Effects of exposure to airborne pollution on human gene expression

by Matt Whitaker

Submission date: 30-Aug-2019 12:37PM (UTC+0100)
Submission ID: 110312498
File name: f_exposure_to_airborne_pollution_on_human_gene_expression_V7.pdf (10.42M)
Word count: 24722
Character count: 140281

Imperial College London

Effects of exposure to airborne pollution on human gene expression

Masters in Health Data Analytics and Machine Learning

CID Number: 01626516

Word count: 9,764

Date of submission: 30 August 2019

ACKNOWLEDGEMENTS

I want to thank my advisor, [name], for being so patient and generous with his time throughout this project. His guidance was invaluable. Thanks also to Dr Marc Chadeau-Hyam for all his feedback on this work, for taking a punt and accepting me on this course, and for making statistics so much fun; to Barbara Bodinier for turning us all into coders; and to everyone on the 2018–2019 HDA course for all the interesting debates.

Finally, thank you to my parents and to my sister, for their love and support in this and everything else.

Support received by supervisor

[name] provided the data for this study, including modelled exposure values. [name] had also conducted some exploratory analysis of the data and proposed some avenues of enquiry. [name] also provided regular feedback through weekly Skype meetings, and read through one full draft of the report before submission. Dr Marc Chadeau-Hyam also provided feedback on a full draft of the report and guidance during the earlier stages.

ABSTRACT

Introduction

Traffic-related air pollutants (TRAPs) are associated with multiple negative health outcomes, but the precise biological mechanisms are largely unknown. In this report, a cross-sectional study of a twin cohort from the Netherlands (n=2,438) was analysed to attempt to identify the molecular signature of TRAP exposure at a transcriptomic level. Exposure levels for seven of the most commonly studied TRAPs and a number of metallic elemental particles were modelled using land-use regression; transcriptomic data were acquired from blood samples; further data were acquired by questionnaire.

Methods

Univariate regression of all transcripts on all TRAPs, adjusted for confounders, was used to identify transcripts that were significantly associated with TRAP exposure. Multivariate methods were used to examine the role of elemental particles. Stability analysis was conducted to test the robustness of the results to input data and model specification. Unsupervised machine learning was used to cluster the data set across a number of different dimensions, and exposure–transcript relationships were analysed within clusters. Bioinformatics and functional analysis were used to propose biological pathways to phenotypic outcomes.

Results

A total of 374 transcripts were identified as significantly associated with exposure to PM2.5 ($p<0.05$). Associations were found to be strongest among residents of less urban areas, with lower exposure levels. The genes most significantly differentially expressed were ZNF791, BTBD1 and OSBPL8, which were upregulated, and WIPF2, which was downregulated. Sensitivity analysis found the results to be robust. Weaker effects were observed in men and in the 28–33 age cohort. Clustering revealed stable clusters within the data relating to exposure levels, geographical variables and demographic variables, and provided a potentially useful mode of stratification for further analysis.

Discussion

Functional analysis showed significant perturbation of phosphoprotein and alternative splicing pathways, which have been used as biomarker predictors of tumour growth and cancer treatment success. Increased effects in women supports previous research indicating women are more severely affected by air pollution than men, although no conclusions can be drawn about whether the effect is biological or behavioural/environmental. Clustering methods suggest a useful adjunct to traditional analytical methods in computational epidemiology, and highlight some weaknesses in land use regression for exposure modelling.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **2**

ABSTRACT **3**

INTRODUCTION **10**

Air pollution and public health *10*

OMICs *11*

Research context: air pollution and the internal exposome *12*

Research goal *13*

METHODS **15**

Data cleaning and preparation *19*

Statistical methods *20*

Investigating elemental particulate matter *23*

Air pollution and smoking: Similar molecular signatures? *25*

Gene enrichment and functional analysis *25*

Unsupervised machine learning approaches *26*

RESULTS **28**

Descriptive statistics *28*

Exposures *29*

Univariate results *31*

Sensitivity analysis *35*

Stratified analyses *37*

Investigating gender-specific associations *39*

The effect of elemental particulate matter *44*

Air pollution and smoking: similar molecular signatures? *49*

Gene enrichment and functional analysis *50*

Unsupervised machine learning results *53*

DISCUSSION **60**

Findings *60*

Study design and technical comments *68*

REFERENCES **71**

APPENDICES **79**

Appendix A *79*

Appendix B *82*

Appendix C *87*

Appendix D *96*

TABLE OF FIGURES

FIGURE 1 SUMMARY OF RELATIVE RISKS FOR MORTALITY BY DIFFERENT AIR POLLUTANTS FROM MULTIPLE EPIDEMIOLOGICAL STUDIES. RELATIVE RISK IS PER 10 MG/M3 INCREASE IN POLLUTION. 95% CONFIDENCE INTERVALS ARE SHOWN. REPRODUCED, WITH PERMISSION, FROM *GLOBAL SOURCES OF LOCAL POLLUTION*[30], CHAPTER 3, PAGE 73. DATA FROM A WHO TASK GROUP REPORT, 2004.[31] 10

FIGURE 2 DENSITY PLOTS SHOWING DISTRIBUTION OF (LOGGED) TRAPS 30

FIGURE 3 JOINT DISTRIBUTIONS AND CORRELATION SCORES FOR SEVEN PRINCIPAL TRAPS (UNLOGGED) 30

FIGURE 4 BOXPLOTS SHOWING DENOISED EXPRESSION LEVELS IN THE FOUR MOST SIGNIFICANTLY AFFECTED TRANSCRIPTS, PLOTTED AGAINST MODELLED PM2.5 EXPOSURE QUARTILES. WHILE THE ASSOCIATIONS ARE SIGNIFICANT, THE ABSOLUTE CHANGE IN EXPRESSION LEVELS IS QUITE SMALL. 32

FIGURE 5 VOLCANO PLOTS SHOWING BETA COEFFICIENTS AND P-VALUES FOR UNIVARIATE REGRESSION OF ALL MAIN TRAP PARTICLES AND ALL TRANSCRIPTS. THE FOUR MOST SIGNIFICANT ASSOCIATIONS ARE LABELLED IN EACH PLOT WITH THE NAME OF THE GENE ASSOCIATED WITH THE TRANSCRIPT. SIGNIFICANT ASSOCIATIONS, AFTER BH MULTIPLE TESTING ADJUSTMENT, ARE OBSERVED BETWEEN PM2.5 AND 374 TRANSCRIPTS, BUT NOT BETWEEN ANY OTHER EXPOSURES. 33

FIGURE 6 A CLUSTERED HEATMAP SHOWING CORRELATION BETWEEN 374 TRANSCRIPTS SIGNIFICANTLY ASSOCIATED WITH PM2.5 IN UNIVARIATE ANALYSIS. HIERARCHICAL CLUSTERING SHOWS TWO DISTINCT GROUPS OF TRANSCRIPTS: THE 285 UPREGULATED AND THE 89 DOWNREGULATED IN UNIVARIATE ANALYSIS, AS WELL AS CLUSTERS OF HIGHLY CORRELATED TRANSCRIPTS WITHIN THE UPREGULATED CLUSTER 34

FIGURE 7 P-VALUE DISTRIBUTIONS: CAPPED VS UNCAPPED UNIVARIATE MODELS. DISTRIBUTIONS ARE HIGHLY CONSISTENT IN CAPPED AND UNCAPPED MODELS. 36

FIGURE 8 BETA COEFFICIENT DISTRIBUTIONS: CAPPED VS UNCAPPED MODELS. DISTRIBUTIONS ARE HIGHLY CONSISTENT IN CAPPED AND UNCAPPED MODELS. 36

FIGURE 9 PLOTS COMPARING THE DISTRIBUTION OF P-VALUES AND BETA COEFFICIENTS IN STRATIFIED RESULTS VERSUS UNSTRATIFIED, INVESTIGATING THE ASSOCIATIONS BETWEEN PM2.5 AND TRANSCRIPTS. SIGNIFICANT ASSOCIATIONS (BH P-VALUE <0.2 – LOWERED TO REVEAL MORE ASSOCIATIONS IN A LOWER-POWERED, SMALLER N ANALYSIS) ARE HIGHLIGHTED IN RED. FEWER SIGNIFICANT RESULTS ARE OBSERVED IN MEN, AND IN THE 28–33 STRATUM. 38

FIGURE 10 DISTRIBUTIONS OF P-VALUES IN MALE AND FEMALE COHORTS (RED = FEMALE; BLUE = MALE). THICK LINES SHOW P-VALUE DISTRIBUTION IN THE FULL MALE/FEMALE COHORTS. EACH FAINT LINE SHOWS A P-VALUE DISTRIBUTION IN A RANDOM SUB-SAMPLE OF 500 MALES OR FEMALES. FEMALE SUBSAMPLES ARE SHOWN TO HAVE P-VALUE DISTRIBUTIONS MORE WEIGHTED TOWARDS LOWER VALUES (HIGHER SIGNIFICANCE), MORE OFTEN. 39

FIGURE 11 DISTRIBUTION OF NUMBER OF SIGNIFICANT ASSOCIATIONS (P<0.001) FOUND IN RANDOM N=500 SUB-SAMPLES OF MALE AND FEMALE OBSERVATIONS. MORE SIGNIFICANT ASSOCIATIONS ARE OBSERVED IN FEMALE SUBSAMPLES..... 40

FIGURE 12 PLOTS COMPARING THE DISTRIBUTION OF P-VALUES AND BETA COEFFICIENTS IN URBANICITY-STRATIFIED RESULTS VERSUS UNSTRATIFIED, INVESTIGATING THE ASSOCIATIONS BETWEEN PM2.5 AND TRANSCRIPTS. SIGNIFICANT ASSOCIATIONS (BH P-VALUE <0.2) ARE HIGHLIGHTED IN RED..... 42

FIGURE 13 VOLCANO PLOTS SHOWING P-VALUES AND BETA COEFFICIENTS FOR SIGNIFICANT TRANSCRIPTS IN A CONTROLLED UNIVARIATE MODEL, INVESTIGATING THE ASSOCIATIONS BETWEEN PM2.5 AND TRANSCRIPTS, STRATIFIED BY URBANICITY. ONLY 374 SELECTED TRANSCRIPTS ARE SHOWN. STRONGER ASSOCIATIONS ARE SEEN IN AREAS OF LOWER URBANICITY, ESPECIALLY STED 3 AND STED 5. 43

FIGURE 14 VOLCANO PLOTS SHOWING RESULTS FROM UNIVARIATE ANALYSIS OF 374 SIGNIFICANT TRANSCRIPTS AND EIGHT ELEMENTAL PM2.5 PARTICLES. X-AXIS IS SHOWN ON THE SAME SCALE AS THE VOLCANO PLOTS IN THE UNIVARIATE ANALYSIS OF TOTAL PM2.5 IN **FIGURE 5**, TO DEMONSTRATE THE DIFFERENCE IN EFFECT SIZE. THE MOST SIGNIFICANT RESULTS CAN BE SEEN IN COPPER (CU), SULPHUR (S), SILICON (SI) AND IRON (FE)..... 45

