Sex-specific associations between particulate matter exposure and gene expression in independent discovery and validation cohorts of middle-aged men and women

Vrijens, Karen; Winckelmans, Ellen; Tsamou, Maria; Baeyens, Willy; De Boever, Patrick; Jennen, Danyel; de Kok, Theo M.; Den Hond, Elly; Lefebvre, Wouter; Plusquin, Michelle; Reyners, Hans; Schoeters, Greet; van Larebeke, Nicolas; Vanpoucke, Charlotte; Kleinjans, Jos; Nawrot, Tim S

Published in: Environmental Health Perspectives

DOI: 10.1289/EHP370

Publication date: 2017

Document version: Publisher's PDF, also known as Version of record

Document license: CC BY-NC-SA

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 17. dec.. 2018
Sex-Specific Associations between Particulate Matter Exposure and Gene Expression in Independent Discovery and Validation Cohorts of Middle-Aged Men and Women

Karen Vrijens,1 Ellen Winckelmans,1 Maria Tsamou,1 Willy Baeyens,2 Patrick De Boever,1,3 Danyel Jennen,4 Theo M. de Kok,5 Elly Den Hond,2,5 Wouter Lefebvre,3 Michelle Plusquin,1 Hans Reyners,6 Greet Schoeters,3,7,8 Nicolas Van Larebeke,9 Charlotte Vanpoucke,10 Jos Kleinjans,4 and Tim S. Nawrot1,11

1Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium; 2Department of Analytical and Environmental Chemistry, Free University of Brussels, Brussels, Belgium; 3Environmental Risk and Health Unit, Flemish Institute for Technological Research (VITO), Mol, Belgium; 4Department of Toxicogenomics, Maastricht University, Maastricht, Netherlands; 5Provincial Institute for Hygiene, Antwerp, Belgium; 6Environment, Nature and Energy Department, Flemish Government, Brussels, Belgium; 7Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium; 8University of Southern Denmark, Institute of Public Health, Department of Environmental Medicine, Odense, Denmark; 9Department of Radiotherapy and Nuclear Medicine, Ghent University, Ghent, Belgium; 10Belgian Interregional Environment Agency (IRCEL), Brussels, Belgium; 11Department of Public Health and Primary Care, Leuven University, Leuven, Belgium

Background: Particulate matter (PM) exposure leads to premature death, mainly due to respiratory and cardiovascular diseases.

Objective: Identification of transcriptomic biomarkers of air pollution exposure and effect in a healthy adult population.

Methods: Microarray analyses were performed in 98 healthy volunteers (48 men, 50 women). The expression of eight sex-specific candidate biomarker genes (significantly associated with PM10 in the discovery cohort and with a reported link to air pollution-related disease) was measured with qPCR in an independent validation cohort (75 men, 94 women). Pathway analysis was performed using Gene Set Enrichment Analysis. Average daily PM2.5 and PM10 exposures over 2-years were estimated for each participant’s residential address using spatiotemporal interpolation in combination with a dispersion model.

Results: Average long-term PM10 was 25.9 (± 5.4) and 23.7 (± 2.3) μg/m3 in the discovery and validation cohorts, respectively. In discovery analysis, associations between PM10 and the expression of individual genes differed by sex. In the validation cohort, long-term PM10 was associated with the expression of DNAJB5 and EAPP in men and ARHGP4 (p = 0.053) in women. AKAP6 and LIMK1 were significantly associated with PM10 in women, although associations differed in direction between the discovery and validation cohorts. Expression of the eight candidate genes in the discovery cohort differentiated between validation cohort participants with high versus low PM10 exposure (area under the receiver operating curve = 0.92; 95% CI: 0.85, 1.00; p = 0.0002 in men, 0.86; 95% CI: 0.76, 0.96; p = 0.004 in women).

Conclusions: Expression of the sex-specific candidate genes identified in the discovery population predicted PM10 exposure in an independent cohort of adults from the same area. Confirmation in other populations may further support this as a new approach for exposure assessment, and may contribute to the discovery of molecular mechanisms for PM-induced health effects.


Background
Particulate matter (PM) is a complex mixture of small particles and liquid droplets that contains a number of components, including acids, organic chemicals, metals, and soil or dust particles. PM exposure is known to increase overall mortality and morbidity, mainly due to its effect on the cardiorespiratory system (Alfarro-Moreno et al. 2007; Pope et al. 2004). Exposure to PM may disturb normal physiological pathways that maintain homeostasis and this may activate cellular processes that mediate the adverse effects of PM (Kleensang et al. 2014). Gene expression changes play an important role in the activation of pathways of toxicity and gene signatures have the potential to serve as biomarkers of exposure (van Leeuwen et al. 2008; van Breda et al. 2015) and recent reports demonstrate their potential use as biomarkers of effect (La Rocca et al. 2014; Fink et al. 2014). As it has been shown previously that transcriptomic responses to diverse environmental stimuli (i.e., chemical exposure, smoking) can be significantly different between men and women (De Coster et al. 2013; Paul and Amundson 2014), we have opted to perform a sex-specific analysis.

Several studies have suggested that elevated oxidative stress may mediate toxic effects of air pollutants (Donaldson et al. 2005; Nel et al. 2001). The systemic inflammatory response following acute inhalation exposure to PM can induce leukocytosis and monocyte release from the bone marrow (Fujii et al. 2002). Controlled exposure studies of recent diesel exhaust exposure (Perrit et al. 2012) and recent exposure to ultra-fine particles (Huang et al. 2010) have reported evidence of altered gene expression in leukocytes but, to our knowledge, associations between patterns of gene expression and long-term particulate air pollution have not been studied in general populations.

