in which chronic exposure to air pollution can lead to oxidative stress, which in turn activates inflammatory responses mainly involving the cytokine signaling pathway, and finally leads to increased risk of CVD risk (Fiorito et al., 2018). The Meet-in-the-Middle approach has the potential to open new avenues for disease prevention by identifying the specific environmental factors involved in the disease process (Vineis et al., 2013). The application to the case of air pollution, however, raises major challenges both theoretical and practical in terms of study design (Kyrtopoulos, 2013). One important challenge to the approach pertains to the timing of biological sample collection. In

general, cohort studies can be used to collect prospective repeated biological samples to evaluate how changes in biomarkers overtime relate to disease risk. However, many large cohort studies focused on air pollution have only been able to collect a single biological sample at one point in time. Depending on the half-life of the biomarker, one prospective sample might not reflect both the exposure and outcome of interest in an optimal way, making it more difficult to detect a mediating role. In the presence of exposure to air pollution that vary from day to day, for biomarkers that have relatively short half-lives (e.g. gene expression) the timing of the biological sample collection is more important than for markers with longer half-lives (e.g. DNA-methylation). Therefore Meet-in-the-Middle approaches incorporating biomarkers with longer half-lives are more likely to succeed.

A potential pitfall for the Meet-in-the-Middle approach is to only focus on population level associations between the biomarker and the exposure and outcome of interest. For a biomarker to play a true role in the pathway from exposure to disease, such associations need to exist on an individual level. Therefore, improved statistical techniques such as mediation analyses should be used to illustrate the overlap between biomarker of exposure and biomarker of disease by simultaneously seeking the overlaps between distributions of disease, exposure, and biomarker. Traditionally, method development for mediation analysis has focused on univariate settings. However, such methods do not accommodate high dimensional OMICS data. PLS-PM applied in chapter 5 is an example of a methodology that is capable of group-level mediation analysis. Mediation analysis for high dimensional data is currently an active field of development and (van Kesteren and Oberski, 2018; Zhang, 2018) such methodologies should find an application in future OMICS studies.

Challenges in reproducibility of OMICS findings

An important issue in air pollution-OMICS studies has been the number of statistically significant findings that are not replicated in subsequent studies. The issue is mainly attributed to a high probability of false positive findings, which may be caused by insufficient sample size, imprecise air pollution exposure assessment, and less than optimal statistical approaches (Huang et al., 2010; Peretz et al., 2008; Pettit et al., 2012; Wittkopp et al., 2016). The studies presented in this thesis have attempted to address some of these issues, but the challenge still remains. In this section we discuss these challenges and provide some recommendations for future air pollution-OMICS studies.

Sample size

OMICS biomarkers have shown convincing and robust signals in response to strong environmental risk factors such as tobacco smoking. Considering the fact that both tobacco smoke and urban air pollution consist mostly of combustion product chemicals, we expected to observe more modest associations between OMICS biomarkers and air pollution similar to those observed for smoking. However, the air pollution results to date have been few. The reason likely lies in the fact that, apart from being a different type of exposure, air pollution exposure assessment is more prone to measurement error than smoking assessment. Levels of exposure to air pollution are also generally much lower than those related to the active inhalation of tobacco smoke. Moreover, OMICS smoking analyses have typically been performed in much larger studies than those on air pollution and OMICs markers (Justice et al., 2017; Sung et al., 2018). Therefore, in order to move the field further it is imperative to move towards large sample sizes to achieve sufficient power in analyses of OMICS biomarker in association with air pollution.
Due to the current high cost of OMICS technology, it is not possible to enlarge the population size to hundreds to thousands of subjects in one single study. Collaboration between studies and combined meta-analysis available data are needed to optimize the use of resources and to increase the likelihood of detecting perturbation of OMICS marker in response to air pollution exposure. In recent years many OMICS-air pollution studies include a replication strategy, such as a meta-analytical approach in which data from several studies have been analyzed by consortia. Research in large-scale consortia is powerful but demands high quality data with harmonization of air pollution exposure assessment, OMICS measurement, and study design over participating study centers. Although larger consortia such as the Normative Aging Study and KORA (N=2,300) (Panni et al., 2016) and Pregnancy And Childhood Epigenetics (PACE) (39 studies; N=29 000) (Felix et al., 2017) have been established, there are still opportunities to combine the present studies and build larger studies. This would allow for meta-analysis (and pooled-analysis) that could provide better exposure contrast and higher statistical power. Such studies may also be able to shed light on pollutant-specific pathways involved in exposure-disease relationships.

Exposure assessment

High-quality exposure assessment is key for identifying the relatively small changes in OMICS markers in response to air pollution exposure. In chapter 4, we used an innovative approach to monitor 24 hour personal exposure to air pollution by equipping volunteers with environmental sensors and GPS. We also estimated annual average ambient concentrations of air pollution for each individual using the LUR models from the ESCAPE and EXPOsOMICS (Eeftens et al., 2012a; Nunen et al., 2017) as a proxy for long-term modeled

ambient exposure. Personal exposure measurements in this study were more strongly associated with DNA methylation than long-term modeled exposure estimates. This observation could be an indication that acute changes in methylation due to exposure to air pollution identified in this study did not reflect a long-term association between air pollution and DNA-methylation.
However, differences in measurement error between personal exposure measurements and long-term modeled exposure estimates might also have contributed to the observed results. Long-term modeled exposure relies on retrospective estimates over a long period of time, which is inherently prone to considerable exposure measurement error. Considering the limited sample size of the study in chapter 4, larger measurement error would have reduced the statistical power to detect significant associations. However, although personal exposure monitoring is a precise exposure assessment method for current or recent exposures, due to the high cost and other logistic challenges it is not suitable to be applied in larger studies such as cohorts to detect chronic health effect of air pollution. Furthermore exposure measurements will only provide accurate estimates for short periods of time (24 hours in chapter 4), while estimates of long-term average exposure levels are likely better characterized by models. LUR models have successfully been used in large epidemiological studies to identify robust associations between a long-term exposure to air pollution and adverse health effects (Hoek et al., 2013; Beelen et al., 2014). Therefore, it would follow that with sufficient power these models should allow the identification of biological perturbations related to air pollution and health effects.

Statistical analysis

While the approach of p-value correction to account for the false positive rate (e.g. FDR, Bonferroni) has made the univariate model a prominent method in OMICS data analysis, this method has downsides in terms of loss of statistical power to detect true positives. Different multivariate approaches have been proposed to overcome this limitation. As explained in the introduction, application of multivariate approaches is not as straightforward in the exposure-OMICS analyses where OMICS is the outcome variable. These types of analyses were on of the main focuses in this thesis. One multivariate approach consists of a two-stage procedure: a model is first fitted to estimate the effect of the potential confounders and multivariate methods are subsequently applied to the residuals of the model while considering OMICS data as a predictor matrix and exposure as an outcome. We applied two multivariate approaches (Elastic-Net and GUESS) as complementary methods to univariate analysis in chapter 3 to assess the association between air pollution and genome-wide gene expression. Application of multivariate approaches to some extend reduces the risk of false-positive findings, but at the cost of reduced sensitivity (Agier et al., 2015). The primary goal of the OMICS data analysis in this thesis was to discover potential new biomarker targets of ambient air pollution. Therefore we preferred

a higher sensitivity to a lower risk of false-positive findings to analyze OMICS data and continued on using the univariate approach with lower FDR threshold in further analyses of OMICS markers.
In addition to univariate and multivariate approaches, different bioinformatics approaches have been proposed to increase the statistical power by reducing the dimension of the OMICS data. We applied DMR which overcome statistical limitations in the assessment of perturbations by combining information from multiple OMICS markers rather than assessing them one by one. It has been shown that the identification of regional differences across several probes provides more robust findings and is more likely to be replicated than individual CpG differences (Wright et al., 2016). Using DMR analysis in chapter 4 of this thesis resulted in much richer signals than univariate analysis. A potential approach to gain insight into the biological role of DMRs associated with air pollution involves mapping them to nearby genes and conducting functional enrichment analysis on these genes. Further functional interpretation of the identified regions requires moving towards regulatory enrichment analysis, incorporating not only mapping of nearby genes, but also accounting for trans-acting effects of DNA-methylation on gene expression (Wright et al., 2016).

Integration of more biological layers

Each OMICS platform (e.g. proteomics, transcriptomics, methylomics) provides a partial snapshot of the internal exposome. We expect that integrating response measures from different OMICS platforms would help bridge the gap between the numerous layers of interactions present in the human biological system and will contribute towards obtaining a more complete view of cellular regulation in response to environmental changes (Bersanelli et al., 2016). The cross-OMICS assessment of the interrelation between different OMICS relies on the selection of influential markers and determining their pairwise correlations. In chapter 4, we observed some evidence for complex interactions between DNA-methylation, gene expression, and immune markers by showing the correlations between these markers. The three correlation clusters point towards shared biological function, though further analyses are required to confirm this interpretation. No goldstandard is available for cross-OMICS integration and this work will likely remains challenging until further development in analytical strategies (Ritchie et al., 2015). In a study of occupational exposure to trichloroethylene (TCE), integration of untargeted metabolomics with established biomarkers of immune function and renal damage identified unknown metabolites of TCE associated with biological response, which was not found using traditional methods (Walker et al., 2016). This demonstrates how a limited number of biological response markers can provide insight on biological changes from chemical exposures. Continued development of statistical approaches to identify interactions among biological response networks and the application of multi-OMICS approaches to characterize human

exposures in cohort studies will likely spur discoveries in the coming years (Niedzwiecki et al., 2019).

Conclusion

In this dissertation, we studied three OMICS platforms and associated regulatory mechanisms in order to identify the biological processes that link air pollution exposure to health effects. Overall, the research conducted provided evidence for perturbation of different OMICS markers in association with both short-term and long-term exposures to different air pollutants. The work also identified changes in OMICS markers in association with exposures that may reflect mechanisms of action involved in potential disease development. The specific exposure-OMICS associations in this dissertation require replication before robust inferences can be drawn. Although evidence for the effect of exposure to air pollution on OMICS marker has been provided by our study, priorities for future work include increasing sample size, improving in the functional interpretation of identified OMICS markers, and integrating results from multiple OMICS platforms.

CHAPTER 7

APPENDICES

7.1 Summary

The majority of chronic diseases are likely to be the result of the combination of environmental exposures and human genetics. The exposome concept was proposed to increase our knowledge regarding the potential role of environment in the causation of disease. The exposome is a paradigm involving the study of all environmental exposures (e.g., air pollutants, chemical contaminants, diet or life style factors) and associated biological responses from conception until death. Once environmental exposures enter the human body they and their biological consequences become part of the internal exposome, which can be measured by OMICS technologies. Investigating perturbations in the internal exposome can provide information on direct measures of exposure or on physiological perturbations that are indicative of a certain exposure. Studying the internal exposome can provide biological underpinnings for empirically observed exposure-disease associations. In this thesis, we focused on air pollution, a high priority environmental pollutant responsible for 7 million premature deaths worldwide each year. Air pollution is recognized as a human carcinogen associated with lung cancer. It is also a major risk factor for other acute and chronic diseases including cardiovascular disease (CVD) and chronic respiratory diseases, such as asthma and chronic obstructive pulmonary disease (COPD). Though numerous epidemiological studies have observed consistent associations between exposure to ambient air pollution and adverse health effects, the underlying biological mechanisms of these effects have not yet been fully established. Oxidative stress and immune/inflammation responses are commonly considered as putative underlying mechanisms. Furthermore, there are persistent scientific questions regarding which constituents of ambient air pollution drive these biological processes and associated diseases. For example, it has been postulated that ultrafine particulates (UFP) may be more detrimental to health than larger particulates due to its small size enabling interstitialization, or absorption directly into the bloodstream.