FIGURE 15 P-VALUES FOR UNIVARIATE REGRESSION MODELS FOR ELEMENTAL PM2.5 PARTICLES PLOTTED AGAINST P-VALUES FOR UNIVARIATE REGRESSION OF PM2.5 ONLY THE 374 TRANSCRIPTS SELECTED BY UNIVARIATE ANALYSIS ARE SHOWN..... 46

FIGURE 16 BETA COEFFICIENTS FOR UNIVARIATE REGRESSION MODELS FOR ELEMENTAL PM2.5 PARTICLES PLOTTED AGAINST BETA COEFFICIENTS FOR UNIVARIATE REGRESSION OF PM2.5. EFFECTS ARE HIGHLY CONSISTENT IN DIRECTION, BUT SMALLER IN SIZE IN ELEMENTAL PARTICLES THAN IN TOTAL PM2.5. ONLY THE 374 TRANSCRIPTS SELECTED BY UNIVARIATE ANALYSIS ARE SHOWN. 46

FIGURE 17 RESULTS OF STABILITY ANALYSIS, REGRESSING 4 SIGNIFICANT TRANSCRIPTS SEPARATELY ONTO PM25 ELEMENTAL PARTICLES USING REPEATED ELASTICNET REGRESSION..... 47

FIGURE 18 POSTERIOR DISTRIBUTION OF THE REGRESSION COEFFICIENT FOR PM2.5_CU (ELEMENTAL COPPER PARTICULATE MATTER). PLOTS SHOW UPREGULATION OF TRANSCRIPTS CODED BY ZNF791, BTBD1 AND OSBPL8, AND DOWNREGULATION OF A TRANSCRIPT CODED BY WIPF2. 48

FIGURE 19 MPPI OUTPUT FROM R2GUESS WHEN EIGHT ELEMENTAL PM2.5 PARTICLES ARE USED AS PREDICTORS FOR FOUR SIGNIFICANT TRAPS. ELEMENTAL COPPER PM2.5 IS THE ONLY SELECTED PREDICTOR VARIABLE... 48

FIGURE 20 GENE ONTOLOGY BIOLOGICAL PROCESS ANALYSIS USING BINGO[96] THE NUMBER OF GENES INVOLVED IN THE PROCESS IS INDICATED BY THE SIZE OF THE NODE. THE SIGNIFICANCE OF THE OVER-REPRESENTATION OF THE SELECTED GENE SET IN THE RELEVANT BIOLOGICAL PROCESS IS REPRESENTED THROUGH THE COLOUR OF THE NODE (DARKER = LOWER ADJUSTED P-VALUE)..... 51

FIGURE 21 A SANKEY DIAGRAM SHOWING HOW STUDY PARTICIPANTS ARE GROUPED WITHIN CLUSTERS ACROSS FOUR DIMENSIONS FROM LEFT TO RIGHT, DIMENSIONS ARE: TRAPS, GEOGRAPHY, TRAFFIC, DEMOGRAPHIC. THE DIAGRAM CLEARLY SHOWS HOW A VERY SIMILAR GROUP OF PARTICIPANTS IN CLUSTERED TOGETHER ACROSS THREE DIMENSIONS (THE SECOND CLUSTER IN TRAPS, GEOGRAPHY AND TRAFFIC). 54

FIGURE 22 JOINT DISTRIBUTION OF EXPOSURE LEVELS IN GEOGRAPHICAL CLUSTERS, FOR SEVEN PRINCIPAL TRAPS. FOR ALL TRAPS, EXPOSURE IS HIGHEST IN CLUSTER 2, AND LOWEST IN CLUSTER 1, ALTHOUGH THERE IS LESS VARIATION IN PM2.5 LEVELS BETWEEN CLUSTERS THAN THE OTHER TRAPS. PLOTS ALONG THE TOP AND LEFT MARGINS SHOW THE MARGINAL DISTRIBUTIONS OF EACH EXPOSURE, GROUPED BY GEOGRAPHICAL CLUSTER. SUB-DIAGONAL PLOTS SHOW JOINT DISTRIBUTIONS FOR ALL EXPOSURES, WITH THE COLOUR OF EACH DATA POINT INDICATING WHICH GEOGRAPHICAL CLUSTER THE POINT BELONGED TO. NUMBERS REPRESENT CORRELATIONS BETWEEN EXPOSURES, BROKEN DOWN WITHIN GEOGRAPHICAL CLUSTERS. 55

FIGURE 23 JOINT DISTRIBUTION OF EXPOSURE LEVELS IN GEOGRAPHICAL CLUSTERS, FOR ELEMENTAL PM2.5 PARTICLES. EXPOSURE LEVELS FOLLOW A SIMILAR PATTERN TO THE PRINCIPAL TRAPS – WITH CLUSTER 2 BEING THE HIGHEST EXPOSED AND CLUSTER 1 THE LOWEST – WITH THE EXCEPTION OF ZINC AND POTASSIUM. PLOTS ALONG THE TOP AND LEFT MARGINS SHOW THE MARGINAL DISTRIBUTIONS OF EACH ELEMENTAL PARTICLE EXPOSURE, GROUPED BY GEOGRAPHICAL CLUSTER. SUB-DIAGONAL PLOTS SHOW JOINT DISTRIBUTIONS FOR ALL EXPOSURES, WITH THE COLOUR OF EACH DATA POINT INDICATING WHICH GEOGRAPHICAL CLUSTER THE POINT BELONGED TO. NUMBERS REPRESENT CORRELATIONS BETWEEN ELEMENTAL PARTICLE LEVELS, BROKEN DOWN WITHIN GEOGRAPHICAL CLUSTERS. 56

FIGURE 24 VARIABLE IMPORTANCE IN RANDOM FOREST REGRESSION MODELS, USING GEOGRAPHICAL VARIABLES TO PREDICT WHICH TRAP CLUSTER A PERSON WILL BE ASSIGNED TO. THE MODEL NOTABLY SELECTS THE VARIABLES WHICH OPERATE AT THE BROADEST GEOGRAPHICAL SCALE. 57

FIGURE 25 SCATTER PLOTS SHOWING LOG10 P-VALUES FOR ADJUSTED UNIVARIATE ANALYSIS OF 374 TRANSCRIPTS AGAINST 7 PRINCIPAL TRAPS, STRATIFIED BY CLUSTER. RESULTS SHOW STRONG ASSOCIATIONS IN DEMOGRAPHIC CLUSTER 2, GEOGRAPHICAL CLUSTER 1, TRAFFIC CLUSTER 1 AND TRAP CLUSTER 1. 59

FIGURE 26 VARIANCE OF MEASUREMENTS OF O_3 (LEFT) NO_2 (MIDDLE) AND PM10 (RIGHT), PLOTTED AGAINST BETA-PARAMETERS FROM LAND USE REGRESSION MODELS. VARIANCE INCREASES AS TRAP EXPOSURE INCREASES, AND RURAL AREAS ARE SHOWN TO HAVE THE LOWEST VARIANCE. REPRODUCED, WITH PERMISSION, FROM: JANSSEN *ET AL*. SPATIAL INTERPOLATION OF AIR POLLUTION MEASUREMENTS USING CORINE LAND COVER DATA. ATMOSPHERIC ENVIRONMENT; VOLUME 42, ISSUE 20, JUNE 2008, PAGES 4884-4903.[152] 64

FIGURE 27 CAR OWNERSHIP IN THE NETHERLANDS IN 2015 REPRODUCED FROM *TRENDS IN THE NETHERLANDS 2018*, CENTRAL BUREAU OF STATISTICS NETHERLANDS, 2018[163] 66

FIGURE 28 CHART SHOWING MILES TRAVELLED ANNUALLY IN CARS, PER PERSON PER YEAR, IN URBAN AND RURAL AREAS IN THE UK. DATA TAKEN FROM UK GOVERNMENT LAND MANAGEMENT STATISTICS: *TRANSPORT AND TRAVEL IN RURAL AREAS*, 2016.[167] 66

FIGURE 29 BOXPLOTS SHOWING DISTRIBUTION OF MODELLED MAIN TRAP EXPOSURE LEVELS, BY URBANICITY 69

INTRODUCTION

Air pollution and public health

The impact of airborne pollutants on human health is well established in the literature.[1–6] Air pollutants have been shown to cause and exacerbate both acute and chronic conditions, affecting the respiratory[5,7–9] and cardiovascular[1,10–13] systems and increasing the risk of multiple types of cancer,[14–23] as well as increasing risk from all-cause mortality[24,25] (see **Figure 1**). According to the World Health Organization (WHO), 4.2 million people die each year as a direct consequence of ambient (outdoor) air pollution, and 91% of the global population lives in locations where the air quality falls below WHO quality guideline levels.[26] The importance of this issue to global public health will only increase in coming decades as the global population becomes increasingly urbanised[27] and the burning of fossil fuels remains the dominant mode of energy production.[28,29]

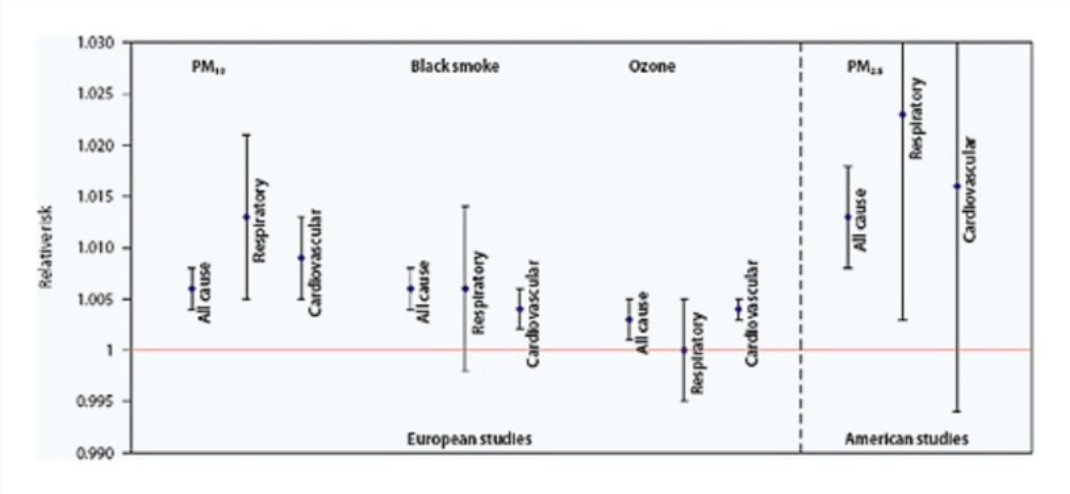

Figure 1 Summary of relative risks for mortality by different air pollutants from multiple epidemiological studies. Relative risk is per 10 μg/m3 increase in pollution. 95% confidence intervals are shown.
Reproduced, with permission, from *Global Sources of Local Pollution*[30], Chapter 3, page 73. Data from a WHO task group report, 2004.[31]

The International Energy Agency (IEA) estimates that 85% of airborne respirable particulate pollution, and nearly all NO_2 emissions, are generated by energy-related fossil-fuel combustion

(predominantly in richer countries) and biomass burning (predominantly in poorer countries).[32] This cluster of fossil-fuel-originated pollutants – include gaseous compounds such as nitrogen dioxide (NO_2) and nitric oxide (NO), as well as various types of particulate matter (PM) – are often grouped and studied as traffic-related air pollutants (TRAPs).