Materials and Methods
Study Design
As our goal was to identify transcriptomic biomarkers of exposure and effect in a healthy adult population, we started by applying microarray analysis in a discovery cohort of 98 adults for which we modelled particulate matter exposure. On the resulting dataset containing significantly modulated genes and pathways, we applied a literature and bioinformatics approach to identify potential exposure effect biomarkers. Subsequently,

Address correspondence to T.S. Nawrot, Centre for Environmental Sciences, Hasselt University, Agoralaan gebouw D, B-3590 Diepenbeek, Belgium. Telephone: 0032/11-26.83.82. E-mail: tim.nawrot@uhasselt.be

Supplemental Material is available online (http://dx.doi.org/10.1289/EHP370).

The project was funded by the Environment, Nature and Energy Department of the Flemish government (LNE/OL201100023/13034/M&G), Steunpunt Milieu- en Gezondheid and European Research Council (ERC-2012-StG 310898). K.V. is a postdoctoral fellow of the Research Foundation—Flanders (12D7714N).

The authors declare they have no actual or potential competing financial interests.

Received: 21 December 2015; Revised: 12 August 2016; Accepted: 22 August 2016; Published: 14 October 2016.

Note to readers with disabilities: EHP strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in EHP articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehponline@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.
these were validated using qPCR analysis in an independent cohort with similar characteristics as the discovery population (Figure 1). Study protocols for the discovery and validation cohort were approved by the Institutional Review Board and the Ethical Committee of Antwerp University and informed consent was obtained from all participants.

**Study Population**

**Discovery cohort.** The original study population was described previously (van Leeuwen et al. 2008) and consisted of 398 participants from eight different regions of residence in Flanders (Belgium), as part of the first Flemish Environment and Health Survey (FLEHS I) during the period 2001–2006. Participants were recruited in several communities based on random sampling. Inclusion criteria were age 50–65 years, living in Flanders > 5 years, and being able to complete questionnaires in Dutch. Prior to blood collection, informed consent was obtained from all individuals. A subset of 98 samples was selected for microarray analysis based on previously measured exposure levels to several pollutants including cadmium, lead, polychlorinated biphenyls (PCBs) (138, 153, and 180), dioxins, polycyclic aromatic hydrocarbons (PAHs), and benzene. The overall exposure to these pollutants was estimated using a z-score for each pollutant, and study participants with both low- and high-exposure levels were chosen for inclusion. Z-scores were not correlated with long-term PM$_{10}$ exposure ($r^2 = 0.0012$). Smokers were excluded from the study population. PAXgene tubes (PreAnalytiX GmbH, Hombrechtikon, Switzerland) were used for RNA collection.

**Validation cohort.** The quantitative polymerase chain reaction (qPCR) validation study was performed in an independent cohort of 175 adults who were part of the third Flemish Environment and Health Survey (FLEHS III) during the period 2012–2015. Healthy volunteers who were between 50 and 65 years of age, living at the same residential address for at least 10 years, and able to complete questionnaires in Dutch were recruited through registers of general medical practices. Prior to blood collection, informed consent was obtained from all individuals. Participants completed a questionnaire covering age, sex, and smoking habits, among other demographic characteristics; they donated blood and urine samples; and subclinical measurements including height, weight, and blood pressure were determined. The sampling campaign lasted from May 2014 until 30 December 2014. We used PAXgene tubes (PreAnalytiX GmbH, Hombrechtikon, Switzerland) to stabilize whole blood RNA for storage.

**Exposure Estimates**

The PM$_{10}$ and PM$_{2.5}$ (10 and 2.5 μm in aerodynamic diameter) concentrations for participants’ residential addresses were calculated using a spatial temporal interpolation method (Kriging) that takes into account land cover data from satellite images [CORINE (coordination of information on the environment) land cover data set; http://www.eea.europa.eu/publications/COR0-landcover] for interpolating the measurement data of the monitoring stations from the Belgian telemetric air quality network as described previously (Maheu et al. 2013; Jacobs et al. 2010; Janssen et al. 2008). Validation statistics of the interpolation tool gave a temporal explained variance of $> 0.7$ for hourly PM$_{10}$ averages as well as for annual mean PM$_{10}$ (Maheu et al. 2013). In combination with the Immission Frequency Distribution Model (IFDM) using emissions from line sources and point sources, the model chain provides averaged as well as for annual mean PM$_{10}$ and PM$_{2.5}$ values on a $25 \times 25$ m receptor grid (Lefebvre et al. 2013). Our model is based on input data from 38 monitoring stations in the study area. The Initiative on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Use in Europe was the incentive for intensive model intercomparison. IFDM was thoroughly compared with other models currently in use for regulatory purposes in Europe (Olesen 1995; Maes et al. 1995; Cosemans et al. 1995, 2001; Mensink and Maes 1996).

Mean daily temperatures and relative humidity for the study region were provided by the Royal Meteorological Institute (Brussels, Belgium), and apparent temperature was calculated (Steadman 1979; Kalkstein and Valimont 1986).

All our estimates were annual mean exposures over a 2-year period because we were interested in developing biomarkers for long-term exposure. For the discovery cohort, annual means were based on 2011–2012 as these were the earliest years for which detailed $25 \times 25$ m grid information became available. Distribution patterns were used for the year 2008. We assumed that relative differences in annual mean concentrations of particulate matter were generally consistent from year to year. For the validation cohort, annual means were based on the 2 years prior to blood sampling (i.e., 2012–2013).

**RNA Isolation**

Total RNA was isolated from 2.5 mL whole blood from PAXgene Blood RNA vacutainers using the PAXgene Blood RNA system (PreAnalytiX, Qiagen, Hilden, Germany), according to the manufacturer’s instructions. A globin reduction assay (GLOBINclear™ Kit by Ambion, Austin, TX, USA) was performed in order to remove hemoglobin mRNA from samples that were submitted to microarray analysis. RNA integrity was assessed using the BioAnalyzer (Agilent, Palo Alto, CA, USA) and purity was measured.
spectrophotometrically. Labeled samples were checked for specific activity and dye incorporation.