The overall aim of this thesis is to study three OMICS platforms (proteomics, transcriptomics, and DNA-methylomics) in relation to exposure to air pollution in order to have a better understanding of the biological processes that link air pollution exposure to health effects. When successful such analyses might contribute to generating insights on air pollution induced health effects, identifying air pollutant specific biological pathways, and detecting biomarkers of early disease. In turn, such insights can contribute to causal inference and the development of preventative strategies.

The specific aims, each addressed in one of the constituent papers, were:

- To evaluate the impact of long-term exposure to air pollution on two OMICS platforms (immune markers and gene expression)
- To evaluate the relationship between 24hr personal exposure measurement of different air pollution measures and genome-wide DNA-methylation

- To investigate to what extent inflammatory markers mediate the association between long-term exposure to air pollution and adult-onset asthma

In the first two chapters we assessed the chronic effect of air pollution on OMICS markers in a relatively large (N=500) cohort from two European countries (Sweden and Italy). Annual modeled outdoor concentrations of nitrogen oxides (NO_x) at the study participants baseline home-address were available from the European Study of Cohorts for Air Pollution Effects (ESCAPE) project and biological samples were available from the Genomics Biomarkers of Environmental Health project (EnviroGenoMarkers). In **Chapter 2**, perturbation in plasma concentration of a large panel of inflammatory markers (N=28) was studied. Long-term exposure to NO_x was associated with decreased levels of four inflammatory markers: interleukin (IL)-2, IL-8, IL-10 and tumor necrosis factor-α in Italy, but not in Sweden. Effects were stronger in the Italian part of the cohort than the Swedish, potentially caused by the difference in air pollution levels between the two cohorts. These results might contribute to future elucidation of the pathways through which long-term exposure to air pollution induced adverse health effects.
In **Chapter 3**, we studied the alteration in genome-wide gene expression associated with long-term average exposure to NO_x. We applied several statistical methods to reduce the risk of false positive findings in this high dimensional data analysis. Associations were observed between exposure to NO_x and seven gene transcripts: A_23_P252075 ($AHCYL2$), A_24_P406830 ($MTMR2$), A_32_P175313 ($TNRC6B$), A_32_P44961 ($LARP1B$), A_32_P156373, A_32_P61298, and A_23_P217280 ($NOX1$). For genes $AHCYL2$ and $MTMR2$, changes were also seen in the DNA-methylome. We observed enrichment between genes downregulated due to cigarette smoking and the most strongly downregulated genes from our study. This observation points toward a shared biologic pathway of the effects of cigarette smoke and air pollution on the transcriptome which is of interest due to overlap between health outcomes that have been related to tobacco smoking and air pollution. Our results contribute to the further elucidation of the pathways through which long-term exposure to air pollution induces adverse health effects.
In **Chapter 4**, we address the acute effect of exposure to different air pollutions. Personal and ambient exposure measurements of particulate matter ($PM_{2.5}$), $PM_{2.5}$ absorbance, and ultrafine particles (UFP) and peripheral blood samples were collected from healthy non-smoking adults (N=157) living in four European countries. We investigated the association of personal and ambient air pollution exposure measures with genome-wide DNA-methylation perturbations at single CpG (cytosine-guanine dinucleotide) sites and a selected number of immune markers. Positive associations were observed between different immune markers and personal $PM_{2.5}$ and UFP concentrations. There were also several CpG sites that were significantly associated with personal $PM_{2.5}$ concentration. These analyses were extended by looking at differentially methylated regions (DMRs),

where we identified several associations between both air pollutants and DMRs. Analysis of differentially methylated regions provides a promising avenue to further explore the subtle impact of environmental exposures on DNA-methylation.

In **Chapter 5**, interplay between air pollution, immune markers, and adult-onset asthma was investigated following the meet-in-the-middle concept. We considered long-term exposure to multiple air pollutions (i.e. $PM_{2.5}$, PM_{10}, NO_2, UFP) and a panel of 13 immune markers in peripheral blood samples collected from adults (N=339) using a case-control study nested in the prospective SAPALDIA cohort. Using group level mediation analysis (partial least square path modeling), we observed that 15% of the effect of UFP on adult-onset asthma could be explained by the measured immune modulation. This points towards the possibility that the measured immune markers might play a role in the biological pathway that connects exposure to UFP to adult-onset asthma. This would also point towards a role of these immune markers as biomarkers of early disease for asthma.

While generating useful insights my studies also highlighted that there are challenges for the successful application of OMICS markers to study the internal exposome. These primarily pertain to increasing the sample size of studies, improving the quality of exposure assessment, and developing new methods for statistical analysis and integration between multiple biological (OMICS) layers.

In conclusion, the research described in this thesis provides evidence for perturbations of different OMICS markers in associations with both short-term and long-term exposure to different markers of air pollution. Observed perturbations in OMICS marker in association with air pollution exposure might play a role in the biological pathway that connects air pollution to health effects. However, the specific exposure-OMICS associations reported in this dissertation require replication before robust inferences can be drawn about effects. Priorities for future work include increasing sample size, improvement in the interpretation of functionality of identified OMICS markers, and understanding the integration of multiple OMICS layers.

7.2 Samenvatting

Het merendeel van hedendaagse chronische ziekten is waarschijnlijk het gevolg van een combinatie van blootstelling aan het milieu en erfelijke aanleg. Het Exposoom paradigma omvat het bestuderen van alle blootstellingen aan het milieu (bijvoorbeeld luchtverontreiniging, chemische verontreinigingen, voedings- of levensstijlfactoren) en daarmee samenhangende biologische reacties vanaf de conceptie tot de dood. Het interne Exposoom wordt gedefinieerd als de milieublootstellingen die in het lichaam meetbaar zijn en de biologische veranderingen die samenhangen met deze blootstellingen. Het interne Exposoom kan (deels) gemeten worden met behulp van zogeheten OMICS-technologien. Deze metingen kunnen direct informatie geven over blootstelling of over fysiologische verstoringen die wijzen op een bepaalde blootstelling. Het bestuderen van het interne Exposoom kan daarmee een biologische onderbouwing bieden voor empirisch waargenomen relaties tussen blootstelling en ziekte.
In dit proefschrift hebben we ons gericht op luchtverontreiniging, een milieuvervuiling met hoge maatschappelijke relevantie. Jaarlijks sterven wereldwijd 7 miljoen mensen vroegtijdig ten gevolge van luchtverontreiniging. Luchtverontreiniging is een erkend humaan carcinogeen en is causaal gerelateerd aan longkanker. Daarnaast is het een belangrijke risicofactor voor andere acute en chronische ziekten, o.a. hart- en vaatziekten, en chronische aandoeningen van de luchtwegen, zoals astma en COPD (chronische obstructieve long ziekte). Hoewel talrijke epidemiologische studies consistente associaties hebben aangetoond tussen blootstelling aan luchtverontreiniging en gezondheidseffecten, worden de onderliggende biologische mechanismen van deze effecten nog niet volledig begrepen. Over het algemeen worden oxidatieve stress en immuun- en ontstekingsreacties als mogelijke onderliggende mechanismen beschouwd. Verder zijn er wetenschappelijke vragen over welke bestanddelen van luchtverontreiniging deze biologische processen en de daarmee samenhangende ziekten veroorzaken. Een voorbeeld is ultrafijn stof (UFP), waarvan wordt aangenomen dat het schadelijker is voor de gezondheid is dan grotere stofdeeltjes omdat het door haar geringe grootte diep in de luchtwegen kan doordringen en via het bloed de kans heeft andere organen te bereiken.
Het algemene doel van dit proefschrift is om drie OMICS-platforms (*proteomics*, *transcriptomics* en *DNA-methylomics*) in relatie tot blootstelling aan luchtvervuiling te bestuderen om zo beter inzicht te krijgen in de biologische processen die luchtvervuiling en gezondheidseffecten met elkaar verbinden. Als dit succesvol is kunnen dergelijke analyses bijdragen aan het genereren van kennis over gezondheidseffecten van luchtverontreiniging, het identificeren van specifieke biologische mechanismen en het vinden van biomerkers voor vroege stadia van ziekte welke kunnen bijdragen aan het ontwikkelen van preventieve maatregelen.
De specifieke doelen, elk beschreven in een hoofdstuk, waren:

- Het evalueren van de impact van langdurige blootstelling aan luchtverontreiniging

door gebruik te maken van twee OMICS methoden (*proteomics* en *transcriptomics*)

- Het onderzoeken van de relatie tussen 24 uurs persoonlijke blootstelling van verschillende luchtverontreinigingscomponenten en genoomwijde *DNA-methylomics*
- Het onderzoeken tot op welke hoogte ontstekingsmerkers de associatie tussen langdurige blootstelling aan luchtverontreiniging en astma bij volwassenen kunnen voorspellen

In de eerste twee hoofdstukken beschrijven we de chronische effecten van luchtverontreiniging op het biologisch systeem door het meten van OMICS biomerkers in een relatief groot cohort (N=500) uit twee Europese landen (Zweden en Itali). Jaarlijkse gemodelleerde buitenconcentraties van stikstofoxiden (NO_x) op het woonadres van de studiedeelnemers waren beschikbaar via het *European Study of Cohorts for Air Pollution Effects* (ESCAPE) project en de biologische monsters via het *Genomics Biomarkers of Environmental Health* project (EnviroGenoMarkers). In **hoofdstuk 2** worden de plasmaconcentraties van een groot aantal ontstekingsmerkers (n=28) beschreven in relatie tot luchtverontreiniging. Langdurige blootstelling aan NO_x was in onze studie gerelateerd aan vier ontstekingsmerkers: interleukin (IL)-2, IL-8, IL-10 en tumor necrosis factor-α. De effecten waren groter in het Italiaanse deel van het cohort dan in het Zweedse deel, mogelijk veroorzaakt door de verschillen in luchtverontreiniging tussen de beide cohorten. De resultaten van dit onderzoek dragen bij aan onze kennis over hoe langdurige blootstelling aan luchtverontreiniging kan leiden tot nadelige gezondheidseffecten.
In **hoofdstuk 3** beschrijven we de verandering in genoomwijde *transcriptomics* in relatie tot langdurige blootstelling aan NO_x. Om het risico op vals-positieve resultaten te verminderen hebben we verschillende statistische methoden toegepast. We vonden relaties tussen blootstelling aan NO_x en 7 transcriptiefactoren: A_23_P252075 ($AHCYL2$ gene), A_24_P406830 ($MTMR2$), A_32_P175313 ($TNRC6B$), A_32_P44961 ($LARP1B$), A_32_P156373 (annotatie onbekend), A_32_P61298 (annotatie onbekend), en A_23_P217280 ($NOX1$). Voor de genen $AHCYL2$ en $MTMR2$ werden er ook veranderingen gevonden in het DNA-methyloom. Interessant was dat genen die in het verleden in verband zijn gebracht met roken verrijkt waren in de resultaten van ons onderzoek. Dit duidt op een mogelijke overeenkomst in biologische mechanismen van de effecten van sigarettenrook en luchtverontreiniging. Dit is van belang vanwege de al eerder geobserveerd overlap in gezondheidseffecten van sigarettenrook en luchtverontreiniging. De resultaten dragen dan ook bij aan de kennis over de oorzaak en gevolgen van langdurige blootstelling aan luchtverontreiniging en gezondheid.
In **hoofdstuk 4** beschrijven we de acute effecten van blootstelling aan verschillende vormen van luchtverontreiniging en het effect op genoomwijde DNA-methylatie en immuunmerkers. Persoonlijke en omgevingsblootstellingsmetingen van deeltjes ($PM_{2.5}$), $PM_{2.5}$-absorptie en UFP en bloedmonsters werden verzameld van gezonde niet-rokende volwassenen (N = 157) in vier Europese landen. Een positief verband werd aangetoond

tussen verschillende immuunmerkers en persoonlijke $PM_{2.5}$- en UFP-concentraties. Er waren ook verschillende CpG-plaatsen die significant gerelateerd waren aan de persoonlijke $PM_{2.5}$ blootstelling. Deze analyses werden uitgebreid door te kijken naar differentieel gemethyleerde gebieden (DMR's), waar we verschillende verbanden vonden tussen luchtverontreinigende stoffen en DMR's. De analyse van differentieel gemethyleerde gebieden biedt een veelbelovende mogelijkheid om de subtiele impact van milieublootstellingen op DNA-methylatie verder te onderzoeken.