While the association between these traffic-related air pollutants and negative health outcomes is well established, the biological pathways linking these exposures and subsequent health outcomes are not well understood.

OMICs

Developments in high-throughput computational methods have allowed the biological pathways between exposure and health outcomes to be explored with increasing precision. The emergent field of 'OMICs' involves the identification of biomarkers at several molecular levels that are associated either with a particular exposure or with a particular phenotypic outcome.

The methodologies in the OMICs field evolved from genomics, where genome-wide analysis techniques sought to identify relationships between genetic variations and phenotypic outcomes.[33] In the past two decades the methodologies developed in genomics have been applied to other sets of markers at different molecular levels, the most well established being the epigenome, transcriptome, proteome, metabolome and microbiome.

Unlike the genome, which is heritable and immutable, the human profile at other OMICs levels is subject to varying degrees of response to environmental factors: exposures.[34,35] The concept of the 'exposome' to describe the totality of these external exposures was proposed by Wild[36] in 2005, and the concept has since been refined to describe:

1. The **external** exposome: the totality of general and specific external (ie non-genetic) environmental factors to which a person is exposed; and
2. The **internal** exposome: the biological perturbations occurring at a molecular level in response to external exposures.[37,38]

There are, therefore, two 'ways in' to mapping biological pathways between exposure and health outcome: one can either ask how the external exposome impacts the internal exposome; or, one can ask which internal exposome features might represent a marker of a phenotypic outcome. (There

have been some analyses which have taken both ways in at the same time, a 'meet-in-the-middle' approach,[39] but to date these are the exception rather than the rule.)

Statistical methods have been developed to address the specific challenges of each 'way in' to OMICs analysis.[40] For internal exposome–phenotype analysis, the principal challenge is solving the problem of dimensionality: most OMICs levels have tens of thousands of correlated biomarkers that might be used to 'predict' a health outcome, and between 10 and a few thousand observations, depending on the study design. Penalised multivariate methods, Bayesian variable selection and dimensionality reduction methods such as principal component analysis help to address this problem.

For external exposome–internal exposome analysis – the analysis used in this study – the challenge is different. The study may involve either one or a handful of 'predictor' variables, the external exposures, and many thousands of dependent variables – the individual markers in whichever OMICs level is under analysis. To date, the gold standard statistical approach is large-scale univariate analysis, where each marker is regressed separately onto the exposure or exposures (controlling for potential confounders by including them as covariates in the regression), and the significance of the association determined using a p-value derived from the regression. This approach brings its own problems. Running thousands of tests under the same hypothesis inflates the risk of false positive findings, and the multiple-testing correction methods used to control for this eventuality diminish the statistical power of the models and increase the likelihood of missing relevant associations (false negatives). This is particularly important when the dependent variables are highly correlated and the correction method used does not account for this fact.[41] The univariate model also necessarily captures a simplified picture of what are usually highly complex relationships between OMICs markers and multiple exposures. Therefore, an initial univariate analysis is often complemented by more sophisticated techniques – multivariate methods, network analysis or bioinformatics – to capture more of the detail of the patterns of association.

Research context: air pollution and the internal exposome

A growing body of literature exists that deploys these statistical methods to analyse the impact of TRAP exposure at various OMICs levels, including the genomic,[42] the epigenomic,[43–49] the transcriptomic,[43,45,50–60] the proteomic[42] and the metabolomic.[43,61] There are two broad approaches to study design in this kind of research:

1. Epidemiological investigation of chronic, long-term exposure. This usually involves a cross-sectional or cohort study design, with a larger number of participants. Air pollution exposure may be estimated or modelled using land-use regression models or location-based measurements.
2. Investigation of short-term, acute exposure. This may involve the use of personal environmental monitors to get individual-specific exposure measurements for a small cohort over a short space of time, or controlled interventional studies where some individuals are deliberately exposed and compared to a control group.

Both modes of investigation come with problems. Large epidemiological studies are subject to influence from unmeasured confounders and may suffer from a lack of precision in measurement; acute exposure studies are expensive to conduct and therefore usually involve small numbers of observations, with a concomitant lack of statistical power.

This situation, combined with the innate complexity and scale of OMICs-based analysis, has resulted in a relative lack of replication in this young field of study. Among studies analysing the transcriptome, distinct molecular signatures have been found to be associated with exposure to NO_x[52,60] and, more commonly, PM2.5.[45,51,53,56–58] However, there is little overlap in the patterns of biomarkers identified in each study. The bulk of previous research has fallen into the second of the two categories, and analysed acute transcriptomic responses to short-term exposure; to date, only one published study has looked at the impact of long-term air pollution exposure at a transcriptomic level.[52]

Research goal

This study falls into the first of the two categories described above. The data comes from a large cohort of twins and twin families from the Netherlands, from whom transcriptomic profiles and biological data were gathered and analysed in relation to air pollution exposures based on land-use regression (LUR) models. The goal of this research is to identify the molecular signature of long-term exposure to air pollution at a transcriptomic level, and to identify biological pathways that are affected by these exposures.

As a first approach to the analysis, 'classical' epidemiological methods were employed. Univariate regression models were used to attempt to identify an initial set of transcripts that were significantly associated with exposure to any of seven principal TRAPs, which consisted of particulate matter of

varying sizes, and nitrogen oxides. Stability analysis was used to test the sensitivity of any findings to the features of the population and determine whether any observed effects are attributable to TRAP exposure or whether TRAP exposure may be functioning as a marker for, or being confounded by, other variables relating to environment. The composition of particulate matter was also investigated: the precise make-up of particulate matter pollutants can vary significantly from place to place and can affect the biological response,[62] so multivariate regression methods were used to analyse the role that the elemental composition of the particulate matter played in any significant associations.

To complement the classical approaches, machine learning techniques were used to cluster the data using different groups of variables, and to investigate whether clustering can reveal further patterns of the associations linking exposures and transcriptomic response.

The use of a twin cohort in this study was intended to help ameliorate some of the problems of confounding that epidemiological studies suffer from. Twins share DNA (100% of it in the case of monozygotic twins, 50% for dizygotic twins) and are highly likely to have had similar upbringings and early-life environmental exposures, so twin cohorts provide a natural, highly controlled setting for epidemiological studies, and allow more powerful exploration of variables that differ between twin pairs.

Gene enrichment analysis and functional analysis was performed to propose biological pathways that may be perturbed by air pollution exposure, and, in conjunction with existing research findings in the literature, to hypothesise likely health outcomes and to provide insight into the pathways between TRAP exposure and known health outcomes. Biomarkers and pathways identified in this analysis were also compared with previously published research in the literature to attempt to identify any consistent findings.

METHODS

Data for this study came from the Netherlands Twin Register[63] biobank, a large-scale biobank designed to provide a resource for genetic epidemiological studies. All patients were healthy and gave informed consent, and the NTR study protocol has been approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam (IRB/ institute codes: NTR 03-180). A total of 2,438 participants were included in the final analysis. **Table 1** describes the study population.

Table 1 Study participants and key characteristics

Variable	Number
Gender	
Male	825
Female	1613
Smoking status (%)	
Never smoked	1337
Current smoker	515
Former smoker	586
BMI	
BMI 20–30	1987
BMI <20	250
BMI >30	201
Twin status	
Monozygotic twin	1,104
Dizygotic twin	971
Twin family member	363

DATA GATHERING

Blood samples were collected from adult participants from twin families between January 2004 and July 2008. Both monozygotic ("identical") twins and dizygotic twins were included, as well as other family members of the twins. Additional phenotypic data, including smoking status and BMI, was gathered at the same time as blood sampling via questionnaire. The blood samples were analysed for haematology and for transcriptomic profile.

TRANSCRIPTOMIC PROFILING

RNA isolation and extraction methods for the NTR biobank are described in depth in Willemsen *et al.*[63] To summarise: blood samples were taken from participants between 7am and 10am and stored in PAXgene Blood RNA tubes (Qiagen, Valencia, Florida, USA) at -20°C, before being sent to Rutgers University Cell and DNA Repository. RNA was extracted with Qiagen Universal liquid handling and analysed spectroscopically.

Samples were hybridized to Affymetrix U219 array plates (GeneTitan, Affymetrix, Santa Clara, California, USA). Array hybridisation, washing, scanning and staining was performed in an Affymetrix Genetitan Instrument.

Standard quality control processes were applied according to the Affymetrix Expression Console Software documentation. A total of 44,241 probe sets remained after the quality control process. Multi-array average normalization was implemented in Affymetrix Power Tools (v 1.12.0) to derive expression values. Expression values are a measurement of fluorescence intensity across a probe set, which functions as a measure of the relative expression level of that transcript. Samples with the incorrect sex chromosome expression were removed, as were samples that showed an average Pearson correlation of less than 0.8 with all other samples of the same transcript.

URBAN EXPOSOME ESTIMATES

Participants' addresses at the time of blood sampling were geocoded and used to estimate levels of exposure to air pollutants, as well as other geographical and exposure-related variables.

Estimating TRAP exposure

Estimates of air pollution levels were generated using Land Use Regression (LUR) models pre-established in the literature.

Exposure levels were modelled for seven air pollutants commonly associated with health impacts: nitrogen oxides (NO_x), nitrogen dioxide (NO_2), particulate matter (PM10), fine particulate matter (PM2.5), the absorbance coefficient of PM2.5 (PM25_abs – used as a proxy for elemental carbon levels),[64,65] coarse particulate matter (PM_coarse), and ultra-fine particles (UFP). For ultrafine particles (UFP), 10-fold cross validation was used to generate estimates.

Elemental particles

Particulate matter is a heterogenous mixture of many different constituent particles, the precise constitution of which varies significantly from place to place and over time.[66] Previous research has observed that the elemental make-up of particulate matter is a significant factor in determining the health impact of PM.[62,67,68] Therefore, in addition to modelling the seven most commonly studied air pollutants, the constituent elemental particles that make up particulate matter (PM10 and PM2.5) were also modelled and analysed., **including** elemental particulate matter.

Table 2 shows the models used for each particle estimate, including elemental particulate matter.

Table 2 Models used for TRAP exposure and elemental particle estimates

Exposure	Description	Reference
NO_2	Nitrogen dioxide; primary air pollution marker	*Atmospheric Environment* 2013; **72:** pp10–23.[69]
NO_x	Nitrogen oxides; Primary air pollution marker	*As above*
PM25	Particulate matter with a diameter <2.5 um; Primary air pollution marker	*Environ Sci Technol* 2012; **46(20):** pp11,195–11,205.[70]
PM25_abs	Absorbance measured in pm2.5	
PM10	Particulate matter with a diameter <10 um; Primary air pollution marker	*As above*
PM_coarse	Particulate matter with a diameter 2.5–10 um;	*As above*
UFP	Ultrafine particle count (Particulate matter with a diameter <0.1 um) *(Ten-fold cross-validation models for UFP)*	*Environ Sci Technol* 2017; **51(6):** 3,336–3,345.[71]
Elemental particles		
PM10_Cu PM10_Fe PM10_K PM10_Ni PM10_S PM10_Si PM10_V PM10_Zn	Elemental composition in PM10	*Environ Sci Technol* 2013; **47(11):** pp 5,778–5,786.[72]
PM25_Cu PM25_Fe PM25_K PM25_Ni PM25_S PM25_Si PM25_V PM25_Zn	Elemental composition in PM25	*As above*

Estimating urbanicity and land use

Land use, population density and traffic were estimated using a range of databases and resources: the EU CORINE Land Cover database,[73] the Dutch National Road Database (Nationaal WegenBestand).[74] and the Woningen populatiegrid 2009.[75] A table showing all modelled variables and units of measurement is included in Appendix C.