Microarray Preparation and Hybridization

We used 0.2 μg total RNA from each sample to synthesize dye-labeled cRNA (Cy3) following the Agilent one-color Quick-Amp labeling protocol (Agilent Technologies). Individual samples were hybridized on Agilent 4 × 44 K Whole Human Genome microarrays (design ID 014850).

Microarray Data Analysis

Microarrays were scanned on an Agilent G2505C DNA Microarray Scanner (Agilent Technologies, Amstelveen, Netherlands). Raw data on pixel intensities were extracted from the scanned images using Agilent Feature Extraction Software (version 10.7.3.1; Agilent Technologies, Amstelveen, Netherlands), protocol GE1_107_sep09. Raw data were pre-processed using an in-house developed quality control pipeline in R (version 2.15.3; R Project for Statistical Computing) as follows: local background correction, flagging of bad spots, controls and spots with intensities below background, log2 transformation and quantile normalization. The R-scripts of the pipeline and additional information on the flagging can be found at https://github.com/BiGcat-um/arrayQC_Module. From the processed data files genes were omitted showing more than 30% flagged data, after which the data files were transferred to the Gene Expression Pattern Analysis Suite, GEPAS 2010 (Montaner et al. 2006) for further preprocessing, including merging replicates (based on median), and imputing missing values by means of K-nearest neighbor imputation (K = 15). Filtering for flat peaks was used with root mean square value 0.25. The filtered data, containing 28,786 genes, were used for further statistical analyses. Microarray gene expression data were analyzed and stratified for sex. In the original microarray data that set initially 28,786 unique Agilent probe IDs (out of 43,376 Agilent probe IDs) were annotated to 22,390 EntrezGene IDs. A further measurement. A quantitative real-time polymerase chain reaction (qPCR) was set up by adding 2 μL of a 10 ng/μL dilution of cDNA together with TaqMan Fast Advanced Master Mix (Life Technologies, Foster City, CA, USA) and PrimeTime™ assay (Integrated DNA Technologies, Coralville, IA, USA), in a final reaction volume of 10 μL. Standard cycling conditions were used to analyze samples in a 7900HT Fast Real-Time PCR system (Life Technologies, Foster City, CA, USA). Expression of eight candidate biomarker genes for each sex was studied and Cq values were collected with SDS 2.3 software. Minimum Information on qPCR Experiments (MIQE) guidelines were taken into account (Bustin et al. 2009). Amplification efficiencies were between 90% and 110% for all assays. Raw data were processed to normalized relative gene expression values with qBase plus (Biogazelle, Zwijnaarde, Belgium) (Hellemans et al. 2007). Triplicates were run for all samples; technical replicates were included when the difference in Cq value was < 0.5. A set of three genes was used for data normalization, namely Hypoxanthine Phosphoribosyltransferase 1 (HPRT), Importin 8 (IP08) and tyrosine 3-monoxygenase/tryptophan 5-monoxygenase activation protein, zeta (YWHAZ).

Data analysis. Statistical analyses were carried out using SAS software (version 9.3, SAS Institute Inc., Cary, NC, USA). Continuous data were presented as mean ± standard deviation (SD) and categorical data as percentages (%) and frequencies. Models were adjusted for age, body mass index (BMI), SES, smoking (categorized as smokers, former smokers and never smokers), white blood cell counts (absolute number of leukocytes and percentage of neutrophils), time of day (< 1200 hours, 1200–1500 hours, 1500–1800 hours, > 2000 hours) and season (October–March or April–September) of blood sampling. p-Values ≤ 0.05 were considered statistically significant.

Gene Expression Analysis

Using linear regression models adjusted for age, body mass index (BMI), socioeconomic status (SES, classified in three groups: no high school degree, high school degree, or further education degree), daytime and season of blood sampling, we obtained estimates for each gene as the log2-fold change in gene expression for an increment of 5 μg/mL exposure. p-Values < 0.05 were considered statistically significant. p-Values were corrected for multiplicity using the Benjamini–Hochberg false discovery rate (FDR) correction. p-Values corrected for multiple testing are referred to as q-values.

Pathway Analysis

Gene Set Enrichment Analysis was performed utilizing the online pathway analysis tool Consensus Pathway Data Base (CPDB) (http://consensuspathdb.org/). CPDB contains ∼5,200 pathways including protein complexes, metabolic, signaling, and gene regulatory pathways, as well as drug-target interactions. Data originate from 32 public resources curated from the literature (Kamburov et al. 2013). Gene Set Enrichment Analysis was performed in a sex-specific manner using the log2-fold changes of the gene expression data for all genes analyzed at the gene expression level, without preselection. For every predefined gene set in each pathway, a Wilcoxon signed-rank test was calculated testing the null hypothesis that the distribution of their fold changes was around zero. As input, all genes without a priori selection (EntrezGene IDs) were uploaded with their fold changes in their gene expression. We selected the biological processes using pathways as output. The p-values were corrected for multiplicity and were presented as q-values. We defined significant biological processes and pathways by a threshold on the adjusted p-value (q < 0.05 or FDR 5%) and we included gene sets with a size between 5 and 100 members.