In **hoofdstuk 5** werd het samenspel van luchtvervuiling, immuunmerkers en astma op volwassenleeftijd onderzocht volgens het *meet-in-the-middle-concept*. We hebben gekeken naar de langdurige blootstelling aan meerdere vormen van luchtverontreiniging (d.w.z. $PM_{2.5}$, PM_{10}, NO_2, UFP) en een panel van 13 immuunmerkers in perifere bloedmonsters verzameld bij volwassenen (N = 339) die waren gencludeerd in een case-control studie genest in het prospectieve SAPALDIA-cohort. Met behulp van mediatie-analyse op groepsniveau (*partial least squares path* modelering), zagen we dat 15% van het effect van UFP op astma op volwassenleeftijd kon worden verklaard door de gemeten immuunverandering. Dit wijst op de mogelijkheid dat de gemeten immuunmerkers een rol zouden kunnen spelen in de biologische route die blootstelling aan UFP relateert met astma bij volwassenen. Dit zou ook wijzen op een rol van deze immuunmerkers als biomerkers van vroege astma.

Naast deze nuttige inzichten, laten mijn studies ook zien dat er uitdagingen zijn voor de succesvolle toepassing van OMICs-merkers om het interne Exposoom te bestuderen. Deze hebben voornamelijk betrekking op het vergroten van de steekproefomvang van de onderzoeken, het verbeteren van de kwaliteit van de blootstellingskarakterisering en het ontwikkelen van nieuwe methoden voor statistische analyse en integratie tussen meerdere biologische (OMICS) lagen.

Concluderend, het onderzoek beschreven in dit proefschrift levert bewijs voor veranderingen van het biologisch-systeem als gemeten met OMICS-merkers in relatie tot zowel kortstondige als langdurige blootstelling aan verschillende vormen luchtverontreiniging. De specifieke blootstellings-OMICS-associaties die in dit proefschrift worden gerapporteerd, vereisen echter replicatie voordat robuuste conclusies over effecten kunnen worden getrokken. Prioriteiten voor toekomstige werkzaamheden zijn het vergroten van de steekproefomvang, verbetering van de interpretatie van functionaliteit van gedentificeerde OMICS-merkers en inzicht in de integratie van meerdere OMICS-lagen.

7.3 Affiliation of contributors

Jelle Vlaanderen, Marc Chadeau Hyam, Rob Beelen†, Gerard Hoek, Roel Vermeulen, Lutzen Portengen, Bert Brunekreef, Erik van Nunen
Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, 3584 CM, Utrecht, the Netherlands
Bert Brunekreef
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht
Paolo Vineis, Marc Chadeau Hyam, John Gulliver, David Morley
Medical Research Council Health Protection Agency Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom
Paolo Vineis, Sonia Tarallo, Alessio Naccarati
Italian Institute for Genomic Medicine (IIGM), Turin, Italy
Lars Modig, Ingvar A. Bergdahl
Department of Public Health and Clinical Medicine, Occupational and Environmental Medicine, University of Umea, 901 87,Umea, Sweden
Soterios A. Kyrtopoulos, Panagiotis Georgiadis
National Hellenic Research Foundation, Institute of Biology, Pharmaceutical Chemistry and Biotechnology, Athens, Greece
Domenico Palli
Molecular and Nutritional Epidemiology Unit, Cancer Prevention and Research Institute (ISPO), Florence, Italy
Dennie G.A.J. Hebels, Jos C.S. Kleinjans
Department of Toxicogenomics, Maastricht University, Universitiessingel 50 6229 Maastricht, The Netherlands
Vittorio Krogh
Epidemiology Unit, Instituto Tumori, Milan, Italy
Akram Ghantous, Zdenko Herceg
Epigenetics Group, International Agency for Research on Cancer, Lyon, France;
Nicole Probst Hensch, Medea Imboden, Ayoung Jeong
Swiss Tropical and Public Health (TPH) Institute, University of Basel, Basel, Switzerland
Manolis Kogevinas
Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

† Deceased

7.4 Publication list

This thesis

- **Mostafavi N**, Vlaanderen J, Chadeau-Hyam M, Beelen R, Modig L, Palli D, Bergdahl IA, Vineis P, Hoek G, Kyrtopoulos S, Vermeulen R. Inflammatory markers in relation to long-term air pollution. *Environment International*, 2015.
- **Mostafavi N**, Vlaanderen J, Portengen L, Chadeau-Hyam M, Modig L, Palli D, Bergdahl IA, Brunekreef B, Vineis P, Hebels D.G.A.J, Kleinjans J.C.S, Krogh V, Hoek G, Georgiadis P, Kyrtopoulos S, Vermeulen R. Associations Between Genome-wide Gene Expression and Ambient Nitrogen Oxides. *Epidemiology*, 2016.
- **Mostafavi N**, Vermeulen R, Ghantous A, Hoek G, Probst-Hensch N, Herceg Z, Tarallo S, Naccarati A, Kleinjans J.C.S, Imboden M, Jeong A, Morley D, Kogevinas M, van Nunen E, Gulliver J, Chadeau-Hyam M, Vineis P, and Vlaanderen J. Acute changes in DNA methylation in relation to 24hr personal air pollution exposure measurements: A panel study in four European countries. *Environment International*, 2018.
- **Mostafavi N**, Jeong A, Vlaanderen J, Imboden M, Vineis P, Jarvis D, Kogevinas M, Probst-Hensch N, Vermeulen R. The mediating effect of immune markers on the association between ambient air pollution and adult-onset asthma. *Scientific reports*, (submitted 2018).

Other publications

- Sikorska K, **Mostafavi N**, Uitterlinden A, Rivadeneira F, Eilers PH, Lesaffre E. GWAS with longitudinal phenotypes: performance of approximate procedures. *European Journal of Human Genetics*, 2015.
- Jeong A, Fiorito G, Keski-Rahkonen P, Imboden M, Kisse A, Robinot N, Gmuender H, Vlaanderen J, Vermeulen R, Kyrtopoulos S, Herceg Z, Ghantous A, Lovison G, Galassi C, Ranzi A, Krogh V, Grioni S, Agnoli C, Sacerdote C, **Mostafavi N**, Naccarati A, Scalberte A, Vineis P, Probst-Hensch N. Perturbation of metabolic pathways mediates the association of air pollutants with asthma and cardiovascular diseases. *Environment International*, 2018.
- Westra WM, Lutzke LS, **Mostafavi N**, Roes AL, Calpe S, Wang KK, Krishnadath KK. Smokeless Tobacco and Cigar and/or Pipe Are Risk Factors for Barrett Esophagus in Male Patients With Gastroesophageal Reflux Disease. *Mayo Clinic Proceedings*, 2018.
- Bergman JJ, De Groof AJ, Pech O, Ragunath K, Armstrong D, **Mostafavi N**, Lundell L, Dent J, Vieth M, Tytgat GN, Sharma P. An interactive web-based educational

tool improves detection and delineation of Barretts oesophagus related neoplasia (BORN): The BORN Project. *Gastroenterology*, 2019.

- Lowenberg M, Vermeire S, **Mostafavi N**, Hoentjen F, Franchimont D, Bossuyt P, Hindryckx P, Rispens T, De Vries A, van der Woude CJ, Berends S, Ambarus CA, Mathot R, Clasquin E, Baert F, DHaens G. Mucosal healing with vedolizumab and associations with serum levels in Crohns disease patients. *Gastroenterology*, (Submitted, 2018).
- Belghazi K, Pouw RE, Koch AD, Weusten BLAM, Schoon EJ, Curvers WL, Gotink AW, **Mostafavi N**, Haidry RJ, Pech O, Bergman JJ, Bisschops R. Self-sizing radiofrequency ablation balloon for eradication of Barrett's esophagus: results of an international multicenter randomized trial comparing three different treatment regimens. *Gut*, (Submitted, 2019).

7.5 Acknowledgements

The Ph.D. path was a great adventure for me. A period of hard work and beautiful memories. Many people have contributed to making these years exciting and joyful. I am very grateful to all those who inspired, encouraged, supported, and taught me during this journey.

First and foremost, I would like to express my sincere gratitude to my promotors; Prof. dr. Bert Brunekreef and Prof. dr. Roel Vermeulen, thank you for giving me the opportunity to pursue my Ph.D. at your department. Dear Bert, you are a dedicated person to your team, constantly supporting them with positive energy and high optimism. I am forever thankful for your encouragement and support. Your guidance and feedback were always valuable input for this thesis. Dear Roel, thank you very much for trusting in me and involving me in different exciting projects. You have a special ability to motivate people to work hard. When I look at this thesis, I realize the importance of your positive criticisms and valuable suggestions on my work. I enjoyed our scientific discussions that brought insight to my perspective and shaped my ideas.

I am particularly grateful for the assistance given by my co-promotor, dr. Jelle Vlaanderen. Dear Jelle, thank you for all the long hours of insightful discussions. Your detailed explanation on every stage of the projects enabled me to fill the gaps on my background knowledge in environmental epidemiology smoothly. I am very grateful not only for your concrete guidance but also for your compassion and understanding through all the struggles. Thank you for being so professional and kind at the same time.

My special gratitude is extended to Prof. dr. Emmanuel Lesaffre. Dear Emmanuel, I started my scientific journey in the Netherlands under your supervision, you have been supporting me continuously and unconditionally. For that I am forever indebted to you and I will always consider you as my scientific father. It is my great honor to have you in my Ph.D. committee.

I would like to express my appreciation to other Ph.D. committee members; dear Prof. dr. E.A.M. Sanders, Prof. dr. M.J.C. Eijkemans, Prof. dr. J.H. Veldink, Prof. dr. T. de Kok, and Prof. dr. R.F. Cassee, thank you for taking the time to review and evaluate my thesis manuscript. It is a great privilege to have you as a member of my Ph.D. committee.

I would like to take this opportunity and thank all my co-authors in particular dr. Lutzen Portengen, dr. Gerard Hoek, Prof. dr. Paolo Vineis, and Prof. dr. Marc Chadeau-Hyam, for sharing your time and invaluable knowledge. Thank you for your fruitful and pleasant collaboration.