An overall measure of 'urbanicity' was also modelled using the Netherlands' Central Bureau of Statistics' urbanicity calculation, which classifies geographical locations into one of five levels of urbanicity based on the building density in the surrounding 1km^2.[76]

Estimating greenness

The 'greenness' of each address – the amount of land surrounding the address that was covered by grass, vegetation, forest, etc – was estimated using data from two databases: the normalised difference vegetation index (NDVI)[77] and the TOP10NL – a topographic file developed by the Dutch Cadastre, Land Registry and Mapping Agency that distinguishes between 'agricultural greenness', 'natural greenness' and 'urban greenness'. Estimates were generated for radii of 100, 300, 500, 1,000, and 3,000 metres. A table showing all modelled variables and units of measurement is included in Appendix C.

Data cleaning and preparation

MISSING AND MISCODED DATA

The data set contained a small number of missing observations (numbers shown in **Table 3**). Missing data were imputed with a multiple imputation (MI) algorithm using predictive mean matching (PMM) in the R package 'Mice'.[78] MI has been shown to produce asymptotically unbiased and efficient estimates of missing data.[79] A total of 363 twin IDs ('twzyg') were missing and were not imputed. Participants with missing twin IDs were excluded from further analysis.

For some data, implausible values (eg -999999999 for some geographical variables) or factor values that did not correspond to classes in the data dictionary (eg -1 for smokers) were assumed missing, recoded with NA and imputed using PMM.

Table 3 Missing data by variable

Variable	NA count	Variable	NA count
bioheight	1	TLOA_50	2
bioweight	9	TLOA_100	1
biobmi	10	HLOA_50	4
biowbc	34	HLOA_100	4
bioneut	36	HLOA_300	2
biolymp	36	DINVNEARC2	2
biomono	36	DINVMAJOC1	1
bioeos	36	DINVMAJOC2	5
biobaso	36	OPP_WATER	761
biotbc	30	P_ONGEHUWD	4
twzyg	363	P_GEHUWD	4
INTINVD2	2	P_GESCHEID	4
HINTINVD	6	P_VERWEDUW	4
DINVNEAR1	2	P_WEST_AL	4
DINVNEAR2	2	P_N_W_AL	4
HINTINVD2	2	P_MAROKKO	197
TRAFNEAR	3	P_ANT_ARU	197
INTINVD	6	P_SURINAM	197
DINVMAJOR2	18	P_TURKIJE	197
DINVMAJOR1	1	P_OVER_NW	197
INTMINVD	6	WOZ	15
TRAFMAJOR	2	P_KOOPWON	3
INTMINVD2	6	P_HUURWON	3
TMLOA_50	2	P_LAAGINKH	24
TMLOA_100	1	P_HOOGINKH	24
HMLOA_50	4	P_LKOOPKRH	35
HMLOA_100	6	P_SOCMINH	35
HMLOA_300	2	P_WWB_UIT	12

Statistical methods

UNIVARIATE METHODS

Initial analysis of univariate relationships between exposures and gene expression was conducted using linear mixed models. For each pair of transcript and exposure, mixed models were used to model gene expression (the dependent variable) as a function of TRAP exposure. Exposures were log transformed to reduce the skewness of the exposure distributions. Two versions of each model were fitted, one with only random effects (denoted as 'CRUDE' in **Table 4**), and one with random effects plus a set of fixed-effect covariates ('ADJ' in **Table 4**). Covariates for the controlled model were selected from a large range of available variables in the NTR data set on the basis of previous research hypothesising which variables were likely to be confounders of the association between exposure and gene expression.[80] The final regression specification was:

$$y\,[transcript] = \overline{\beta_0} * \;\; [matrix\ of\ covariates] + \;\beta_1 \; * \; X\,[TRAP]$$

Where β_0 is a vector of coefficients relating to the covariates, and β_1 is a regression coefficient relating to the variable of interest (TRAP exposure).

Table 4 Covariates for univariate regression models

Covariate	Variable_name	form	Imputed	CRUDE	ADJ
Plate	RNAexpr_plate	random		X	X
Well	RNAexpr_well	random		X	X
Family ID	FamilyID	random		X	X
Month of sampling	biomonth	random		X	X
Sex	sex	Fixed cat			X
Age	Bioage	Fixed lin			X
Smoking status	biosmoke	Fixed cat	X		X
BMI	BMI_cat	Fixed cat	X		X
Lymphocytes count	biolympn	Fixed lin	X		X
Monocytes count	biomonon	Fixed lin	X		X
Neutorphils count	bioneutn	Fixed lin	X		X
Eosinophils count	bioeosn	Fixed lin	X		X
Basophils count	biobason	Fixed lin	X		X
Hemoglobin level	biohgb	Fixed lin	X		X
Days between sampling and extraction	RNAexpr_ndays	Fixed lin			X
Days between extraction and amplification	RNAexpr_ndays_ext_amp	Fixed lin			X
Sampling time, antr biobank	biotime_mod	Fixed lin	X		X

The statistical significance of the regression coefficient measuring the effect of a given exposure on a given transcript was calculated using a likelihood ratio test comparing the fit of the models with and

without the variable of interest (TRAP exposure, in this case). Multiple testing correction was performed by controlling the False Discovery Rate (FDR) at a 0.05 level using the Benjamini-Hochberg (BH) procedure.[81]

SENSITIVITY ANALYSIS

A variety of statistical methods were used to test the robustness of the univariate findings and assess their sensitivity to the input data and model calibration.

Winsorising

To check the sensitivity of the univariate regression results to outliers, significant TRAP–transcript associations identified by univariate regression modelling were re-analysed with (log transformed) exposure distribution capped at 5% and 95% – ie any exposure readings above the 95th percentile or below the 5th percentile were re-coded to the value of the 95th and 5th percentile, respectively (a process known as Winsorising).[82]

Assessing redundancy across TRAPs

TRAP levels often correlate strongly with one another. To assess the effect of each of the TRAPs conditionally on all the others, versions of the univariate model were built which included all principal TRAPs, except the one under analysis, as extra confounders. For example, when analysing NO_2:

$$y\,[transcript] = \overline{\beta_0} * \;\; [matrix\ of\ covariates, incl.\ all\ TRAPs\ except\ NO2] + \overline{\beta_1} * \; X\,[NO2]$$

Additional environmental confounders

To test the extent to which air pollution might be functioning as a proxy for other environmental factors, extended regression models were also tested with additional environmental variables included as confounders: 'top10NL_300m' – the proportion of green land in a 300m radius of the address, based on the TOP10NL database;[83] 'NDVI_300m' – the average normalized difference vegetation index in a 300m radius of the address; and 'P_LKOOPKRH' – a measure of the purchasing power of the household. For example, when analysing NO_2:

$$y\,[transcript] = \overline{\beta_0} * \;\; [covariates, incl.\ all\ TRAPs\ except\ NO2, plus\ 3\ environ.\ covariates] + \overline{\beta_1} * \; X\,[NO2]$$

STRATIFIED ANALYSIS

To investigate the stability of the full-dataset regression results and the role of some key phenotypic and geographical covariates on the TRAP–gene-expression relationship, the same univariate analyses were conducted on data stratified by sex, age, and 'urbanicity' (as represented by 'STED', a variable measuring the housing density in a $1km^2$ area around each study participant's address), and results compared to the results from unstratified analysis. In addition to testing the robustness of the full-cohort univariate findings, stratified analysis addressed three specific hypotheses:

1. The location data for each individual was based on their current address. No information was available on the length of time that an individual had lived at that address, and we may expect younger people to move around more frequently, diluting results.
   Hypothesis: more significant associations will be found in the older age brackets than the younger.
2. There is evidence in the literature of differential impacts of air pollution between men and women, with the majority of findings indicating greater effects in women.[84–88]
   Hypothesis: different effects and strength of associations may be observed in men and women.
3. As previously discussed, the extent to which TRAP exposure is a marker for other factors relating to urbanicity is a key question in this investigation and all studies of air pollution. Stratifying by urbanicity will help to show the extent to which uncontrolled confounders relating to urbanicity may be interfering with the analysis.
   Hypothesis: different effects and strength of associations may be observed in areas of different levels of urbanicity.

Power differences in stratified analysis

Where population sizes in different strata were not balanced there is a resultant imbalance in statistical power, making findings difficult to compare. In the gender-stratified analysis, for example, there were 1,613 women and only 825 men, resulting in a significantly lower-powered analysis for the male stratum. To discover whether any strata differences were actual differential effects or just a result of power differences, a form of stability analysis was conducted using repeated sub-sampling from the imbalanced strata, setting population sizes equal (at n=500) in each sample.

INTERACTION EFFECTS

To further investigate interactions between exposure, gene expression and other covariates (sex, age, smoking status, urbanicity and BMI), linear mixed models were fitted for every transcript–TRAP combination, with the covariates from the adjusted (ADJ) model, and additionally with a multiplicative interaction effect term between the exposure and the covariate. The models with the interaction term were compared with the null model using a likelihood-ratio test. The final specification for the regression was:

$$y\,[transcript] = \overline{\beta_0} * \;\; [matrix\ of\ covariates] + \overline{\beta_1} * \; X\,[TRAP] + \overline{\beta_2} * \; X\,[TRAP] * INTERACTION\ VARIABLE\ [sex, age, smoker, etc]$$

Interaction analysis results are included in Appendix B.

Investigating elemental particulate matter

In addition to modelling levels of the seven principal TRAPs, measurements were modelled for elemental particles of copper, zinc, sulphur, iron, potassium, silicon, vanadium and silicon, at both the PM2.5 and PM10 size.

Previous studies have highlighted the different effects that can be induced by elemental constituent particles of PM2.5 and PM10, and noted that the composition of elemental particles within particulate matter may be more important to health outcomes than the level of particulate matter itself.[62,67,68] This was investigated through univariate and multivariate analysis of elemental particle effects on selected transcripts.

First, univariate analysis was conducted on all elemental PM2.5 and PM10 particles, to gauge effect sizes and strength of associations in a univariate framework. Next, PM10 and PM2.5 elemental particles were analysed separately in multivariate regression models to investigate the joint effect that elemental particles may have on expression levels in transcripts, and as a variable selection method to identify the most important elemental particles in a health context.

ElasticNet regression stability analysis

Meinshausen *et al* describe a method of repeated subsampling combined with variable selection algorithms to identify and rank variables of interest in high-dimensional settings[89]. The method mitigates some of the instability that is innate in penalised regression methods, and LASSO regression in particular, owing to the fact that many solutions are possible and that the calibration procedure is stochastic. This stability selection technique was applied to the data using the following steps:

1. Transcriptomic data were denoised for fixed and random effects by taking the residuals from a series of linear mixed models.
2. Penalised regression models were built ElasticNet[90] in R. Parameters alpha (the mix of L1 and L2 penalisation) and lambda (the penalty weight) were calibrated using 10-fold cross-validation. The format of the models was:

$$y[transcript] = \bar{\beta}\ [a\ vector\ of\ coefficients] * X\ [a\ set\ of\ 8\ elemental\ particles]$$

3. Random samples of 80% of the data were taken and the selected predictor variables recorded. This process was repeated 100 times and the selection proportion, ie the number of times each variable was included in a model, was calculated.