Selection of Potential Exposure/Effect Biomarker Genes

We used a modified version of the meet-in-the-middle approach for biomarker identification in relation to clinical relevance (Vineis et al. 2013), a schematic representation is shown in Figure 1. We first identified the top 50 genes associated with PM10 (i.e., with the smallest uncorrected p-values) in men and women, respectively, then performed a literature search using PubMed and ScienceDirect to identify genes within each sex-specific set that have been associated with air pollution-related health outcomes. Specifically, we searched for the name of each gene in combination with any of the following diseases or processes: allergy (Magnussen et al. 1993), chronic obstructive pulmonary disease (COPD) (Ko et al. 2007), asthma (Bowatte et al. 2015), lung cancer (Raaschou-Nielsen et al. 2011), cardiovascular disease (CVD) (Mills et al. 2009), cerebrovascular disease (CeVD) (Johnson et al. 2010), Alzheimer’s disease (Finkelstein and Jerrett 2007) and cognition (Dadvand et al. 2015). Genes with lowest p-values and proven link to air pollution (AP)-related diseases were chosen for validation. For men, DNAJB5, RAC3, EAPP, HDBLP, PRG2, PERI, PIK3R1, and SLA2 were selected for validation, whereas for women the gene list for validation included AKAP6, LIMK1, SIRT7, ARHPGAP4, ATG16L2, TPM3, 5-HTR1B, and PYG02.
p-values corrected for multiple testing referred to as q-values. We plotted residuals for each gene to check whether significance was driven by outliers, these were removed where appropriate. To indicate significance of selected biomarker genes for each sex, we included an interaction term for sex in our main analysis. p-Values for the interaction term sex were calculated for all genes under study, not only those that were significant.

In validation analysis, we examined the association between gene expression and PM10 exposure, stratified by sex using linear regression models for the eight selected genes for each sex.

**ROC Curves Exposure Prediction**

We calculated the ability to predict PM10 exposure based on expression of the set of eight validated genes significantly associated with PM10 exposure in the discovery cohort for each sex. For this purpose, we estimated sensitivity and specificity of the prediction using receiver operating characteristic (ROC) plots. Subjects were stratified according to their long-term PM10 exposure levels with the 75th percentile of exposure as cut-off point (25.7 μg/m³ annual mean for women, 24.5 μg/m³ for men). All long-term PM10 exposure levels with the 75th percentile were plotted. Subjects were stratified according to their long-term PM10 exposure levels, the cut-off point as 16.0 μg/m³ for both men and women.

**Pathway Analysis**

There were 1,251 and 966 pathways significantly associated with PM10 and PM2.5, respectively, in men, and 280 and 182 pathways significantly associated with both exposures. In women there were significant associations between 47 genes and PM10 only, 149 genes and PM2.5 only, and there were 92 genes associated with both exposures. In women there were significant associations between 91 genes and PM10 only, 1,067 genes and PM2.5 only, and there were 498 genes associated with both exposures. We identified two genes in common between long-term PM10 exposure in men and women, namely RAC3 and DNAJB5, respectively ranked as the 209th and 331st most significant genes with PM10 exposure in women (out of 592 genes). Furthermore RAC3 was also significantly associated with long-term PM2.5 exposure in men and DNAJB5 with long-term PM2.5 exposure in women. We did not observe any significant FDR-corrected q-values in the discovery phase of our study.

**Transcriptome Signature in Relation to Long-Term Exposure**

We selected eight genes that were significantly (p < 0.05) associated with long-term PM10 exposure in the microarray study and have a published link with air pollution-related disease (Table 4) for validation in an independent cohort. Of these we could confirm (i.e., they were also significantly associated with PM10 in the validation cohort based on uncorrected p-values, and associations were in the same direction as in the discovery cohort) two
out of eight genes for men [DNA] homolog, subfamily B, member 5 (DNAJB5), and E2F associated phosphoprotein (EAPP) and one out of eight genes for women to be [Rho GTPase Activating protein 4 (ARHGAP4)] borderline significantly (p = 0.0535) associated with PM10 exposure (Table 4). AKAP6 (p = 0.02) and LIMK1 (p = 0.006) were significantly associated with PM10 in women in the validation cohort, albeit with significantly lower expression instead of higher expression as in the discovery cohort. We also tested the same sets of eight genes for each sex for associations with PM2.5 exposure in the validation cohort, since all but one of the candidate genes (PYGO2 in women, which also was not significant for PM10 in the discovery cohort) were significantly associated with long-term PM2.5 exposure in the discovery cohort. For PM2.5 exposure, we could confirm two out of eight genes [DNAJB5 (borderline significant, p = 0.059) and EAPP] for men and four out of eight genes for women [ARHGAP4, PYGO2, sirtuin 7 (SIRT7) and Autophagy related 16-like 2 (ATG16L2)] (see Table S1). Excluding 21 current smokers (14 of 94 women and 7 of 75 men) from the validation cohort did not alter our conclusions, based on the similarity in the effect estimates, apart from expression of ARHGAP4 in association with long-term PM10 exposure (see Table S2).

**Validation Set**

To determine whether gene expression candidate biomarker identified in the discovery cohort were robust exposure markers, we performed ROC curve analysis with long-term PM10 exposure level 24.5 μg/m³ (75th percentile) as cut-off point in men. Figure 3A shows the sensitivity and 1 minus specificity (false positive ratio) of PM10 exposure levels for men in association with the candidate biomarker genes. The model including the eight genes in men had an area under the curve (AUC) value of 0.92 [95% confidence interval (CI): 0.85, 1.00; p = 0.0002]. In women the model including the eight genes had an AUC of 0.86 (95% CI: 0.76, 0.96; p = 0.004) (Figure 3B, cut-off point 25.7 μg/m³). The combined gene set performed better both in men and women than the individual genes. Similarly, for PM2.5 exposure prediction, the model for men had an AUC of 0.91 (95% CI: 0.83, 0.97; p = 0.007) (Figure 3C), the model for women had an AUC of 0.90 (95% CI: 0.81, 0.98; p = 0.0002) (Figure 3D).