My sincere thanks are extended to Myrna de Rooij and Daniella van de Langenberg for accepting to be my paranymphs. Dear Daniella, such a long history and sweet moments we had together over the past six years! You are one of the coolest friends I ever had. Thanks for being supportive when times were not all that great! Dear Myrna, I admire your energy and passion in whatever you do and the smile you never lose! Thank you for

being so kind, caring and supportive!
Thanks to my wonderful (former and current) roommates, our office was always *gezellig*. Virissa Lenters and Maartje Brouwer thank you for all chats and friendly advices. Erik van Nunen, I enjoyed sharing a project with you. Luuk van Wel, I appreciate your friendly attitude and support. Thank you for the kind and motivating talks at coffee breaks. Mariana Simoes, your friendship means the world to me. I always enjoy every moment that I spend with you. Thank you for being such a great friend and listener! Liese Van Gompel, Jeroen Meijer, and Warner van Kersen, thank you for your company. It was a pleasure working with you all. Thank you!
The list of good friends and colleagues goes even further at IRAS in Utrecht. Calvin Ge, Marije Reedijk, Alejandro Dorado, Daniel Figueiredo, George Downward, Roosmarijn Luiken, Gijs Klous, Joseph Bukalasa, Hicham Zilaout, Fatemeh Saberi, Jose Jacobs, Wietske Dohmen, Maciek Strak, Aniek Lotterman, many thanks for the chats, outings, dinners, lunch walks and drinks!
My special thanks to Ingrid Dahmen, Christina Kneefel, and Mieke Lumens for your help in practical matters! Dear Ingrid, I appreciate your continues assistance whenever I needed it and also for the guidance in the preparations for my defense.
I would like to thank my former colleagues in the Dep. of Biostatistics at Erasmus MC in Rotterdam with whom I have shared pleasant moments. Especially, Prof. dr. Emmanuel Lesaffre, Prof. dr. Paul Eilres, Prof. dr. Dimitris Rizopoulos, Joost van Rosmalen, Marek Molas, Sten Willemsen, Veronika Rockova, Bayoue Li, Johan de Rooi, Karolina Sikorska, Susan Bryan, Elrozy Andrinopoulou, Nicole Erler, Ehsan Motazedi, and Eline van Gent, thank you for all the great and sweet memories.
My special thanks are further extended to my current colleagues at Amsterdam UMC. Dear Prof. dr. Paul Fockens, dear Prof. dr. Geert DHaens, and dear Prof. dr. Koos Zwinderman, it is a great pleasure working with you in such a professional and friendly atmosphere.
I wish to thank my parents in-law for their kindness and encouragement through every new step. Dear Maryam, Atefeh and Sadegh, my mind is full of the good memories with you. Atefeh, thanks for designing the cover of this thesis.
My deepest gratitude goes to my parents who supported me throughout all my life. Thank you for being there for me no matter what, for raising me in the warmest loving environment! My dear brothers, Amir and Ali, thanks for all your kindness, support, and company.
Finally, my dearest Kazem (Alireza), thank you for your endless support and incredible patience! We started this journey together and have been each other's motivation, support, and listeners through all the ups and downs. You are truly my best friend!

Nahid Mostafavi
May 2019 - Rotterdam

7.6 About the author

Nahid Mostafavi was born on August 14, 1985 in Tehran. After finishing her high school she started her Bachelor in Mathematics in Shahid Beheshti University. After following few statistical courses she decided to continue her study in Statistics instead and obtained her BSc degree in 2007. She defended her Master thesis entitled "Using BLUP to estimate random effects in linear mixed-effects model, with an application in medical data" and obtained her MSc degree in 2011.
In 2012, she had an opportunity to join the Dep. of Biostatistics at Erasmus MC Rotterdam, where she was introduced to Bayesian statistics and Joint models.
Out of interest in applied statistics and analyzing high dimensional data in 2013, she started her Ph.D. program under supervision of Prof. dr. Bert Brunekreef and Prof. dr. Roel Vermeulen at Division of environmental epidemiology, the Institute for Risk Assessment Science (IRAS), in Utrecht University.
Since March 2018, she has been employed as a statistician at the Dep. of Gastroenterology in Amsterdam University Medical Center.

BIBLIOGRAPHY

Ackermann-Liebrich, U., Kuna-Dibbert, B., Probst-Hensch, N. M., Schindler, C., Dietrich, D. F., Stutz, E. Z., Bayer-Oglesby, L., Baum, F., Brändli, O., Brutsche, M., et al. (2005). Follow-up of the swiss cohort study on air pollution and lung diseases in adults (sapaldia 2) 1991–2003: methods and characterization of participants. *Sozial-und Präventivmedizin/Social and Preventive Medicine*, 50(4):245–263.

Adam, M., Imboden, M., Boes, E., Schaffner, E., Künzli, N., Phuleria, H. C., Kronenberg, F., Gaspoz, J.-M., Carballo, D., and Probst-Hensch, N. (2014). Modifying effect of a common polymorphism in the interleukin-6 promoter on the relationship between long-term exposure to traffic-related particulate matter and heart rate variability. *PloS one*, 9(8):e104978.

Adam, M., Schikowski, T., Carsin, A. E., Cai, Y., Jacquemin, B., Sanchez, M., Vierkötter, A., Marcon, A., Keidel, D., Sugiri, D., et al. (2015). Adult lung function and long-term air pollution exposure. escape: a multicentre cohort study and meta-analysis. *European Respiratory Journal*, 45(1):38–50.

Agier, L., Portengen, L., Chadeau-Hyam, M., Basagaña, X., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J. R., Nieuwenhuijsen, M. J., et al. (2015). A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. In *27th Conference of the International Society for Environmental Epidemiology. São Paulo*.

Ainsworth, B. E., Haskell, W. L., Whitt, M. C., Irwin, M. L., Swartz, A. M., Strath, S. J., O Brien, W. L., Bassett, D. R., Schmitz, K. H., Emplaincourt, P. O., et al. (2000). Compendium of physical activities: an update of activity codes and met intensities. *Medicine and science in sports and exercise*, 32(9; SUPP/1):S498–S504.

Alfaro-Moreno, E., Torres, V., Miranda, J., Martínez, L., García-Cuellar, C., Nawrot, T. S., Vanaudenaerde, B., Hoet, P., Ramírez-López, P., Rosas, I., et al. (2009). Induction of il-6 and inhibition of il-8 secretion in the human airway cell line calu-3 by urban particulate matter collected with a modified method of pm sampling. *Environmental research*, 109(5):528–535.

Alvarado-Cruz, I., Sánchez-Guerra, M., Hernández-Cadena, L., De Vizcaya-Ruiz, A., Mugica, V., Pelallo-Martínez, N. A., de Jesús Solís-Heredia, M., Byun, H.-M., Baccarelli,

A., and Quintanilla-Vega, B. (2017). Increased methylation of repetitive elements and dna repair genes is associated with higher dna oxidation in children in an urbanized, industrial environment. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 813:27–36.

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). Mckusick's online mendelian inheritance in man (omim®). *Nucleic acids research*, 37(suppl 1):D793–D796.

Araujo, J. A., Barajas, B., Kleinman, M., Wang, X., Bennett, B. J., Gong, K. W., Navab, M., Harkema, J., Sioutas, C., Lusis, A. J., et al. (2008). Ambient particulate pollutants in the ultrafine range promote early atherosclerosis and systemic oxidative stress. *Circulation research*, 102(5):589–596.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Auerbach, A. and Hernandez, M. L. (2012). The effect of environmental oxidative stress on airway inflammation. *Current Opinion in Allergy and Clinical Immunology*, 12(2):133.

Baccarelli, A., Wright, R. O., Bollati, V., Tarantini, L., Litonjua, A. A., Suh, H. H., Zanobetti, A., Sparrow, D., Vokonas, P. S., and Schwartz, J. (2009). Rapid dna methylation changes after exposure to traffic particles. *American journal of respiratory and critical care medicine*, 179(7):572–578.

Baglietto, L., Ponzi, E., Haycock, P., Hodge, A., Bianca Assumma, M., Jung, C.-H., Chung, J., Fasanelli, F., Guida, F., Campanella, G., et al. (2017). Dna methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *International journal of cancer*, 140(1):50–61.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2013). Ncbi geo: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., Marcon, A., et al. (2013). Development of no 2 and no x land use regression models for estimating air pollution exposure in 36 study areas in europe–the escape project. *Atmospheric Environment*, 72:10–23.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., et al. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project. *The Lancet*, 383(9919):785–795.

Beineke, P., Fitch, K., Tao, H., Elashoff, M. R., Rosenberg, S., Kraus, W. E., and Wingrove, J. A. (2012). A whole blood gene expression-based signature for smoking status. *BMC medical genomics*, 5(1):1.

Bellavia, A., Urch, B., Speck, M., Brook, R. D., Scott, J. A., Albetti, B., Behbod, B., North, M., Valeri, L., Bertazzi, P. A., et al. (2013). Dna hypomethylation, ambient particulate matter, and increased blood pressure: findings from controlled human exposure experiments. *Journal of the American Heart Association*, 2(3):e000212.

Belshaw, N. J., Pal, N., Tapp, H. S., Dainty, J. R., Lewis, M. P., Williams, M. R., Lund, E. K., and Johnson, I. T. (2010). Patterns of dna methylation in individual colonic crypts reveal aging and cancer-related field defects in the morphologically normal mucosa. *Carcinogenesis*, 31(6):1158–1163.

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)*, pages 289–300.

Bentayeb, M., Simoni, M., Baiz, N., Norback, D., Baldacci, S., Maio, S., Viegi, G., Annesi-Maesano, I., in Europe on Health Effects of Air Quality in Nursing Homes (GERIE) Group, G. S., et al. (2012). Adverse respiratory effects of outdoor air pollution in the elderly [review article]. *The International Journal of Tuberculosis and Lung Disease*, 16(9):1149–1161.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., and Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(2):S15.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., et al. (2011). High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295.

Bind, M.-A., Coull, B., Suh, H., Wright, R., Baccarelli, A., Vokonas, P., and Schwartz, J. (2014a). A novel genetic score approach using instruments to investigate interactions between pathways and environment: application to air pollution. *PloS one*, 9(4):e96000.

Bind, M.-A., Lepeule, J., Zanobetti, A., Gasparrini, A., Baccarelli, A. A., Coull, B. A., Tarantini, L., Vokonas, P. S., Koutrakis, P., and Schwartz, J. (2014b). Air pollution and gene-specific methylation in the normative aging study: association, effect modification, and mediation analysis. *Epigenetics*, 9(3):448–458.

Bonassi, S., Taioli, E., and Vermeulen, R. (2013). Omics in population studies: a molecular epidemiology perspective. *Environmental and molecular mutagenesis*, 54(7):455–460.

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Zeller, T., Liquet, B., Newcombe, P., Yengo, L., Wild, P. S., Schillert, A., Ziegler, A., et al. (2013). Guess-ing polygenic associations with multiple phenotypes using a gpu-based evolutionary stochastic search algorithm. *PLoS Genet*, 9(8):e1003657.