Running this analysis on large numbers of transcripts is impractical, therefore analysis was conducted for the four most significant transcripts identified in the univariate analysis, for both elemental PM10 and, separately, elemental PM2.5.

Bayesian Variable Selection

The GUESS modelling package,[91] and its R implementation, R2GUESS[92] are founded on Evolutionary Stochastic Search (ESS++),[93] which combines Monte Carlo Markov Chain (MCMC) and genetic algorithms to perform Bayesian Variable Selection (BVS) in models with multivariate Y and large X.

R2GUESS is used in this context specifically as it permits multivariate Y (dependent) variables. This allows the investigation of the combined effect of multiple TRAP exposures (independent variables) on a group of transcripts. Models with more Y variables than X variables cannot be analysed in R2GUESS. Therefore, as in ElasticNet analysis, the four most significant transcripts identified in univariate regression were used as outcome variables for both elemental PM10 and PM2.5.

As with ElasticNet, a denoised transcriptomic data set was used to remove the effect of our measured confounders. The format of the model was:

$$y[4\ transcripts] = \bar{\beta}\ [a\ vector\ of\ 8\ coefficients] * X\ [a\ set\ of\ 8\ elemental\ particles]$$

For the final model runs, 55,000 sweeps were performed, with the first 5,000 discarded as burn-in. The R2GUESS output returns Marginal Posterior Probability of Inclusion (MPPI) values for each predictor variable, which was used as a measure of variable importance.

Air pollution and smoking: Similar molecular signatures?

Previous studies have highlighted commonalities between the biological responses to air pollution and biological responses to smoking.[60]

To test the hypothesis that the molecular signature of TRAP exposure may share features with the molecular signature of cigarette smoke exposure, the transcripts selected in univariate screening were compared to a list of genes whose expression has been identified as part of the molecular signature of smoking in previous research by Beineke *et al.*[94] A hypergeometric test was conducted to quantify whether there was a shared transcriptomic profile in the results of this study and the results of studies of smoking.

Gene enrichment and functional analysis

Functional analysis of the genes selected in univariate analysis was conducted using the DAVID[95] bioinformatics tool. Gene-annotation functional analysis using multiple pathway maps including KEGG and GO.

Further analysis of the biological roles of the selected genes was carried out by analysing which gene ontology terms were over-represented in the selected group. The Biological Networks Gene Ontology tool (BiNGO)[96] was used to map the dominant functional themes of the gene set selected at $p<0.1$ after BH correction. The significance threshold was lowered from 0.05 to 0.1 to increase the number of genes included in the both the DAVID and BINGO analysis and thereby increase the power of each.

Unsupervised machine learning approaches

Unsupervised machine learning approaches were also used to investigate underlying patterns and structure in the data set that may help characterise the biological response to air pollution exposure.

Variables characterising the exposome were grouped into five different groups: geographical variables, traffic-related variables, TRAPs, biological variables and demographic variables.

Table 5 summarises the variables in each dimension. Table 1 shows the variables in the TRAP and biological dimensions respectively; the data dictionary in Appendix A includes a full list of variables from the geographical, traffic and demographic dimensions.

Table 5 Groups of variables used for clustering

Dimension	Number of variables	Overview of variables (See data dictionary in Appendix XXX for full details)
Geographical variables	116	Surface are assigned to various types of land use (natural, industrial, residential, urban etc), within radii of various sizes around the home address. Proportion of vegetation and various types of green space within radii of various sizes around the home address. Household and population densities within radii of various sizes around the home address. Traffic intensity and type nearby and density of roads within radii.
Traffic-related variables	37	Traffic intensity and type nearby and density of roads within radii.
TRAPs	22	Air pollutant particles (seven principal particles and elemental PM10 and PM2.5)
Biological variables	10	Height, weight, BMI, white blood cell counts
Demographic variables	20	Average house value in area; proportion of immigrants in nearby population; measures of income in nearby households.

K-means clustering[97] was then applied, using as clustering factors the variables in each of the five different exposome dimensions separately. This generated five different clusterings of the study population. The R package NbClust[98] was used to analyse the optimum number of clusters using 30 different metrics. Different metrics often produce different optimum cluster numbers. Therefore, the final number of clusters was chosen by 'majority rule' – the number that was proposed by the most metrics.

The stability of the clusters was evaluated through bootstrapping using a method described by Hennig.[99] The Jaccard coefficient, which measures similarity between sets, is used to compute the similarity between clusters and clusters generated from repeated bootstrapped samples of the data. The R function 'clusterboot'[100] from the 'fpc' package[101] was used to automate this process, with the number of resamples set to 100.

The exposure profile and other characteristics of the clusters were investigated. To assess the importance of the constituent variables within clusters, random forest regression was used to predict cluster membership using the cluster variables. The 'importance' of the constituent variables to each cluster can then be calculated, using the 'varImp' function from the 'Caret' R package,[102] on the basis of the 'Mean Decrease in Node Purity (Gini)' figure for each variable.

Impurity is the probability of an incorrect classification of a new data point within a variable, if the data point is randomly classified according to the distribution of classes in the existing data set. The formula for impurity is:

$$Impurity = \sum_{i=1}^{C} p(i) \times (1 - p(i))$$

Where C is the number of classes and p(i) is the probability of any class 'i' being chosen.

The 'mean decrease in node purity' is therefore the mean of the decrease in node impurity that a variable introduces over all the nodes it is used in, weighted by the proportion of samples which reach that node in the random forest.

This method or analysing variable importance was also applied 'between' clusters – for example using geographical variables to predict TRAP cluster membership – as a way of understanding relationships between clusters and, in particular, which variables might be of special interest in future LUR-based analyses.

Univariate regressions, stratified by cluster, were also carried out for 374 selected transcripts against the seven principal TRAPs.

RESULTS

Descriptive statistics

After removing observations for individuals for whom data were incomplete or for whom accurate location information could not be determined, a total of 2,438 participants remained in the study for subsequent analysis. **Table 6** breaks down the observations by variable and by gender.

Table 6: Phenotypic and biological data from study participants

Variable	Overall	Male	Female	p-value	Missing observations
n	2438	825	1613	<0.001	
Smoking status (%)				0.001	6
Never smoked	1337 (54.8)	415 (50.3)	922 (57.2)		
Current smoker	515 (21.1)	206 (25.0)	309 (19.2)		
Former smoker	586 (24.0)	204 (24.7)	382 (23.7)		
BMI (%)				0.014	13
BMI 20–30	1987 (81.5)	690 (83.6)	1297 (80.4)		
BMI <20	250 (10.3)	64 (7.8)	186 (11.5)		
BMI >30	201 (8.2)	71 (8.6)	130 (8.1)		
Twin status (%)				<0.001	363
Monozygotic twin (male)	322 (15.5)	322 (48.3)	0 (0.0)		
Dizygotic twin (m-m pair)	172 (8.3)	172 (25.8)	0 (0.0)		
Monozygotic twin (female)	782 (37.7)	0 (0.0)	782 (55.5)		
Dizygotic twin (f-f pair)	418 (20.1)	0 (0.0)	418 (29.7)		
Dizygotic twin (m-f pair)	209 (10.1)	99 (14.8)	110 (7.8)		
Dizygotic twin (f-m pair)	172 (8.3)	74 (11.1)	98 (7.0)		
Age (mean (SD))	36.85 (13.04)	36.61 (14.17)	36.97 (12.42)	0.523	0
Height in cm (mean (SD))	174.07 (9.03)	182.59 (6.98)	169.71 (6.49)	<0.001	1
Weight in kg (mean (SD))	73.50 (14.43)	82.20 (13.37)	69.06 (12.84)	<0.001	7
BMI (mean (SD))	24.22 (4.03)	24.65 (3.73)	24.00 (4.16)	<0.001	13
Blood test data (mean (SD))					
White blood cell count	6.53 (1.76)	6.45 (1.56)	6.56 (1.85)	0.139	34
Percentage neutrofielen (granulocyten)	52.67 (8.73)	52.48 (8.21)	52.77 (8.98)	0.448	36
Percentage lymfocyten (mononucleairen)	35.36 (8.08)	34.62 (7.54)	35.74 (8.33)	0.001	36
Percentage monocyten (mononucleairen)	8.31 (2.16)	9.10 (2.14)	7.91 (2.06)	<0.001	36
Percentage eosinofielen (granulocyte)	3.16 (2.33)	3.37 (2.03)	3.04 (2.47)	0.001	36
Percentage basofielen (granulocyten)	0.50 (1.04)	0.42 (1.17)	0.54 (0.97)	0.007	36
Aantal neutrofielen in 10*9/ L (granulocyten)	4.60 (0.64)	4.99 (0.56)	4.40 (0.58)	<0.001	36

Exposures

Table 7 and **Figure 2** show the distributions of the seven principal TRAP particle measurements. **Figure 3** shows the joint distributions and correlations between the TRAP particles. The same plots for elementary particles are available in Appendix A. There is a high degree of correlation (Pearson's $r > 0.7$) between all of the principal TRAPs, with the exception of PM2.5, which only correlates strongly, as might be expected, with PM2.5_abs (Pearson's $r = 0.791$), the absorption coefficient of PM2.5.