**Discussion**

We identified and validated transcriptome signatures that are associated with long-term exposure to particulate air pollution in apparently healthy men and women. These sets of eight sex-specific genes were predictive of exposure in the validation cohort, and including all eight genes in one model provided a better prediction than the eight genes individually. We found DNAJB5 and EAPP in men and ARHGAP4 in women based on a discovery set and a validation analysis to be significantly associated with PM10 exposure. Besides ARHGAP4, our PM2.5-exposure analysis for women identified PYGO2, SIRT7, and ATG16L2 as significantly associated with particulate matter exposure. However, we cannot assume these

**Figure 2.** Venn diagram showing the overlap of all genes significantly associated with long-term PM10 and PM2.5 exposure in men and women in the discovery cohort.

<table>
<thead>
<tr>
<th>Rank no.</th>
<th>Gene</th>
<th>PM10</th>
<th>PM2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EAPP</td>
<td>1.15 (0.71, 1.24)</td>
<td>2.45 (1.56, 3.78)</td>
</tr>
<tr>
<td>2</td>
<td>DCTN6</td>
<td>1.23 (1.10, 1.38)</td>
<td>1.31 (1.14, 1.50)</td>
</tr>
<tr>
<td>3</td>
<td>DNAJB5</td>
<td>1.36 (1.14, 1.63)</td>
<td>1.42 (1.18, 1.70)</td>
</tr>
<tr>
<td>4</td>
<td>ISL2</td>
<td>1.55 (1.17, 2.06)</td>
<td>1.83 (1.28, 2.62)</td>
</tr>
<tr>
<td>5</td>
<td>KIAA1814</td>
<td>1.23 (1.07, 1.42)</td>
<td>2.24 (1.37, 3.66)</td>
</tr>
<tr>
<td>6</td>
<td>HDLBP</td>
<td>1.14 (0.94, 1.24)</td>
<td>1.62 (1.21, 2.18)</td>
</tr>
<tr>
<td>7</td>
<td>B3GNT3</td>
<td>1.19 (0.65, 1.34)</td>
<td>1.49 (1.17, 1.91)</td>
</tr>
<tr>
<td>8</td>
<td>ATOH8</td>
<td>1.55 (1.14, 2.10)</td>
<td>1.62 (1.20, 2.18)</td>
</tr>
<tr>
<td>9</td>
<td>LSM12</td>
<td>0.86 (0.74, 0.95)</td>
<td>1.40 (1.13, 1.73)</td>
</tr>
<tr>
<td>10</td>
<td>ZNF187</td>
<td>1.16 (1.04, 1.28)</td>
<td>1.39 (1.13, 1.72)</td>
</tr>
<tr>
<td>11</td>
<td>ARHGAP25</td>
<td>1.11 (0.93, 1.33)</td>
<td>1.35 (1.11, 1.64)</td>
</tr>
<tr>
<td>12</td>
<td>SERF1B</td>
<td>0.83 (0.72, 0.96)</td>
<td>1.34 (1.10, 1.64)</td>
</tr>
<tr>
<td>13</td>
<td>ANXA1</td>
<td>1.19 (0.95, 1.36)</td>
<td>1.27 (1.11, 1.69)</td>
</tr>
<tr>
<td>14</td>
<td>TKT1</td>
<td>1.36 (0.97, 1.71)</td>
<td>1.78 (1.20, 2.62)</td>
</tr>
<tr>
<td>15</td>
<td>PRG2</td>
<td>1.29 (1.07, 1.56)</td>
<td>1.30 (0.93, 1.55)</td>
</tr>
<tr>
<td>16</td>
<td>PER1</td>
<td>1.19 (0.95, 1.43)</td>
<td>0.77 (0.64, 0.92)</td>
</tr>
<tr>
<td>17</td>
<td>GUCAC2B</td>
<td>1.36 (0.89, 1.75)</td>
<td>1.57 (1.15, 2.13)</td>
</tr>
<tr>
<td>18</td>
<td>ST14</td>
<td>1.20 (0.81, 1.57)</td>
<td>1.40 (1.11, 1.78)</td>
</tr>
<tr>
<td>19</td>
<td>COX3</td>
<td>0.64 (0.73, 0.96)</td>
<td>1.27 (1.07, 1.49)</td>
</tr>
<tr>
<td>20</td>
<td>TTC20B</td>
<td>1.20 (0.94, 1.37)</td>
<td>1.46 (1.12, 1.89)</td>
</tr>
</tbody>
</table>

**Table 2.** Top 20 significant genes in association with 5-μg/m³ increase in long-term PM10 and PM2.5 exposure for men and women.

**Note:** Rank no. gene indicates its hierarchy for that particular exposure and sex based on level of significance of the identified association, so gene ranked as no. 1 has the lowest p-value. FC, fold change.
associations indicate causal relations due to the observational nature of our study. ROC analysis revealed excellent separation between individuals with high and low exposure to long-term particulate air pollution using the genes selected for validation. We believe gene expression levels have the potential to be used as biomarkers of exposure and effect with high specificity to link particulate air pollution to its health consequences, as these can be measured at the personal level rather than be obtained through exposure modeling at the population level. Further studies looking at different age and ethnic groups are warranted to explore the capabilities of gene expression levels as predictors in more depth. Longitudinal studies that monitor disease incidence, exposure and gene expression over time would be excellent to provide more insights.

We observed different transcriptomic expression levels in association with particulate air pollution exposure in men and women. Sex-specific differences may be explained by differences in inflammatory responses between men and women. Immunologic differences between men and women have been reported based on gene expression profiles in blood between smokers and nonsmokers, where women seem to have a more specific (involving less extensive pathways) immunologic response to smoking than men (Faner et al. 2014). Furthermore, sex-specific associations were also reported for microarray expression profiles in relation to environmental exposure to diverse compounds such as PCBs, dioxin, benzene, and PAHs (De Coster et al. 2013). The sex-specific associations between PM and gene expression that we observed are in line with previous reports of sex-specific associations with other exposures. As such, prenatal exposure to bisphenol A (BPA) led to differential responses in murine placentae of female and male embryos (Imanishi et al. 2003). Prenatal stress exposure in rats was associated with sex-specific differences in gene expression and behavioral effects in male and female offspring (Van den Hove et al. 2013). This study clearly shows the same biological exposure (i.e., prenatal stress) leads to a highly differential response in male and female offspring.