Breton, C. V., Marsit, C. J., Faustman, E., Nadeau, K., Goodrich, J. M., Dolinoy, D. C., Herbstman, J., Holland, N., LaSalle, J. M., Schmidt, R., et al. (2017). Small-magnitude effect sizes in epigenetic end points are important in childrens environmental health studies: the childrens environmental health and disease prevention research centers epigenetics working group. *Environmental health perspectives*, 125(4):511.

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., et al. (2010). Particulate matter air pollution and cardiovascular disease an update to the scientific statement from the american heart association. *Circulation*, 121(21):2331–2378.

Brunekreef, B. and Holgate, S. T. (2002). Air pollution and health. *The lancet*, 360(9341):1233–1242.

Bruniquel, D. and Schwartz, R. H. (2003). Selective, stable demethylation of the interleukin-2 gene enhances transcription by an active process. *Nature immunology*, 4(3):235.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3).

Cai, Y., Zijlema, W. L., Doiron, D., Blangiardo, M., Burton, P. R., Fortier, I., Gaye, A., Gulliver, J., de Hoogh, K., Hveem, K., et al. (2017). Ambient air pollution, traffic noise and adult asthma prevalence: a bioshare approach. *European Respiratory Journal*, 49(1):1502127.

Cesaroni, G., Porta, D., Badaloni, C., Stafoggia, M., Eeftens, M., Meliefste, K., and Forastiere, F. (2012). Nitrogen dioxide levels estimated from land use regression models several years apart and association with mortality in a large cohort study. *Environmental Health*, 11(1):1.

Chadeau-Hyam, M., Athersuch, T. J., Keun, H. C., De Iorio, M., Ebbels, T. M., Jenab, M., Sacerdote, C., Bruce, S. J., Holmes, E., and Vineis, P. (2011). Meeting-in-the-middle using metabolic profiling–a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers*, 16(1):83–88.

Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B., and Vermeulen, R. C. (2013). Deciphering the complex: Methodological

overview of statistical models to derive omics-based biomarkers. *Environmental and molecular mutagenesis*, 54(7):542–557.

Chadeau-Hyam, M., Vermeulen, R., Hebels, D., Castagné, R., Campanella, G., Portengen, L., Kelly, R., Bergdahl, I., Melin, B., Hallmans, G., et al. (2014). Prediagnostic transcriptomic markers of chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. *Annals of Oncology*, 25(5):1065–1072.

Chen, R., Meng, X., Zhao, A., Wang, C., Yang, C., Li, H., Cai, J., Zhao, Z., and Kan, H. (2016). Dna hypomethylation and its mediation in the effects of fine particulate air pollution on cardiovascular biomarkers: A randomized crossover trial. *Environment international*, 94:614–619.

Chi, G. C., Liu, Y., MacDonald, J. W., Barr, R. G., Donohue, K. M., Hensley, M. D., Hou, L., McCall, C. E., Reynolds, L. M., Siscovick, D. S., et al. (2016). Long-term outdoor air pollution and dna methylation in circulating monocytes: results from the multi-ethnic study of atherosclerosis (mesa). *Environmental Health*, 15(1):119.

Chuang, K.-J., Yan, Y.-H., Chiu, S.-Y., and Cheng, T.-J. (2010). Long-term air pollution exposure and risk factors for cardiovascular diseases among the elderly in taiwan. *Occupational and environmental medicine*, pages oem–2009.

Chung, F. (2001). Anti-inflammatory cytokines in asthma and allergy: interleukin-10, interleukin-12, interferon-γ. *Mediators of inflammation*, 10(2):51–59.

Clark, I. A. (2007). How tnf was recognized as a key mechanism of disease. *Cytokine & growth factor reviews*, 18(3-4):335–343.

Consortium, U. et al. (2014). Uniprot: a hub for protein information. *Nucleic acids research*, page gku989.

Cui, Y., Balshaw, D. M., Kwok, R. K., Thompson, C. L., Collman, G. W., and Birnbaum, L. S. (2016). The exposome: embracing the complexity for discovery in environmental health. *Environmental health perspectives*, 124(8):A137.

Cyrys, J., Eeftens, M., Heinrich, J., Ampe, C., Armengaud, A., Beelen, R., Bellander, T., Beregszaszi, T., Birk, M., Cesaroni, G., et al. (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.

Cyrys, J., Heinrich, J., Hoek, G., Meliefste, K., Lewné, M., Gehring, U., Bellander, T., Fischer, P., van Vliet, P., Brauer, M., et al. (2003). Comparison between different traffic-related particle indicators: elemental carbon (ec), pm 2.5 mass, and absorbance. *Journal of Exposure Science and Environmental Epidemiology*, 13(2):134.

Dai, L., Mehta, A., Mordukhovich, I., Just, A. C., Shen, J., Hou, L., Koutrakis, P., Sparrow, D., Vokonas, P. S., Baccarelli, A. A., et al. (2017). Differential dna methylation and pm2. 5 species in a 450k epigenome-wide association study. *Epigenetics*, 12(2):139–148.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., Partchev, I., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, 39(12):1–28.

de FC Lichtenfels, A. J., Van Der Plaat, D. A., de Jong, K., van Diemen, C. C., Postma, D. S., Nedeljkovic, I., van Duijn, C. M., Amin, N., La Bastide-van Gemert, S., De Vries, M., et al. (2018). Long-term air pollution exposure, genome-wide dna methylation and lung function in the lifelines cohort study. *Environmental health perspectives*, 126(2):027004.

de Hoogh, K., Gulliver, J., van Donkelaar, A., Martin, R. V., Marshall, J. D., Bechle, M. J., Cesaroni, G., Pradas, M. C., Dedele, A., Eeftens, M., et al. (2016). Development of west-european pm2. 5 and no2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental research*, 151:1–10.

De Marco, R., Poli, A., Ferrari, M., Accordini, S., Giammanco, G., Bugiani, M., Villani, S., Ponzio, M., Bono, R., Carrozzi, L., et al. (2002). The impact of climate and traffic-related no2 on the prevalence of asthma and allergic rhinitis in italy. *Clinical & Experimental Allergy*, 32(10):1405–1412.

De Prins, S., Koppen, G., Jacobs, G., Dons, E., Van de Mieroop, E., Nelen, V., Fierens, F., Panis, L. I., De Boever, P., Cox, B., et al. (2013). Influence of ambient air pollution on global dna methylation in healthy adults: a seasonal follow-up. *Environment international*, 59:418–424.

DeBord, D. G., Carreón, T., Lentz, T. J., Middendorf, P. J., Hoover, M. D., and Schulte, P. A. (2016). Use of the exposome in the practice of epidemiology: a primer on-omic technologies. *American journal of epidemiology*, 184(4):302–314.

Demetriou, C. A., Raaschou-Nielsen, O., Loft, S., Møller, P., Vermeulen, R., Palli, D., Chadeau-Hyam, M., Xun, W. W., and Vineis, P. (2012). Biomarkers of ambient air pollution and lung cancer: a systematic review. *Occupational and environmental medicine*, pages oemed–2011.

Ding, R., Jin, Y., Liu, X., Ye, H., Zhu, Z., Zhang, Y., Wang, T., and Xu, Y. (2017). Dose-and time-effect responses of dna methylation and histone h3k9 acetylation changes induced by traffic-related air pollution. *Scientific reports*, 7:43737.

Dobreva, Z. G., Miteva, L. D., and Stanilova, S. A. (2009). The inhibition of jnk and p38 mapks downregulates il-10 and differentially affects c-jun gene expression in human monocytes. *Immunopharmacology and immunotoxicology*, 31(2):195–201.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587.

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. *Environmental health perspectives*, pages 992–998.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., de Nazelle, A., et al. (2012a). Development of land use regression models for pm2. 5, pm2. 5 absorbance, pm10 and pmcoarse in 20 european study areas; results of the escape project. *Environmental science & technology*, 46(20):11195–11205.

Eeftens, M., Beelen, R., Fischer, P., Brunekreef, B., Meliefste, K., and Hoek, G. (2011). Stability of measured and modelled spatial contrasts in no2 over time. *Occupational and environmental medicine*, pages oem–2010.

Eeftens, M., Meier, R., Schindler, C., Aguilera, I., Phuleria, H., Ineichen, A., Davey, M., Ducret-Stich, R., Keidel, D., Probst-Hensch, N., et al. (2016). Development of land use regression models for nitrogen dioxide, ultrafine particles, lung deposited surface area, and four other markers of particulate matter pollution in the swiss sapaldia regions. *Environmental Health*, 15(1):53.

Eeftens, M., Tsai, M.-Y., Ampe, C., Anwander, B., Beelen, R., Bellander, T., Cesaroni, G., Cirach, M., Cyrys, J., de Hoogh, K., et al. (2012b). Spatial variation of pm 2.5, pm 10, pm 2.5 absorbance and pm coarse concentrations between and within 20 european study areas and the relationship with no 2–results of the escape project. *Atmospheric Environment*, 62:303–317.

Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, M., Grankvist, K., Johansson, M., Assumma, M. B., Naccarati, A., et al. (2015). Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nature communications*, 6:10192.

Fasoli, M., Dal Santo, S., Zenoni, S., Tornielli, G. B., Farina, L., Zamboni, A., Porceddu, A., Venturini, L., Bicego, M., Murino, V., et al. (2012). The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *The Plant Cell*, pages tpc–112.

Felix, J. F., Joubert, B. R., Baccarelli, A. A., Sharp, G. C., Almqvist, C., Annesi-Maesano, I., Arshad, H., Baïz, N., Bakermans-Kranenburg, M. J., Bakulski, K. M., et al. (2017).

Cohort profile: pregnancy and childhood epigenetics (pace) consortium. *International journal of epidemiology*, 47(1):22–23u.

Fiorito, G., Vlaanderen, J., Polidoro, S., Gulliver, J., Galassi, C., Ranzi, A., Krogh, V., Grioni, S., Agnoli, C., Sacerdote, C., et al. (2018). Oxidative stress and inflammation mediate the effect of air pollution on cardio-and cerebrovascular disease: A prospective study in nonsmokers. *Environmental and molecular mutagenesis*, 59(3):234–246.

Forbes, L. J., Patel, M. D., Rudnicka, A. R., Cook, D. G., Bush, T., Stedman, J. R., Whincup, P. H., Strachan, D. P., and Anderson, R. H. (2009). Chronic exposure to outdoor air pollution and markers of systemic inflammation. *Epidemiology*, 20(2):245–253.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Georgiadis, P., Hebels, D. G., Valavanis, I., Liampa, I., Bergdahl, I. A., Johansson, A., Palli, D., Chadeau-Hyam, M., Chatziioannou, A., Jennen, D. G., et al. (2016). Omics for prediction of environmental health effects: Blood leukocyte-based cross-omic profiling reliably predicts diseases associated with tobacco smoking. *Scientific reports*, 6.

Ghantous, A., Saffery, R., Cros, M.-P., Ponsonby, A.-L., Hirschfeld, S., Kasten, C., Dwyer, T., Herceg, Z., and Hernandez-Vargas, H. (2014). Optimized dna extraction from neonatal dried blood spots: application in methylome profiling. *BMC biotechnology*, 14(1):60.

Ghio, A. J., Carraway, M. S., and Madden, M. C. (2012). Composition of air pollution particles and oxidative stress in cells, tissues, and living systems. *Journal of Toxicology and Environmental Health, Part B*, 15(1):1–21.