The lower correlation between PM2.5 and the other TRAPs is attributable to the different sources of the particles. For example: the proportion of NO_2 produced by traffic – especially diesel cars – is greater than the proportion of PM2.5 produced by traffic.[103,104] PM2.5 levels also tend to vary less within-region, while other exposures are more prone to local variation.[105]

Table 7 Average modelled TRAP measurement

TRAP	Measurement mean (SD)	Units
no2	23.68 (6.51)	ug/m3
nox	34.56 (11.12)	ug/m3
pm25	16.48 (0.75)	ug/m3
pm25_abs	1.27 (0.24)	10-5/m
pm10	24.99 (1.37)	ug/m3
pmcoarse	8.33 (0.81)	ug/m3
UFP	9702.04 (2041.66)	particles/cm3

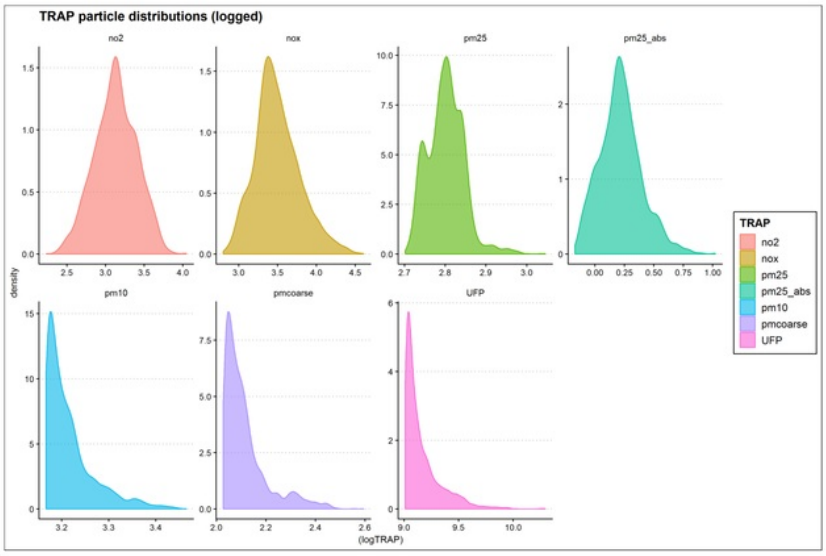

Figure 2 Density plots showing distribution of (logged) TRAPs

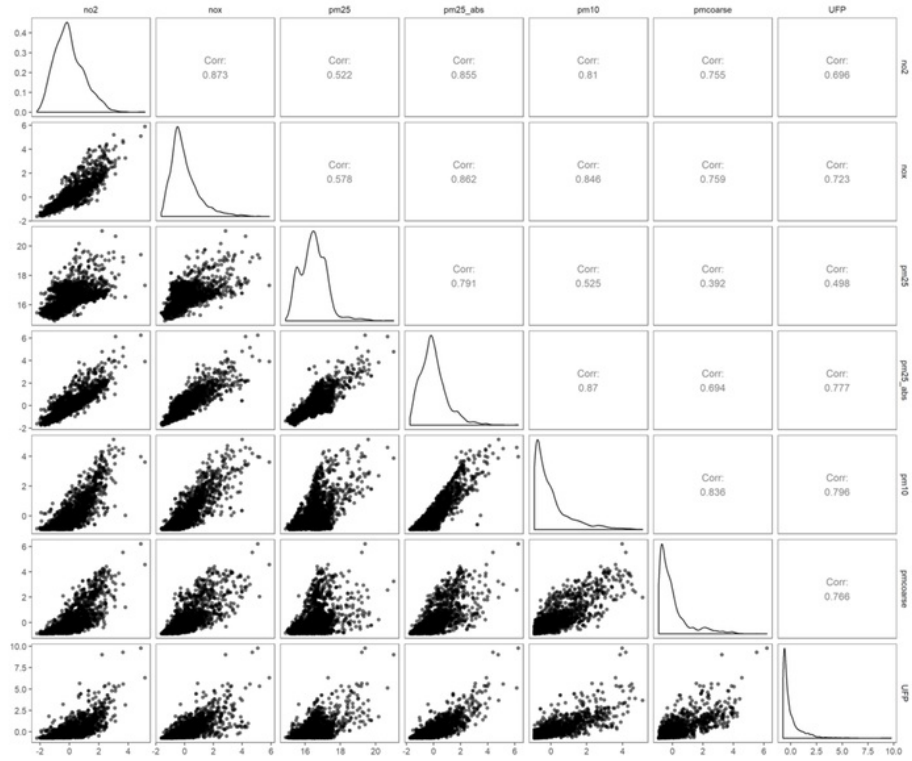

Figure 3 Joint distributions and correlation scores for seven principal TRAPs (unlogged)

Univariate results

UNIVARIATE ANALYSIS OF ALL PRINCIPAL PARTICLES

Table 8 shows the number of significant associations discovered for each of the seven principal TRAPs (NO_2, NO_x, PM2.5, PM2.5_abs, PM10, PMcoarse and UFP) after multiple testing correction at varying levels of stringency.

In the adjusted model, significant results were only found in PM2.5 – a total of 374, after applying the Benjamini-Hochberg correction procedure for controlling the false discovery rate at $\alpha = 0.05$. Of these, 8 were also significant at a level of $\alpha = 0.01$.

Table 8 Number of significant associations for univariate models

BH = Benjamini-Hochberg multiple testing correction

	Crude (unadjusted) model							Adjusted model						
	Log NO2	Log NOX	Log PM25	Log PM25abs	Log PM10	Log PMcoarse	Log UFP	Log NO2	Log NOX	Log PM25	Log PM25abs	Log PM10	Log PMcoarse	Log UFP
Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pval <0.05	3736	2959	7943	3940	3704	3430	2731	2810	2082	5940	3011	2107	2385	2001
BH20	40	2	6020	182	286	52	17	0	0	2560	1	0	0	1
BH5	0	0	2122	6	8	5	0	0	0	374	0	0	0	0
BH1	0	0	607	0	0	0	0	0	0	8	0	0	0	0

Covariates (random effects): plate, well, family ID, month of sampling. (Used in CRUDE and ADJ models)
Covariates (fixed effects): sex, age, smoking status, BMI, white blood cell counts, days between sampling and extraction, days between extraction and amplification, sampling time. (Used in F12 model only)

Logging the TRAPs resulted in a larger number of significant associations. The logged model was chosen to be the default model as testing revealed that the R^2 figure was consistently marginally higher in the log models than the unlogged, indicating a better fit (the distribution of R^2 differentials between the logged and unlogged models is shown in **Figure 33** in Appendix B).

Figure 4 shows denoised expression levels for the four most significant transcripts, plotted against modelled PM2.5 exposure levels. Pearson's r correlation statistics between exposure and expression are around 0.1.

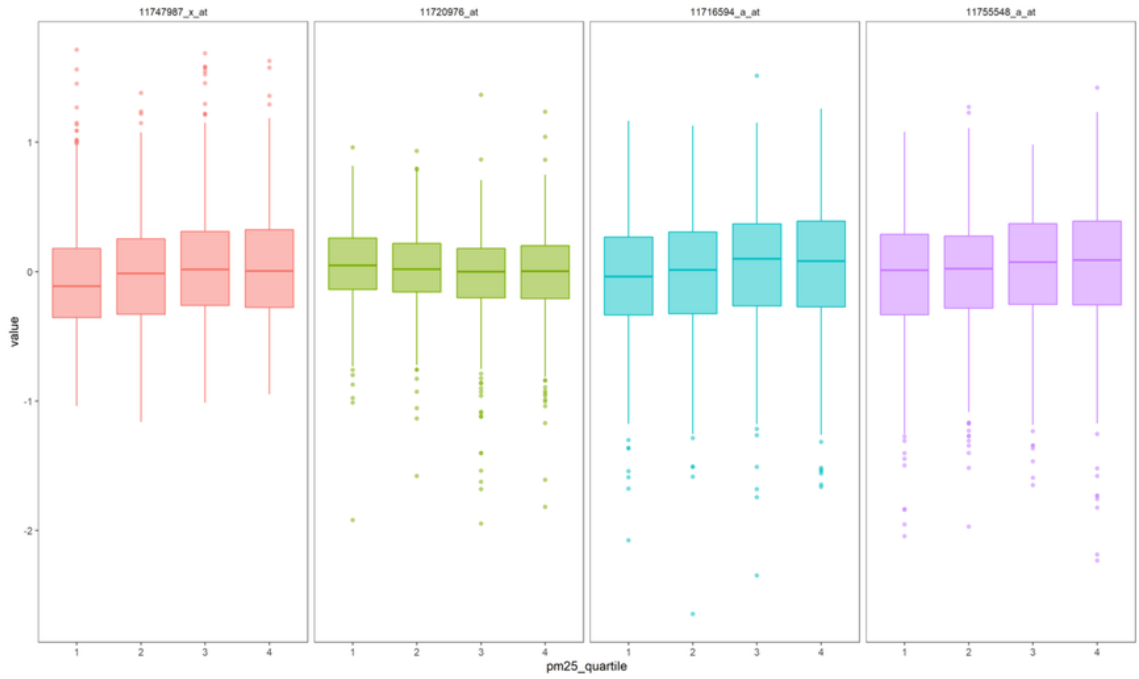

Figure 4 Boxplots showing denoised expression levels in the four most significantly affected transcripts, plotted against modelled PM2.5 exposure quartiles. While the associations are significant, the absolute change in expression levels is quite small.

Figure 5 is a series of volcano plots showing the distributions of betas and p-values for each exposure, for the adjusted (ADJ) models. The five most significant genes are labelled in every plot. The large majority of significant genes – 285 in total – were upregulated, while 89 were downregulated. **Figure 6**, a clustered heatmap, shows the levels of correlation between the 374 transcripts identified as significantly associated with PM2.5, and the hierarchical clustering clearly shows the groups of upregulated and downregulated transcripts, as well as revealing high levels of correlation among some clusters of transcripts within the two larger clusters.

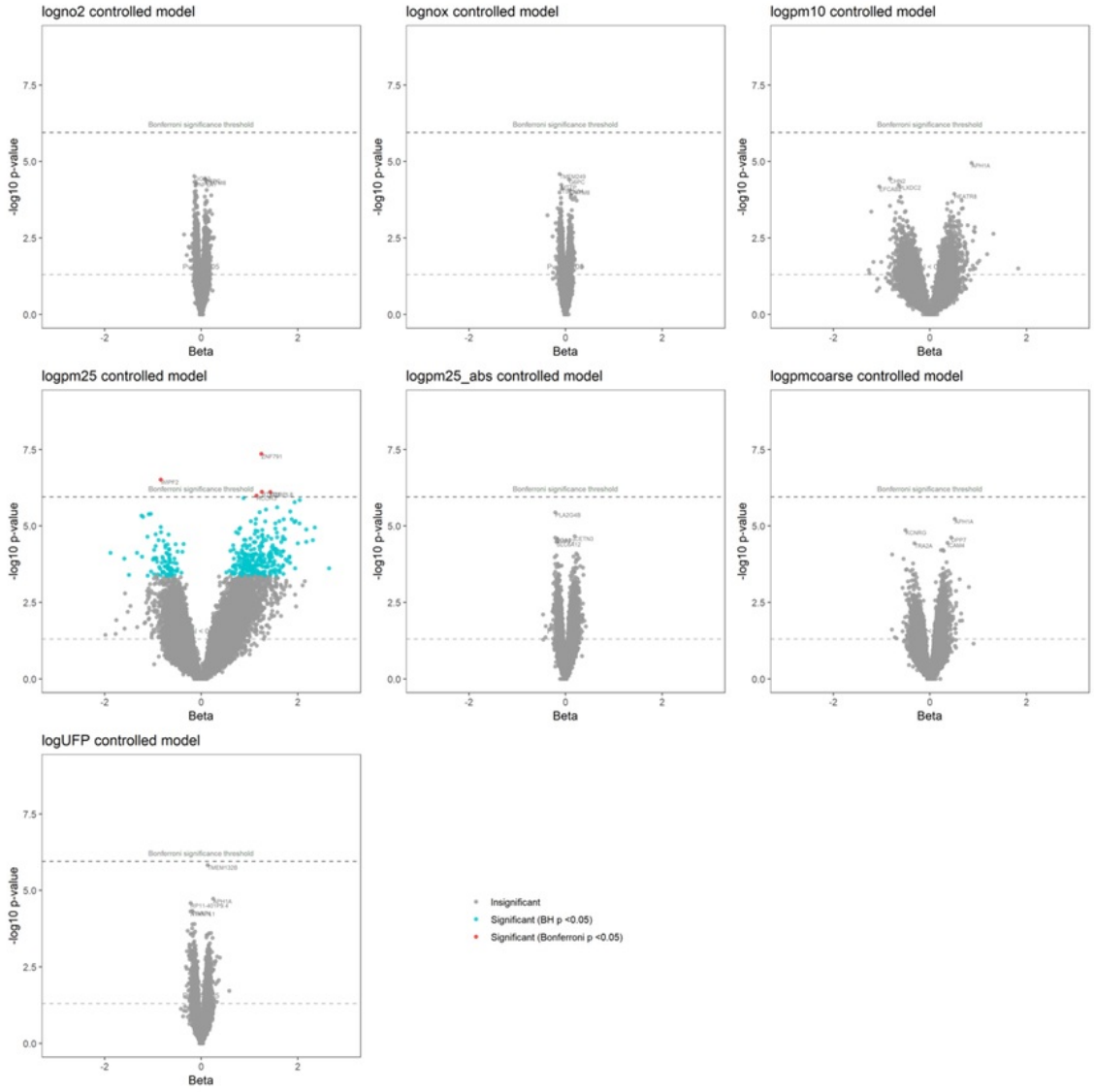

Figure 5 Volcano plots showing beta coefficients and p-values for univariate regression of all main TRAP particles and all transcripts. The four most significant associations are labelled in each plot with the name of the gene associated with the transcript. Significant associations, after BH multiple testing adjustment, are observed between PM2.5 and 374 transcripts, but not between any other exposures.