To date, limited human data is available on microarray gene expression profiling in response to air pollution exposure. However, in an attempt to study the effects of in utero carcinogenic exposures, gene expression profiles in cord blood from 111 babies participating in the Norwegian BraMat cohort were assessed and correlation analyses of gene expression levels with biomarkers of exposure measured showed variable numbers of significantly correlating genes. Overall, separate analyses for male and female newborns resulted in higher numbers of significantly

<table>
<thead>
<tr>
<th>Exposure/pathway</th>
<th>q-Value</th>
<th># measured/ # genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response to elevated platelet cytosolic Ca2+</td>
<td>3.11E-07</td>
<td>75/87</td>
</tr>
<tr>
<td>Prolactin signaling pathway</td>
<td>5.78E-07</td>
<td>61/72</td>
</tr>
<tr>
<td>Platelet degranulation</td>
<td>5.90E-07</td>
<td>71/82</td>
</tr>
<tr>
<td>Leukocyte transendothelial migration</td>
<td>1.25E-06</td>
<td>98/118</td>
</tr>
<tr>
<td>Signaling by insulin receptor</td>
<td>5.18E-06</td>
<td>89/109</td>
</tr>
<tr>
<td>PM10s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-cell communication</td>
<td>1.35E-08</td>
<td>95/130</td>
</tr>
<tr>
<td>Chagas disease (American trypanosomiasis)</td>
<td>1.40E-06</td>
<td>92/104</td>
</tr>
<tr>
<td>Signaling by type 1 insulin-like growth factor 1 receptor (IGF1R)</td>
<td>1.40E-06</td>
<td>76/96</td>
</tr>
<tr>
<td>Signaling by insulin receptor</td>
<td>1.93E-06</td>
<td>96/120</td>
</tr>
<tr>
<td>Insulin receptor signaling cascade</td>
<td>2.33E-06</td>
<td>74/76</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PM10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins</td>
<td>2.08E-04</td>
<td>88/97</td>
</tr>
<tr>
<td>Packaging of telomere ends</td>
<td>3.90E-04</td>
<td>46/53</td>
</tr>
<tr>
<td>Electron transport chain</td>
<td>8.11E-04</td>
<td>94/103</td>
</tr>
<tr>
<td>Respiratory electron transport</td>
<td>9.59E-04</td>
<td>71/76</td>
</tr>
<tr>
<td>Telomere maintenance</td>
<td>1.50E-03</td>
<td>72/81</td>
</tr>
<tr>
<td>PM10s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory electron transport</td>
<td>9.07E-04</td>
<td>81/92</td>
</tr>
<tr>
<td>Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins</td>
<td>1.77E-03</td>
<td>99/113</td>
</tr>
<tr>
<td>Packaging of telomere ends</td>
<td>4.54E-03</td>
<td>45/52</td>
</tr>
<tr>
<td>Proteasome</td>
<td>4.93E-03</td>
<td>41/44</td>
</tr>
<tr>
<td>Transcriptional regulation by small RNAs</td>
<td>4.93E-03</td>
<td>95/106</td>
</tr>
</tbody>
</table>

**Table 3.** The top five significant pathways defined by gene set enrichment analysis for each indicator of exposure.

Note: Pathways were identified using the Gene Set Enrichment Analysis Tool from the online Consensus Pathway Data Base (http://cpgdb.molgen.mpg.de/). #, number.
In vitro experiments using hematopoietic stem cells from sirtuin 7 (SIRT7) knockout mice have shown SIRT7 regulates mitochondrial activity and its inactivation causes reduced quiescence, increased mitochondrial protein folding stress, and compromised reparative capacity of hematopoietic stem cells (Mohrin et al. 2015; Liu and Chen 2015). Mitochondrial DNA and function have been shown to be associated with chronic air pollution exposure in populations of newborns (Janssen et al. 2012) and elderly men (Zhong et al. 2016), hence NAD-dependent deacetylase SIRT7 might provide insight into a molecular mechanism underlying the mitochondrial damage following air pollution exposure. Autophagy related 16-like 2 (ATG16L2) is a core autophagy gene. Previously, we found in newborns epigenetic damage and stimulates the expression of EAPP itself becomes upregulated after DNA damage and stimulates the expression of p21 independently of p53 (Andorfer and Hata 2000). DNAJB5 contains a cysteine-rich domain which renders the protein sensitive to ROS. The protein forms a multiprotein complex together with Trx1 and class II histone deacetylases (HDACs) that functions as a master negative regulator of cardiac hypertrophy (Ago et al. 2008). E2F-associated phospho-protein (EAPP) is a nuclear phosphoprotein that interacts with the activating members of the E2F transcription factor family. In vitro overexpression of EAPP increased the fraction of G1 cells and led to heightened resistance against DNA damage. EAPP itself becomes upregulated after DNA damage and stimulates the expression of p21 independently of p53 (Andorfer and Rotheneder 2011).

In pathway analyses, we identified several respiratory chain related pathways significantly associated with long-term PM10 and PM2.5 exposure in men. Rossner et al. (2015) reported deregulation of expression of respiratory chain, oxidative phosphorylation, and mitochondrial membrane pathways when comparing gene expression profiles in adult nonsmoking men from a heavily polluted area versus a control region in the Czech Republic across different seasons (winter and summer 2009 and winter 2010).