Gilmour, M. I., Jaakkola, M. S., London, S. J., Nel, A. E., and Rogers, C. A. (2006). How exposure to environmental tobacco smoke, outdoor air pollutants, and increased pollen burdens influences the incidence of asthma. *Environmental health perspectives*, 114(4):627.

González, I., Lê Cao, K.-A., Davis, M. J., and Déjean, S. (2012). Visualising associations between paired omics data sets. *BioData mining*, 5(1):19.

Gruzieva, O., Xu, C.-J., Breton, C. V., Annesi-Maesano, I., Antó, J. M., Auffray, C., Ballereau, S., Bellander, T., Bousquet, J., Bustamante, M., et al. (2017). Epigenome-wide meta-analysis of methylation in children related to prenatal no2 air pollution exposure. *Environmental health perspectives*, 125(1):104.

Guarnieri, M. and Balmes, J. R. (2014). Outdoor air pollution and asthma. *The Lancet*, 383(9928):1581–1592.

Guida, F., Sandanger, T. M., Castagné, R., Campanella, G., Polidoro, S., Palli, D., Krogh, V., Tumino, R., Sacerdote, C., Panico, S., et al. (2015). Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human molecular genetics*, 24(8):2349–2359.

Gulliver, J., de Hoogh, K., Hansell, A., and Vienneau, D. (2013). Development and back-extrapolation of no2 land use regression models for historic exposure assessment in great britain. *Environmental science & technology*, 47(14):7804–7811.

Hajat, A., Allison, M., Diez-Roux, A. V., Jenny, N. S., Jorgensen, N. W., Szpiro, A. A., Vedal, S., and Kaufman, J. D. (2015). Long-term exposure to air pollution and markers of inflammation, coagulation, and endothelial activation: a repeat-measures analysis in the multi-ethnic study of atherosclerosis (mesa). *Epidemiology (Cambridge, Mass.)*, 26(3):310.

Hallmans, G., Ågren, Å., Johansson, G., Johansson, A., Stegmayr, B., Jansson, J.-H., Lindahl, B., Rolandsson, O., Söderberg, S., Nilsson, M., et al. (2003). Cardiovascular disease and diabetes in the northern sweden health and disease study cohort-evaluation of risk factors and their interactions. *Scandinavian journal of public health*, 31(61 suppl):18–24.

Hamra, G. B., Laden, F., Cohen, A. J., Raaschou-Nielsen, O., Brauer, M., and Loomis, D. (2015). Lung cancer and exposure to nitrogen dioxide and traffic: a systematic review and meta-analysis. *Environmental health perspectives*, 123(11):1107–1112.

Hankinson, O. (1995). The aryl hydrocarbon receptor complex. *Annual review of pharmacology and toxicology*, 35(1):307–340.

Hassanvand, M. S., Naddafi, K., Kashani, H., Faridi, S., Kunzli, N., Nabizadeh, R., Momeniha, F., Gholampour, A., Arhami, M., Zare, A., et al. (2017). Short-term effects of particle size fractions on circulating biomarkers of inflammation in a panel of elderly subjects and healthy young adults. *Environmental Pollution*, 223:695–704.

Hayes, A. F. (2009). Beyond baron and kenny: Statistical mediation analysis in the new millennium. *Communication monographs*, 76(4):408–420.

Hebels, D. G., Georgiadis, P., Keun, H. C., Athersuch, T. J., Vineis, P., Vermeulen, R., Portengen, L., Bergdahl, I. A., Hallmans, G., Palli, D., et al. (2013). Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. *Environmental Health Perspectives (Online)*, 121(4):480.

Helm, S., Eggert, A., and Garnefeld, I. (2010). Modeling the impact of corporate reputation on customer satisfaction and loyalty using partial least squares. In *Handbook of partial least squares*, pages 515–534. Springer.

Herceg, Z., Ghantous, A., Wild, C. P., Sklias, A., Casati, L., Duthie, S. J., Fry, R., Issa, J.-P., Kellermayer, R., Koturbash, I., et al. (2017). Roadmap for investigating epigenome deregulation and environmental origins of cancer. *International journal of cancer*.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, 42(33):7561–7578.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental Health*, 12(1):1.

Hoffmann, B., Moebus, S., Dragano, N., Stang, A., Möhlenkamp, S., Schmermund, A., Memmesheimer, M., Bröcker-Preuss, M., Mann, K., Erbel, R., et al. (2009). Chronic residential exposure to particulate matter air pollution and systemic inflammatory markers. *Environmental Health Perspectives*, 117(8):1302.

Hofmann, J. N., Yu, K., Bagni, R. K., Lan, Q., Rothman, N., and Purdue, M. P. (2011). Intra-individual variability over time in serum cytokine levels among participants in the prostate, lung, colorectal, and ovarian cancer screening trial. *Cytokine*, 56(2):145–148.

Hosnijeh, F. S., Boers, D., Portengen, L., Bueno-de Mesquita, H. B., Heederik, D., and Vermeulen, R. (2012). Plasma cytokine concentrations in workers exposed to 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin (tcdd). *Frontiers in oncology*, 2.

Hosnijeh, F. S., Krop, E. J., Scoccianti, C., Krogh, V., Palli, D., Panico, S., Tumino, R., Sacredote, C., Nawroly, N., Portengen, L., et al. (2010). Plasma cytokines and future risk of non-hodgkin lymphoma (nhl): a case-control study nested in the italian european prospective investigation into cancer and nutrition. *Cancer Epidemiology and Prevention Biomarkers*, pages 1055–9965.

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(1):86.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57.

Huang, S.-L., Hsu, M.-K., and Chan, C.-C. (2003). Effects of submicrometer particle compositions on cytokine production and lipid peroxidation of human bronchial epithelial cells. *Environmental health perspectives*, 111(4):478.

Huang, Y.-C. T., Karoly, E. D., Dailey, L. A., Schmitt, M. T., Silbajoris, R., Graff, D. W., and Devlin, R. B. (2011). Comparison of gene expression profiles induced by coarse, fine, and ultrafine particulate matter. *Journal of Toxicology and Environmental Health, Part A*, 74(5):296–312.

Huang, Y.-C. T., Schmitt, M., Yang, Z., Que, L. G., Stewart, J. C., Frampton, M. W., and Devlin, R. B. (2010). Gene expression profile in circulating mononuclear cells after exposure to ultrafine carbon particles. *Inhalation toxicology*, 22(10):835–846.

Jacobsen, M. J., Mentzel, C. M. J., Olesen, A. S., Huby, T., Jørgensen, C. B., Barrès, R., Fredholm, M., and Simar, D. (2016). Altered methylation profile of lymphocytes is concordant with perturbation of lipids metabolism and inflammatory response in obesity. *Journal of diabetes research*, 2016.

Jacquemin, B., Schikowski, T., Carsin, A. E., Hansell, A., Krämer, U., Sunyer, J., Probst-Hensch, N., Kauffmann, F., and Künzli, N. (2012). The role of air pollution in adult-onset asthma: a review of the current evidence. In *Seminars in respiratory and critical care medicine*, volume 33, pages 606–619. Thieme Medical Publishers.

Jacquemin, B., Siroux, V., Sanchez, M., Carsin, A.-E., Schikowski, T., Adam, M., Bellisario, V., Buschka, A., Bono, R., Brunekreef, B., et al. (2015). Ambient air pollution and adult asthma incidence in six european cohorts (escape). *Environmental health perspectives*, 123(6):613.

Jeong, A., Fiorito, G., Keski-Rahkonen, P., Imboden, M., Kiss, A., Robinot, N., Gmuender, H., Vlaanderen, J., Vermeulen, R., Kyrtopoulos, S., et al. (2018). Perturbation of metabolic pathways mediates the association of air pollutants with asthma and cardiovascular diseases. *Environment international*, 119:334–345.

Ji, H. and Hershey, G. K. K. (2012). Genetic and epigenetic influence on the response to environmental particulate matter. *Journal of Allergy and Clinical Immunology*, 129(1):33–41.

Ji, H., Myers, J. M. B., Brandt, E. B., Brokamp, C., Ryan, P. H., and Hershey, G. K. K. (2016). Air pollution, epigenetics, and asthma. *Allergy, Asthma & Clinical Immunology*, 12(1):51.

Jiang, R., Jones, M. J., Sava, F., Kobor, M. S., and Carlsten, C. (2014). Short-term diesel exhaust inhalation in a controlled human crossover study is associated with changes in dna methylation of circulating mononuclear cells in asthmatics. *Particle and fibre toxicology*, 11(1):71.

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., Guan, W., Xu, T., Elks, C. E., Aslibekyan, S., et al. (2016). Epigenetic signatures of cigarette smoking. *Circulation: Cardiovascular Genetics*, 9(5):436–447.

Joubert, B. R., Håberg, S. E., Bell, D. A., Nilsen, R. M., Vollset, S. E., Midttun, Ø., Ueland, P. M., Wu, M. C., Nystad, W., Peddada, S. D., et al. (2014). Maternal smoking and dna methylation in newborns: in utero effect or epigenetic inheritance? *Cancer Epidemiology and Prevention Biomarkers*.

Justice, A. E., Winkler, T. W., Feitosa, M. F., Graff, M., Fisher, V. A., Young, K., Barata, L., Deng, X., Czajkowski, J., Hadley, D., et al. (2017). Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nature communications*, 8:14977.

Kelly, F. J. (2003). Oxidative stress: its role in air pollution and adverse health effects. *Occupational and environmental medicine*, 60(8):612–616.

Kelly, R. S., Lundh, T., Porta, M., Bergdahl, I. A., Palli, D., Johansson, A.-S., Botsivali, M., Vineis, P., Vermeulen, R., Kyrtopoulos, S. A., et al. (2013). Blood erythrocyte concentrations of cadmium and lead and the risk of b-cell non-hodgkins lymphoma and multiple myeloma: A nested case-control study. *PloS one*, 8(11):e81892.

Koeth, R. A., Levison, B. S., Culley, M. K., Buffa, J. A., Wang, Z., Gregory, J. C., Org, E., Wu, Y., Li, L., Smith, J. D., et al. (2014). γ-butyrobetaine is a proatherogenic intermediate in gut microbial metabolism of l-carnitine to tmao. *Cell metabolism*, 20(5):799–812.

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., Britt, E. B., Fu, X., Wu, Y., Li, L., et al. (2013). Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature medicine*, 19(5):576.

Krauskopf, J., Caiment, F., van Veldhoven, K., Chadeau-Hyam, M., Sinharay, R., Chung, K. F., Cullinan, P., Collins, P., Barratt, B., Kelly, F. J., et al. (2018). The human circulating mirnome reflects multiple organ disease risks in association with short-term exposure to traffic-related air pollution. *Environment international*, 113:26–34.

Kyrtopoulos, S. A. (2013). Making sense of omics data in population-based environmental health studies. *Environmental and molecular mutagenesis*, 54(7):468–479.

Landrigan, P. J., Fuller, R., Acosta, N. J., Adeyi, O., Arnold, R., Baldé, A. B., Bertollini, R., Bose-O'Reilly, S., Boufford, J. I., Breysse, P. N., et al. (2017). The lancet commission on pollution and health. *The Lancet*.

Lanki, T., Hampel, R., Tiittanen, P., Andrich, S., Beelen, R., Brunekreef, B., Dratva, J., De Faire, U., Fuks, K. B., Hoffmann, B., et al. (2015). Air pollution from road traffic and systemic inflammation in adults: a cross-sectional analysis in the european escape project. *Environmental health perspectives*, 123(8):785.