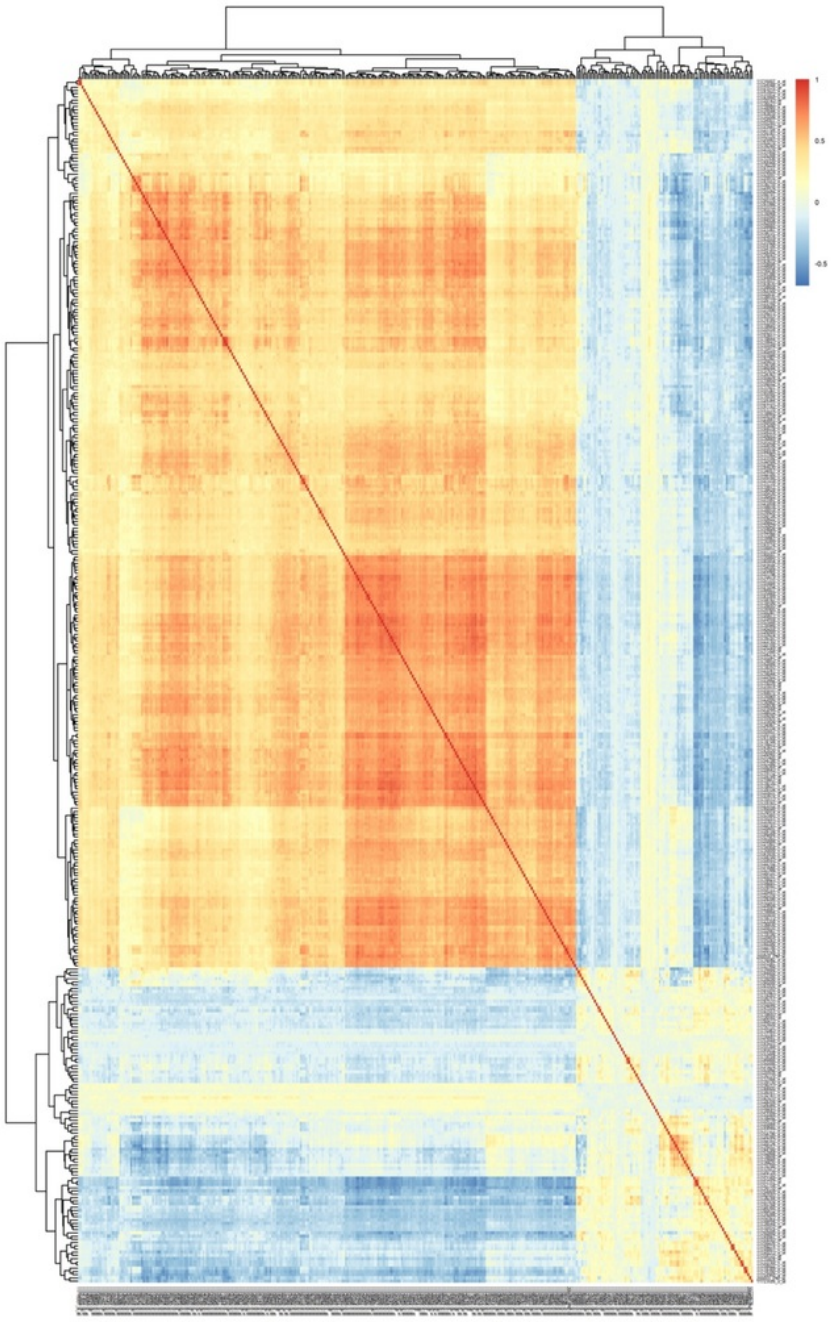

Figure 6 A clustered heatmap showing correlation between 374 transcripts significantly associated with PM2.5 in univariate analysis. Hierarchical clustering shows two distinct groups of transcripts: the 285 upregulated and the 89 downregulated in univariate analysis, as well as clusters of highly correlated transcripts within the upregulated cluster.

Sensitivity analysis

Table 9 shows the results of the capped (Winsorised) sensitivity analysis for the 374 transcripts identified as having significant TRAP associations in the controlled model.

Table 9 Sensitivity analysis: capped regression results

	Capped		Not capped		Adjusting for other TRAPs *	Adjusting for other TRAPs and environ. variables**
Exposure	#transcripts p<0.05 CRUDE	#transcripts p<0.05 ADJ	#transcripts p<0.05 CRUDE	#transcripts p<0.05 ADJ	#transcripts p<0.05 ADJ	#transcripts p<0.05 ADJ
NO2	122	120	123	116	29	12
%	32.62	32.09	32.89	31.02	7.75	
NOX	93	112	93	106	5	4
%	24.87	29.95	24.87	28.34	1.34	
PM25	**373**	**374**	**373**	**374**	**257**	**253**
%	**99.73**	**100**	**99.73**	**100**	**68.72**	
PM25abs	230	295	254	326	8	9
%	61.5	78.88	67.91	87.17	2.14	
PM10	44	49	45	43	25	25
%	11.76	13.1	12.03	11.5	6.68	
PMcoarse	26	31	30	25	11	10
%	6.95	8.29	8.02	6.68	2.94	
UFP	33	54	29	62	9	9
%	8.82	14.44	7.75	16.58	2.41	

*NO2, NOX, PM25, PM25abs, PM10, PMcoarse, and UFP in a single model.
** NO2, NOX, PM25, PM25abs, PM10, PMcoarse, UFP, and top10NL_300m, NDVI_300m, P_LKOOPKRH in a single model

The results are largely robust to capping the extreme values in the data set: the number of transcripts identified as significant does not change substantially between the capped and uncapped results. In the multiple regressions, significant results are substantially reduced for all TRAPs except PM2.5. Including environmental variables in the multiple regression also substantially reduces the significance of the associations in all TRAPs except PM2.5.

Figure 7 and **Figure 8** show the relative p-values and beta coefficients for uncapped univariate regression versus capped (at the 5th and 95th percentiles) regression, for the 374 significant transcripts, for each TRAP exposure. The plots show a high degree of consistency of results in uncapped vs capped data – 97.1% of the beta coefficients had the same sign in the capped results as the uncapped – indicating that the findings are not driven by outliers.

There is more variation between the capped and uncapped p-values within PM2.5, but this is better explained by the fact that the 374 significant transcripts were selected on the basis of their association with PM2.5 and the p-values are by definition at the extremum of the distribution within PM2.5, and therefore more prone to variation when the data is disturbed (the trend for variation between capped and uncapped p-values increasing as the p-value decreases is clearly visible in the plots for the other TRAPs too).

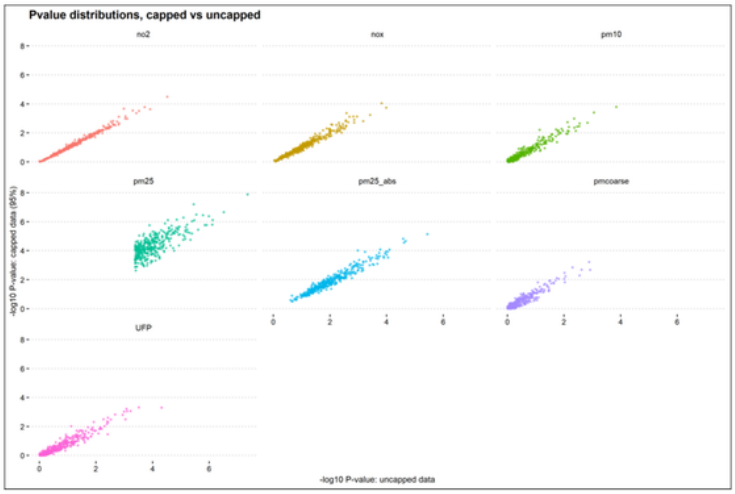

Figure 7 P-value distributions: capped vs uncapped univariate models. Distributions are highly consistent in capped and uncapped models.

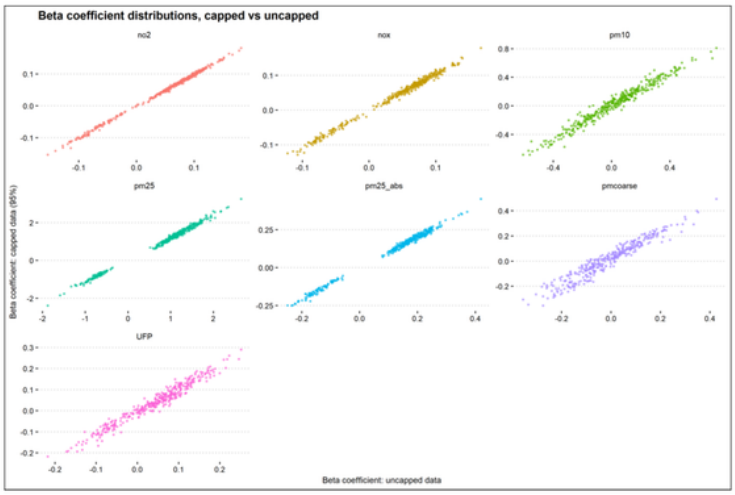

Figure 8 Beta coefficient distributions: capped vs uncapped models. Distributions are highly consistent in capped and uncapped models.

Stratified analyses

STRATIFYING BY AGE AND GENDER

Figure 9 shows the relative distributions of p-values and coefficients in the stratified results within PM2.5, stratifying by age quartiles and gender. Significant results are shown in **Table 10**. Fewer significant results are observed than in the full-cohort analysis, which is at least partially explained by sample size and concomitant loss of power.