Although sex-related differences have been observed for different environmental pollutants, to our knowledge, this is the first study on microarray gene expression profiles in association with long-term air pollution exposure among middle-aged men and women.

Our study has strengths and limitations. We did our investigations in two independent cohorts for discovery and validation, using the same expression modeling and the gold standard qPCR as validation tool (Canales et al. 2006). Although sample size for the discovery cohort was limited, we believe validation in an independent cohort based on a reliable method such as qPCR indicates the robustness of our analyses. Our study also has its limitations inherent to the cross-sectional

### Table 4. Selection of biomarker candidate genes and their fold changes for an increase of 5 μg/m3 long-term PM10 exposure.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene description</th>
<th>Gene function</th>
<th>Link to disease</th>
<th>Discovery cohort FC (95% CI)</th>
<th>p-Value</th>
<th>Validation cohort FC (95% CI)</th>
<th>p-Value</th>
<th>q-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNAJB5</td>
<td>DnaJ (Hsp40) homolog, subfamily B, member 5</td>
<td>Heat shock protein 40</td>
<td>CVD (Ago et al. 2008)</td>
<td>1.36 (1.14, 1.63)</td>
<td>0.0014</td>
<td>1.64 (1.20, 2.23)</td>
<td>0.0026</td>
<td>0.02</td>
</tr>
<tr>
<td>RAC3</td>
<td>Ras-related C3 botulinum toxin substrate 3 (rho family, small GTP binding protein Rac3)</td>
<td>Regulation of cellular responses (cell growth)</td>
<td>Lung cancer (Liu et al. 2015)</td>
<td>1.25 (1.04, 1.51)</td>
<td>0.024</td>
<td>1.26 (0.94, 1.96)</td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td>EAPP</td>
<td>E2F associated phosphoprotein</td>
<td>Cell cycle/apoptosis</td>
<td>Lung cancer (DeMuth et al. 1998)</td>
<td>1.15 (1.0, 1.24)</td>
<td>0.0055</td>
<td>1.18 (1.02, 1.38)</td>
<td>0.028</td>
<td>0.12</td>
</tr>
<tr>
<td>HDLBP</td>
<td>High density lipoprotein binding protein (viginin)</td>
<td>Sterol metabolism</td>
<td>CVD (Husten 1998)</td>
<td>1.14 (1.04, 1.24)</td>
<td>0.0065</td>
<td>1.02 (0.88, 1.19)</td>
<td>0.75</td>
<td>0.86</td>
</tr>
<tr>
<td>PRG2</td>
<td>Proteoglycan 2</td>
<td>Eosinophil major basic protein</td>
<td>CVD (Melchior et al. 2013), asthma (Li et al. 2006)</td>
<td>1.29 (1.07, 1.56)</td>
<td>0.012</td>
<td>1.29 (0.98, 1.71)</td>
<td>0.066</td>
<td>0.18</td>
</tr>
<tr>
<td>PER1</td>
<td>Period homolog 1 (Drosophila)</td>
<td>Circadian rhythm</td>
<td>CVD (Young et al. 2001)</td>
<td>1.19 (1.05, 1.36)</td>
<td>0.012</td>
<td>0.95 (0.74, 1.23)</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td>PKR2R1</td>
<td>Phosphoinositide-3-kinase, regulatory subunit 1 (p85 alpha)</td>
<td>Insulin metabolism</td>
<td>Lung cancer (Lu et al. 2006)</td>
<td>1.22 (1.03, 1.43)</td>
<td>0.023</td>
<td>1.01 (0.82, 1.26)</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>SLA2</td>
<td>SRC-like adaptor 2</td>
<td>SLAP adapter protein</td>
<td>CVD (Chervokova et al. 2015)</td>
<td>1.22 (1.03, 1.44)</td>
<td>0.027</td>
<td>1.16 (0.97, 1.39)</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AKAP6</td>
<td>A kinase (PRKA) anchor protein 6</td>
<td>Regulatory subunit of protein kinase A</td>
<td>CVD (Oti et al. 2006)</td>
<td>1.21 (1.07, 1.36)</td>
<td>0.0036</td>
<td>0.72 (0.55–0.94)</td>
<td>0.017</td>
<td>0.05</td>
</tr>
<tr>
<td>LIMK1</td>
<td>LIM domain kinase 1</td>
<td>Regulation of actin filament dynamics</td>
<td>Lung cancer (Chen et al. 2013), Alzheimer’s (Heredia et al. 2006)</td>
<td>1.28 (1.06, 1.55)</td>
<td>0.01</td>
<td>0.75 (0.61–0.91)</td>
<td>0.0057</td>
<td>0.03</td>
</tr>
<tr>
<td>SIRT7</td>
<td>Sirtuin (silent mating type information regulation 2 homolog) 7 (S. cerevisiae)</td>
<td>Transcription repressor</td>
<td>CVD (Vakhrucheva et al. 2008)</td>
<td>0.89 (0.82, 0.96)</td>
<td>0.0038</td>
<td>0.80 (0.6–1.07)</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>ARHGAP4</td>
<td>Rho GTPase Activating protein 4</td>
<td>Regulation of small GTP-binding proteins from the RAS superfamily</td>
<td>Cognition (Huang et al. 2012)</td>
<td>0.88 (0.81, 0.95)</td>
<td>0.0035</td>
<td>0.62 (0.38–1.00)</td>
<td>0.054</td>
<td>0.11</td>
</tr>
<tr>
<td>ATG16L2</td>
<td>Autophagy related 16-like 2 (S. cerevisiae)</td>
<td>Autophagy</td>
<td>CVD (Magné et al. 2015)</td>
<td>0.81 (0.73, 0.90)</td>
<td>0.00028</td>
<td>0.81 (0.59–1.11)</td>
<td>0.19</td>
<td>0.25</td>
</tr>
<tr>
<td>TPM3</td>
<td>Tropomyosin 3</td>
<td>Actin-binding protein</td>
<td>Lung cancer (Rostila et al. 2012)</td>
<td>0.65 (0.48, 0.88)</td>
<td>0.0086</td>
<td>1.02 (0.83–1.26)</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>5-HTR1B</td>
<td>5-Hydroxytryptamine (serotonin) receptor 1B</td>
<td>Neurotransmitter/vasoconstriction</td>
<td>CVD (Iwabayashi et al. 2012)</td>
<td>1.31 (1.08, 1.59)</td>
<td>0.0097</td>
<td>1.28 (0.49–3.34)</td>
<td>0.62</td>
<td>0.71</td>
</tr>
<tr>
<td>PYG02</td>
<td>Pygophus homolog 2</td>
<td>Related to Wnt signaling</td>
<td>Lung cancer (Liu et al. 2013)</td>
<td>0.93 (0.85, 1.01)</td>
<td>0.097</td>
<td>0.75 (0.61–0.92)</td>
<td>0.0078</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Note: Models adjusted for age, BMI, SES, smoking (validation cohort), leukocyte and neutrophil count, daytime of blood sampling and season. p-Values corrected for multiple testing are represented as q-values.
nature of our study. We used 2010–2012 air pollution data to develop our high-resolution exposure models, which we applied to the participants’ baseline addresses (2004). Studies in the Netherlands (Brauer et al. 2003), Italy (Rosenlund et al. 2008), the United Kingdom (Briggs et al. 2000), and Canada (Vancouver) (Henderson et al. 2007) have shown that during periods of about 10 years and longer, existing land use regression models predicted historic spatial contrasts well. The use of a relatively homogenous population limits the potential generalizability of our study to populations with different ages, races and ethnicities, or locations. Lastly, our study design did not allow to control for cell counts in the discovery phase of the study. As cell counts were not performed on the samples for microarray analysis, and there is no good means for imputation of these values for Agilent 4 × 44 K arrays during data analysis, we were not able to control for this.