Larsson, N., Brown, J., Stenfors, N., Wilson, S., Mudway, I. S., Pourazar, J., and Behndig, A. F. (2013). Airway inflammatory responses to diesel exhaust in allergic rhinitics. *Inhalation toxicology*, 25(3):160–167.

Lauw, F. N., ten Hove, T., Dekkers, P. E., de Jonge, E., van Deventer, S. J., and van der Poll, T. (2000). Reduced th1, but not th2, cytokine production by lymphocytes after in vivo exposure of healthy subjects to endotoxin. *Infection and immunity*, 68(3):1014–1018.

Lazarevic, N., Dobson, A. J., Barnett, A. G., and Knibbs, L. D. (2015). Long-term ambient air pollution exposure and self-reported morbidity in the australian longitudinal study on women's health: a cross-sectional study. *BMJ open*, 5(10):e008714.

Lenters, V., Vermeulen, R., and Portengen, L. (2018). Performance of variable selection methods for assessing the health effects of correlated exposures in case–control studies. *Occup Environ Med*, 75(7):522–529.

Li, N., Alam, J., Venkatesan, M. I., Eiguren-Fernandez, A., Schmitz, D., Di Stefano, E., Slaughter, N., Killeen, E., Wang, X., Huang, A., et al. (2004). Nrf2 is a key transcription factor that regulates antioxidant defense in macrophages and epithelial cells: protecting against the proinflammatory and oxidizing effects of diesel exhaust chemicals. *The Journal of Immunology*, 173(5):3467–3481.

Li, N., Hao, M., Phalen, R. F., Hinds, W. C., and Nel, A. E. (2003). Particulate air pollutants and asthma: a paradigm for the role of oxidative stress in pm-induced adverse health effects. *Clinical Immunology*, 109(3):250–265.

Liao, W., Lin, J.-X., and Leonard, W. J. (2013). Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Immunity*, 38(1):13–25.

Liquet, B., Bottolo, L., Campanella, G., Richardson, S., Chadeau-Hyam, M., et al. (2016). R2guess: A graphics processing unit-based r package for bayesian variable selection regression of multivariate responses. *Journal of Statistical Software*, 69(1):1–32.

Liu, L.-J. S., Curjuric, I., Keidel, D., Heldstab, J., Künzli, N., Bayer-Oglesby, L., Ackermann-Liebrich, U., Schindler, C., and team the SAPALDIA (2007). Characterization of source-specific air pollution exposure for a large population-based swiss cohort (sapaldia). *Environmental health perspectives*, 115(11):1638–1645.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., and Straif, K. (2013). The carcinogenicity of outdoor air pollution. *The lancet oncology*, 14(13):1262–1263.

Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental health perspectives*, pages 1691–1696.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614.

Madrigano, J., Baccarelli, A., Mittleman, M. A., Sparrow, D., Spiro III, A., Vokonas, P. S., Cantone, L., Kubzansky, L., and Schwartz, J. (2012). Air pollution and dna methylation: interaction by psychological factors in the va normative aging study. *American journal of epidemiology*, 176(3):224–232.

Madrigano, J., Baccarelli, A., Mittleman, M. A., Wright, R. O., Sparrow, D., Vokonas, P. S., Tarantini, L., and Schwartz, J. (2011). Prolonged exposure to particulate pollution, genes associated with glutathione pathways, and dna methylation in a cohort of older men. *Environmental health perspectives*, 119(7):977.

Martin, B. W., Ackermann-Liebrich, U., Leuenberger, P., Künzli, N., Stutz, E. Z., Keller, R., Zellweger, J.-P., Wüthrich, B., Monn, C., Blaser, K., et al. (1997). Sapaldia: methods and participation in the cross-sectional part of the swiss study on air pollution and lung diseases in adults. *Sozial-und Präventivmedizin*, 42(2):67–84.

McHale, C. M., Zhang, L., Lan, Q., Vermeulen, R., Li, G., Hubbard, A. E., Porter, K. E., Thomas, R., Portier, C. J., Shen, M., et al. (2011). Global gene expression profiling of a population exposed to a range of benzene levels. *Environmental health perspectives*, 119(5):628.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., et al. (2012). The ucsc genome browser database: extensions and updates 2013. *Nucleic acids research*, 41(D1):D64–D69.

Miller, R. L. and Peden, D. B. (2014). Environmental effects on immune responses in patients with atopy and asthma. *Journal of Allergy and Clinical Immunology*, 134(5):1001–1008.

Morris, T. J. and Beck, S. (2015). Analysis pipelines and packages for infinium humanmethylation450 beadchip (450k) data. *Methods*, 72:3–8.

Morse, D., Pischke, S. E., Zhou, Z., Davis, R. J., Flavell, R. A., Loop, T., Otterbein, S. L., Otterbein, L. E., and Choi, A. M. (2003). Suppression of inflammatory cytokine production by carbon monoxide involves the jnk pathway and ap-1. *Journal of Biological Chemistry*, 278(39):36993–36998.

Mostafavi, N., Vlaanderen, J., Chadeau-Hyam, M., Beelen, R., Modig, L., Palli, D., Bergdahl, I. A., Vineis, P., Hoek, G., Kyrtopoulos, S. A., et al. (2015). Inflammatory markers in relation to long-term air pollution. *Environment international*, 81:1–7.

Mostafavi, N., Vlaanderen, J., Portengen, L., Chadeau-Hyam, M., Modig, L., Palli, D., Bergdahl, I. A., Brunekreef, B., Vineis, P., Hebels, D. G., et al. (2017). Associations between genome-wide gene expression and ambient nitrogen oxides. *Epidemiology*, 28(3):320–328.

Navarro, S. L., Brasky, T. M., Schwarz, Y., Song, X., Wang, C., Kristal, A. R., Kratz, M., White, E., and Lampe, J. W. (2012). Reliability of serum biomarkers of inflammation from repeated measures in healthy individuals. *Cancer Epidemiology Biomarkers & Prevention*, 21(7):1167–1170.

Nawrot, T. S. and Adcock, I. (2009). The detrimental health effects of traffic-related air pollution: a role for dna methylation?

Niedzwiecki, M. M., Walker, D. I., Vermeulen, R., Chadeau-Hyam, M., Jones, D. P., and Miller, G. W. (2019). The exposome: molecules to populations. *Annual review of pharmacology and toxicology*, 59:107–127.

Nieuwenhuijsen, M., Paustenbach, D., and Duarte-Davidson, R. (2006). New developments in exposure assessment: the impact on the practice of health risk assessment and epidemiological studies. *Environment International*, 32(8):996–1009.

Nunen, E. v., Vermeulen, R., Tsai, M.-Y., Probst Hensch, N., Ineichen, A., Davey, M., Imboden, M., Ducret Stich, R., Naccarati, A., Raffaele, D., et al. (2017). Land use regression models for ultrafine particles in six european areas. *Environmental Science & Technology, 2017, vol. 51, num. 6, p. 3336-3345*.

Palli, D., Berrino, F., Vineis, P., Tumino, R., Panico, S., Masala, G., Saieva, C., Salvini, S., Ceroti, M., Pala, V., et al. (2003). A molecular epidemiology project on diet and cancer: the epic-italy prospective study. design and baseline characteristics of participants. *Tumori*, 89(6):586–593.

Panasevich, S., Leander, K., Rosenlund, M., Ljungman, P., Bellander, T., de Faire, U., Pershagen, G., and Nyberg, F. (2009). Associations of long-and short-term air pollution exposure with markers of inflammation and coagulation in a population sample. *Occupational and environmental medicine*.

Panni, T., Mehta, A. J., Schwartz, J. D., Baccarelli, A. A., Just, A. C., Wolf, K., Wahl, S., Cyrys, J., Kunze, S., Strauch, K., et al. (2016). Genome-wide analysis of dna methylation and fine particulate matter air pollution in three study populations: Kora f3, kora f4, and the normative aging study. *Environmental health perspectives*, 124(7):983.

Pedersen, B. S., Schwartz, D. A., Yang, I. V., and Kechris, K. J. (2012). Comb-p: software for combining, analyzing, grouping and correcting spatially correlated p-values. *Bioinformatics*, 28(22):2986–2988.

Pedersen, B. S., Yang, I. V., and De, S. (2013). Cruzdb: software for annotation of genomic intervals with ucsc genome-browser database. *Bioinformatics*, 29(23):3003–3006.

Peretz, A., Sullivan, J. H., Leotta, D. F., Trenga, C. A., Sands, F. N., Allen, J., Carlsten, C., Wilkinson, C. W., Gill, E. A., and Kaufman, J. D. (2008). Diesel exhaust inhalation elicits acute vasoconstriction in vivo. *Environmental health perspectives*, 116(7):937.

Petrarca, F., Russolillo, G., and Trinchera, L. (2017). Integrating non-metric data in partial least squares path models: Methods and application. In *Partial Least Squares Path Modeling*, pages 259–279. Springer.

Pettit, A. P., Brooks, A., Laumbach, R., Fiedler, N., Wang, Q., Strickland, P. O., Madura, K., Zhang, J., and Kipen, H. M. (2012). Alteration of peripheral blood monocyte gene expression in humans following diesel exhaust inhalation. *Inhalation toxicology*, 24(3):172–181.

Philibert, R. A., Beach, S. R., and Brody, G. H. (2012). Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics*, 7(11):1331–1338.

Plusquin, M., Guida, F., Polidoro, S., Vermeulen, R., Raaschou-Nielsen, O., Campanella, G., Hoek, G., Kyrtopoulos, S. A., Georgiadis, P., Naccarati, A., et al. (2017). Dna methylation and exposure to ambient air pollution in two prospective cohorts. *Environment international*, 108:127–136.

Pope III, C. A., Bhatnagar, A., McCracken, J. P., Abplanalp, W., Conklin, D. J., and Otoole, T. (2016). Exposure to fine particulate air pollution is associated with endothelial injury and systemic inflammation. *Circulation research*, 119(11):1204–1214.

Pope III, C. A., Burnett, R. T., Turner, M. C., Cohen, A., Krewski, D., Jerrett, M., Gapstur, S. M., and Thun, M. J. (2011). Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure-response relationships. *Environmental health perspectives*, 119(11):1616.

Rappaport, S. M. (2011). Implications of the exposome for exposure science. *Journal of Exposure Science and Environmental Epidemiology*, 21(1):5.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85.

Romieu, I., Moreno-Macias, H., and London, S. J. (2010). Gene by environment interaction and ambient air pollution. *Proceedings of the American Thoracic Society*, 7(2):116–122.

Rückerl, R., Greven, S., Ljungman, P., Aalto, P., Antoniades, C., Bellander, T., Berglind, N., Chrysohoou, C., Forastiere, F., Jacquemin, B., et al. (2007). Air pollution and inflammation (interleukin-6, c-reactive protein, fibrinogen) in myocardial infarction survivors. *Environmental health perspectives*, pages 1072–1080.

Sanchez, G. (2013). Pls path modeling with r. *Berkeley: Trowchez Editions*, 383.

Schikowski, T., Schaffner, E., Meier, F., Phuleria, H. C., Vierkötter, A., Schindler, C., Kriemler, S., Zemp, E., Krämer, U., Bridevaux, P.-O., et al. (2013). Improved air quality and attenuated lung function decline: modification by obesity in the sapaldia cohort. *Environmental health perspectives*, 121(9):1034.

Seaton, A., Soutar, A., Crawford, V., Elton, R., McNerlan, S., Cherrie, J., Watt, M., Agius, R., and Stout, R. (1999). Particulate air pollution and the blood. *Thorax*, 54(11):1027–1032.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Shen, Y.-S. and Lung, S.-C. C. (2017). Mediation pathways and effects of green structures on respiratory mortality via reducing air pollution. *Scientific Reports*, 7:42854.

Shenker, N. S., Ueland, P. M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., Flanagan, J. M., and Vineis, P. (2013). Dna methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*, 24(5):712–716.

Sill, M., Hielscher, T., Becker, N., Zucknick, M., et al. (2014). c060: Extended inference with lasso and elastic-net regularized cox and generalized linear models. *Journal of Statistical Software*, 62(5):1–22.

Simoni, M., Baldacci, S., Maio, S., Cerrai, S., Sarno, G., and Viegi, G. (2015). Adverse effects of outdoor pollution in the elderly. *Journal of thoracic disease*, 7(1):34–45.

Sousa, A. R., Trigg, C. J., Lane, S. J., Hawksworth, R., Nakhosteen, J. A., Poston, R. N., and Lee, T. H. (1997). Effect of inhaled glucocorticoids on il-1 beta and il-1 receptor antagonist (il-1 ra) expression in asthmatic bronchial epithelium. *Thorax*, 52(5):407–410.

Steenhof, M., Mudway, I. S., Gosens, I., Hoek, G., Godri, K. J., Kelly, F. J., Harrison, R. M., Pieters, R. H., Cassee, F. R., Lebret, E., et al. (2013). Acute nasal pro-inflammatory response to air pollution depends on characteristics other than particle mass concentration or oxidative potential: the raptes project. *Occupational and environmental medicine*, pages oemed–2012.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Sung, Y. J., Winkler, T. W., de Las Fuentes, L., Bentley, A. R., Brown, M. R., Kraja, A. T., Schwander, K., Ntalla, I., Guo, X., Franceschini, N., et al. (2018). A large-scale multi-ancestry genome-wide study accounting for smoking behavior identifies multiple significant loci for blood pressure. *The American Journal of Human Genetics*, 102(3):375–400.

Tang, W. W., Wang, Z., Levison, B. S., Koeth, R. A., Britt, E. B., Fu, X., Wu, Y., and Hazen, S. L. (2013). Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *New England Journal of Medicine*, 368(17):1575–1584.

Tarantini, L., Bonzini, M., Apostoli, P., Pegoraro, V., Bollati, V., Marinelli, B., Cantone, L., Rizzo, G., Hou, L., Schwartz, J., et al. (2009). Effects of particulate matter on genomic dna methylation content and inos promoter methylation. *Environmental health perspectives*, 117(2):217.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., and Lauro, C. (2005). Pls path modeling. *Computational statistics & data analysis*, 48(1):159–205.

Trifunović, J., Miller, L., Debeljak, Ž., and Horvat, V. (2015). Pathologic patterns of interleukin 10 expression–a review. *Biochemia medica*, 25(1):36–48.

Tsai, D.-H., Amyai, N., Marques-Vidal, P., Wang, J.-L., Riediker, M., Mooser, V., Paccaud, F., Waeber, G., Vollenweider, P., and Bochud, M. (2012). Effects of particulate matter on inflammatory markers in the general adult population. *Particle and fibre toxicology*, 9(1):1.

Uppal, K., Go, Y.-M., and Jones, D. P. (2017). xmwas: an r package for data-driven integration and differential network analysis. *bioRxiv*, page 122432.

van Eeden, S. F., Tan, W. C., Suwa, T., Mukae, H., Terashima, T., Fujii, T., Qui, D., Vincent, R., and Hogg, J. C. (2001). Cytokines involved in the systemic inflammatory response induced by exposure to particulate matter air pollutants (pm10). *American journal of respiratory and critical care medicine*, 164(5):826–830.

van Kesteren, E.-J. and Oberski, D. L. (2018). Exploratory mediation analysis with many potential mediators. *arXiv preprint arXiv:1810.06334*.

Van Leeuwen, D., Van Herwijnen, M., Pedersen, M., Knudsen, L. E., Kirsch-Volders, M., Sram, R., Staal, Y., Bajak, E., van Delft, J., and Kleinjans, J. (2006). Genome-wide differential gene expression in children exposed to air pollution in the czech republic. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 600(1):12–22.

van Leeuwen, D. M., Gottschalk, R. W., Schoeters, G., Van Larebeke, N., Nelen, V., Baeyens, W. F., Kleinjans, J., and van Delft, J. H. (2008). Transcriptome analysis in peripheral blood of humans exposed to environmental carcinogens: a promising new biomarker in environmental health studies. *Environmental health perspectives*, 116(11):1519–1525.

VanderWeele, T. J. (2017). Outcome-wide epidemiology. *Epidemiology (Cambridge, Mass.)*, 28(3):399.

Vawda, S., Mansour, R., Takeda, A., Funnell, P., Kerry, S., Mudway, I., Jamaludin, J., Shaheen, S., Griffiths, C., and Walton, R. (2013). Associations between inflammatory and immune response genes and adverse respiratory outcomes following exposure to outdoor air pollution: a huge systematic review. *American journal of epidemiology*, 179(4):432–442.

Vineis, P., Chadeau-Hyam, M., Gmuender, H., Gulliver, J., Herceg, Z., Kleinjans, J., Kogevinas, M., Kyrtopoulos, S., Nieuwenhuijsen, M., Phillips, D., et al. (2017). The exposome in practice: design of the exposomics project. *International journal of hygiene and environmental health*, 220(2):142–151.

Vineis, P., van Veldhoven, K., Chadeau-Hyam, M., and Athersuch, T. J. (2013). Advancing the application of omics-based biomarkers in environmental epidemiology. *Environmental and molecular mutagenesis*, 54(7):461–467.

Vlaanderen, J., Van Veldhoven, K., Font-Ribera, L., Villanueva, C. M., Chadeau-Hyam, M., Portengen, L., Grimalt, J. O., Zwiener, C., Heederik, D., Zhang, X., et al. (2017). Acute changes in serum immune markers due to swimming in a chlorinated pool. *Environment international*, 105:1–11.

Walker, D. I., Uppal, K., Zhang, L., Vermeulen, R., Smith, M., Hu, W., Purdue, M. P., Tang, X., Reiss, B., Kim, S., et al. (2016). High-resolution metabolomics of occupational exposure to trichloroethylene. *International journal of epidemiology*, 45(5):1517–1527.

Wang, C., Chen, R., Cai, J., Shi, J., Yang, C., Tse, L. A., Li, H., Lin, Z., Meng, X., Liu, C., et al. (2016). Personal exposure to fine particulate matter and blood pressure: a role of angiotensin converting enzyme and its dna methylation. *Environment international*, 94:661–666.

Wang, C., Chen, R., Shi, M., Cai, J., Shi, J., Yang, C., Li, H., Lin, Z., Meng, X., Liu, C., et al. (2017). Possible mediation by methylation in acute inflammation following personal exposure to fine particulate air pollution. *American journal of epidemiology*, 187(3):484–493.

Wang, H., Gottfries, J., Barrenäs, F., and Benson, M. (2011a). Identification of novel biomarkers in seasonal allergic rhinitis by combining proteomic, multivariate and pathway analysis. *PLoS One*, 6(8):e23563.

Wang, R., Henderson, S. B., Sbihi, H., Allen, R. W., and Brauer, M. (2013). Temporal stability of land use regression models for traffic-related air pollution. *Atmospheric Environment*, 64:312–319.

Wang, T. W., Vermeulen, R. C., Hu, W., Liu, G., Xiao, X., Alekseyev, Y., Xu, J., Reiss, B., Steiling, K., Downward, G. S., et al. (2015). Gene-expression profiling of buccal epithelium among non-smoking women exposed to household air pollution from smoky coal. *Carcinogenesis*, 36(12):1494–1501.

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., DuGar, B., Feldstein, A. E., Britt, E. B., Fu, X., Chung, Y.-M., et al. (2011b). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, 472(7341):57.

Weichenthal, S., Bai, L., Hatzopoulou, M., Ryswyk, K., Kwong, J. C., Jerrett, M., Donkelaar, A., Martin, R. V., Burnett, R. T., Lu, H., et al. (2017). Long-term exposure to ambient ultrafine particles and respiratory disease incidence in in toronto, canada: a cohort study. *Environmental Health*, 16(1):64.

Whelan, R., Kim, C., Chen, M., Leiter, J., Grunstein, M., and Hakonarson, H. (2004). Role and regulation of interleukin-1 molecules in pro-asthmatic sensitised airway smooth muscle. *European Respiratory Journal*, 24(4):559–567.

WHO (2018). The world health report 2018. Technical report, Genf, Schweiz.

Wild, C. P. (2005). Complementing the genome with an exposome: the outstanding challenge of environmental exposure measurement in molecular epidemiology.

Wild, C. P., Scalbert, A., and Herceg, Z. (2013). Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. *Environmental and molecular mutagenesis*, 54(7):480–499.

Wittkopp, S., Staimer, N., Tjoa, T., Stinchcombe, T., Daher, N., Schauer, J. J., Shafer, M. M., Sioutas, C., Gillen, D. L., and Delfino, R. J. (2016). Nrf2-related gene expression and exposure to traffic-related air pollution in elderly subjects with cardiovascular disease: An exploratory panel study. *Journal of Exposure Science and Environmental Epidemiology*, 26(2):141–149.

Wong, H.-L., Pfeiffer, R. M., Fears, T. R., Vermeulen, R., Ji, S., and Rabkin, C. S. (2008). Reproducibility and correlations of multiplex cytokine levels in asymptomatic persons. *Cancer Epidemiology Biomarkers & Prevention*, 17(12):3450–3456.

Wouters, E. F., Reynaert, N. L., Dentener, M. A., and Vernooy, J. H. (2009). Systemic and local inflammation in asthma and chronic obstructive pulmonary disease: is there a connection? *Proceedings of the American Thoracic Society*, 6(8):638–647.

Wright, M. L., Dozmorov, M. G., Wolen, A. R., Jackson-Cook, C., Starkweather, A. R., Lyon, D. E., and York, T. P. (2016). Establishing an analytic pipeline for genome-wide dna methylation. *Clinical epigenetics*, 8(1):45.

Wu, W., Jin, Y., and Carlsten, C. (2018). Inflammatory health effects of indoor and outdoor particulate matter. *Journal of Allergy and Clinical Immunology*, 141(3):833–844.

Xing, G., Lin, C.-Y., and Xing, C. (2011). A comparison of approaches to control for confounding factors by regression models. *Human heredity*, 72(3):194–205.

Zarogoulidis, P., Katsikogianni, F., Tsiouda, T., Sakkas, A., Katsikogiannis, N., and Zarogoulidis, K. (2014). Interleukin-8 and interleukin-17 for cancer. *Cancer investigation*, 32(5):197–205.

Zhang, Q. (2018). High dimensional mediation analysis with applications to causal gene identification. *bioRxiv*, page 497826.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zuurbier, M., Hoek, G., Oldenwening, M., Meliefste, K., Krop, E., van den Hazel, P., and Brunekreef, B. (2011). In-traffic air pollution exposure and cc16, blood coagulation, and inflammation markers in healthy adults. *Environmental health perspectives*, 119(10):1384.