This is the first time levels of gene expression of candidate genes have been used to accurately predict air pollution exposure levels (PM$_{10}$, PM$_{2.5}$). For this purpose, we have established ROC curves based on the genes selected for validation in an independent cohort and were able to separate low- (< 75th percentile) from high- (> 75th percentile) exposed individuals. ROC curves are commonly used to compare the diagnostic performance of two or more tests, as they give a good indication of both the sensitivity and specificity of the studied test (Greiner et al. 2000). As such, it has been demonstrated that gene expression signatures can predict survival for instance in pancreatic (Newhook et al. 2014) or non-small cell lung cancer (Lu et al. 2006). In 2009, this technique was applied for the first time in an environmental epidemiology setting, showing that specific DNA methylation patterns could accurately predict the relationship between exposure to airborne PAHs and childhood asthma incidence. Perera et al. (2009) investigated PAH levels in cord blood samples from 20 newborns and replicated the association between PAH levels and candidate region methylation in 56 other newborns from the Columbia Center for Children’s Environmental Health (CCCEH) cohort that recruits nonsmoking Dominican and African American women and their children residing in different areas of New York in the United States (Perera et al. 2009). However, the application of this approach to the field of gene expression data in association with air pollution exposure is novel.

In ROC curve analysis, an AUC of 0.80 is considered a ROC curve with good separation characteristics, and an AUC of 0.90 is considered excellent, in its ability to distinguish between true and false positives. We have identified sex-specific gene-sets that fulfill these criteria for PM$_{10}$ and PM$_{2.5}$ exposure. However, we must interpret the current set

![Figure 3. Receiver operating characteristics (ROC) curve for leukocyte gene expression of gene sets distinguishing between high and low long-term PM$_{10}$ or PM$_{2.5}$ exposure, respectively, based on the eight genes selected for validation for each sex. (A) performance of gene set consisting of DNAJB5, RAC3, SLA2, HDLBP, PRG2, PER1, PIK3R1, and EAPP to distinguish between high and low PM$_{10}$ exposure in men (above 75th percentile corresponding to 24.5 μg/m$^3$) and low (< 24.5 μg/m$^3$) and (B) performance of gene set consisting of ARHGAP4, AKAP6, PYGO2, HTR1B, ATG16L2, SIRT7, TPM3 and LIMK1 in women to distinguish between high (above 75th percentile corresponding to: 25.7 μg/m$^3$) and low (< 25.7 μg/m$^3$) long-term residential PM$_{10}$ exposure. (C) Performance of same male-specific gene set in men and (D) female-specific gene set in women to distinguish between high (above 75th percentile corresponding to: 16.0 μg/m$^3$) and low (< 16.0 μg/m$^3$) long-term residential PM$_{2.5}$ exposure.](image-url)
Conclusions

In conclusion, microarray analysis has identified different gene-expression levels in response to long-term air pollution in men and women. From gene-level analysis, candidate biomarker genes with a reported link to AP-related disease were selected and validated (i.e., significantly associated with PM exposure with the same direction of regulation of expression) in an independent cohort. For men, we propose DNAJBS and EAPP as biomarkers of exposure. For women, we identified ARHGPAP4, PYGO2, SIRT7, and ATG16L2 as biomarker genes of exposure.

ROC analysis revealed that the genes were able to predict high or low PM10 exposure accurately. Prospective studies in other populations are needed to confirm our findings with regard to sex-specific expression of these gene associations with PM exposure. Furthermore, it would be highly relevant to analyze the gene expression of these sex-specific gene-sets in cohorts with higher PM exposure as well as in subjects at different stages of life, including the more vulnerable stages such as early childhood and puberty.

REFERENCES