Table 10 Significant transcripts identified in stratified analysis

*Bracketed numbers indicate the number of significant transcripts that were also significant in the unstratified analysis

		Uncontrolled model (RAN)							Controlled model (FULL)						
Stratum	n	no2	nox	pm25	pm25_ abs	pm10	Pm coarse	UFP	no2	nox	pm25	pm25_ abs	pm10	pm coarse	UFP
17-27	609	1	1	1	1	1	1	1	-	-	-	-	-	-	-
28-33	610	-	-	-	-	-	-	-	-	-	-	-	-	-	-
34-44	609	-	-	25 (23)*	-	-	-	-	-	-	1 (1)*	-	-	-	-
45-79	610	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Female	1613	2	-	1708 (1268)*	10 (2)*	-	-	-	1	-	40 (31)*	-	-	1	-
Male	825	-	-	-	-	-	2	2	-	-	-	-	-	2	1
Full cohort	**2438**	**-**	**-**	**2122**	**6**	**8**	**5**	**-**	**-**	**-**	**374**	**-**	**-**	**-**	**-**

Far fewer significant results are observed in women than in men, and in women the effect size and direction of exposure on gene expression levels is much more consistent with the full-cohort results. The results of stability testing to overcome the difference in statistical power between the male and female cohorts are shown below.

Similarly, no significant results are observed in the 29–33 age stratum, and there is almost no agreement between the effect size and direction in that stratum and in the full-cohort results. The results in the other age cohorts are much more consistent with the full-cohort results. The age strata were split by quartiles and are therefore balanced in number of observations, so the effect is not a result of varying statistical power. The implications of the anomalous results in the 28–33 age cohort are addressed in the Discussion.

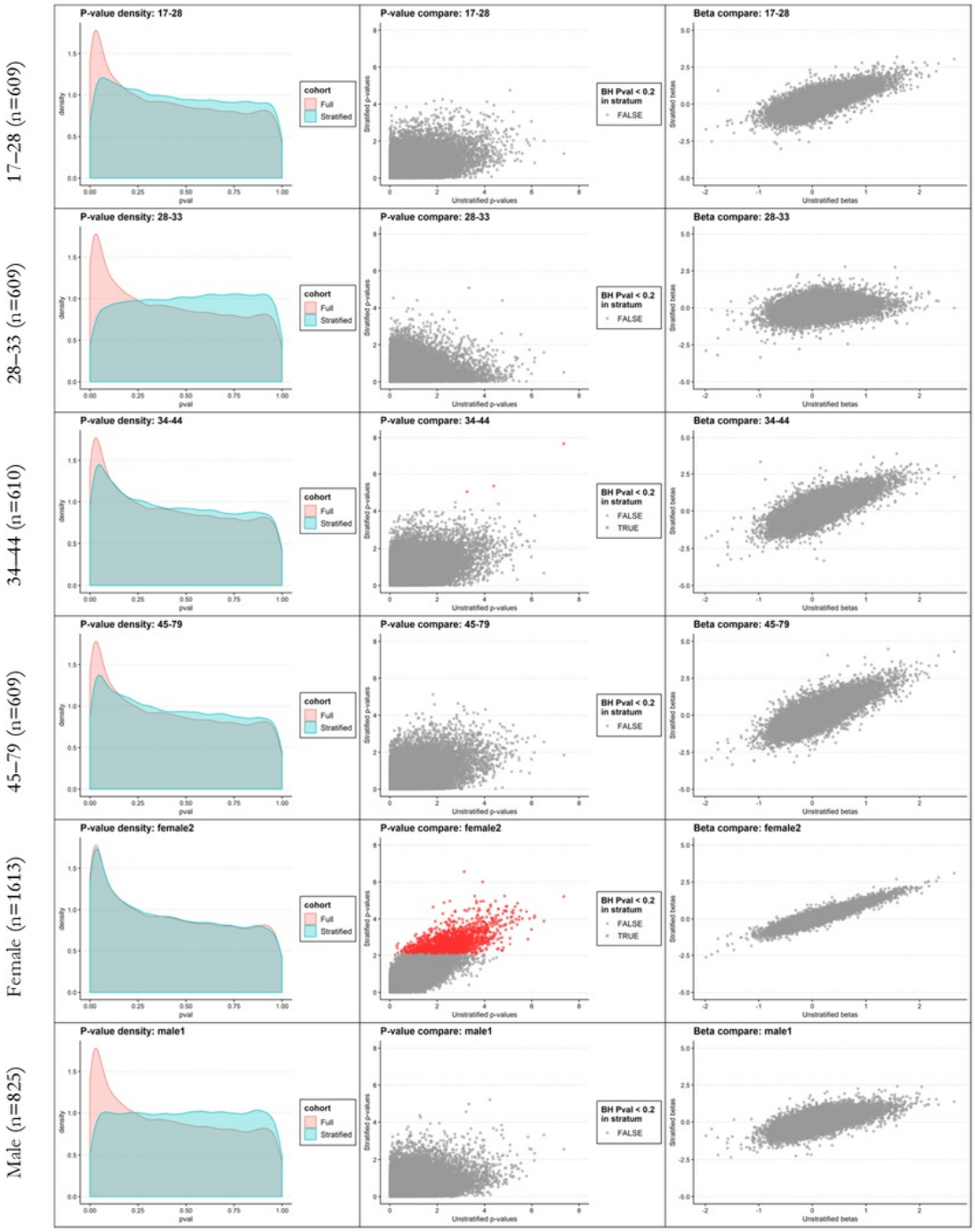

Figure 9 Plots comparing the distribution of p-values and beta coefficients in stratified results versus unstratified, investigating the associations between PM2.5 and transcripts. Significant associations (BH p-value <0.2 – lowered to reveal more associations in a lower-powered, smaller n analysis) are highlighted in red.
Fewer significant results are observed in men, and in the 28–33 stratum.

Investigating gender-specific associations

The stratified univariate analysis revealed much stronger TRAP–transcript associations in the female cohort than the male. **Figure 10** shows the distributions of p-values in 60 repeated subsamples of male and female populations with sample size n=500, over-plotted and compared with the distribution of p-values in the full male (n=825) and female (n=1,613) populations.

Subsampling reveals more significant results in the female stratum than in the male. This is further demonstrated in **Figure 11**.

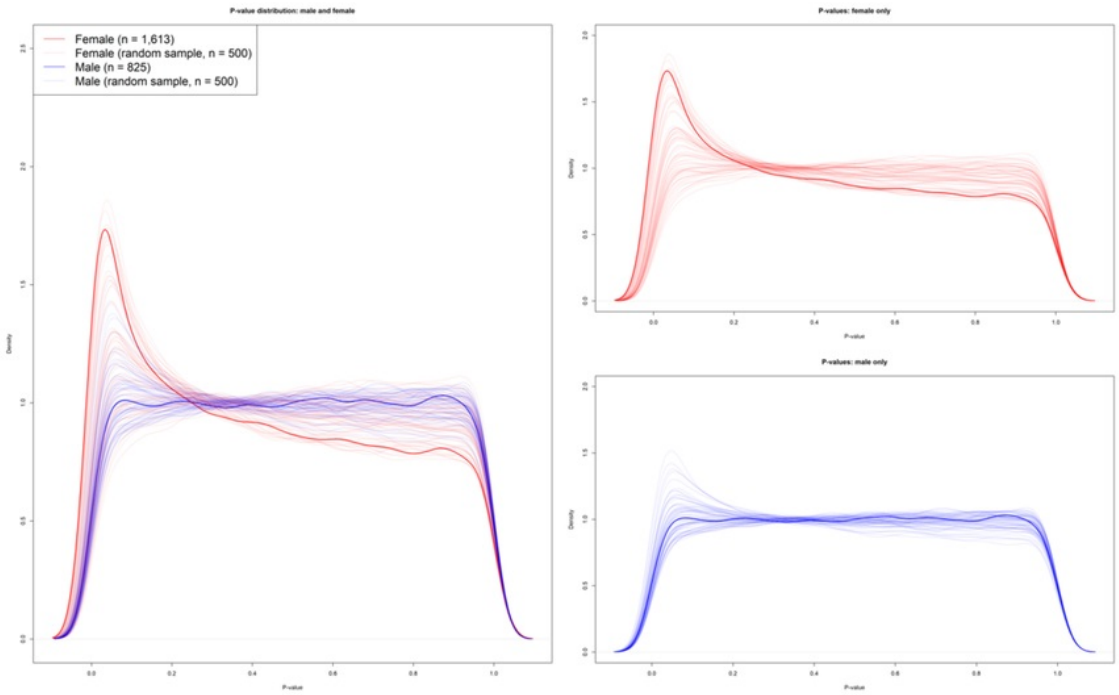

Figure 10 Distributions of p-values in male and female cohorts (red = female; blue = male).
Thick lines show p-value distribution in the full male/female cohorts. Each faint line shows a p-value distribution in a random sub-sample of 500 males or females. Female subsamples are shown to have p-value distributions more weighted towards lower values (higher significance), more often.

Figure 11 shows the distribution of significant results from the sub-sampling analysis at a $p<0.001$ threshold (uncorrected). More significant associations are found in female subsamples.

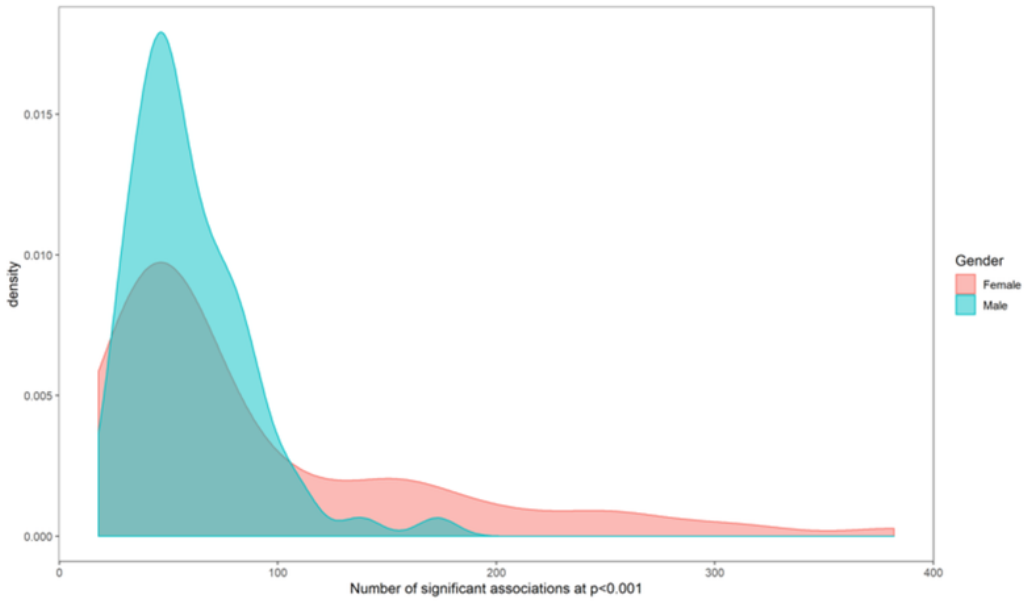

Figure 11 Distribution of number of significant associations ($p<0.001$) found in random $n=500$ sub-samples of male and female observations. More significant associations are observed in female subsamples.

STRATIFYING BY URBANICITY

Using the Netherlands' Central Bureau of Statistics' urbanicity calculation, each address in the dataset was categorised into one of 5 categories of 'urbanicity', on the basis of population density within 1km^2 (see **Table 11**).[76]

Table 11 Address density ranges for each urbanicity category

Urbanicity Category	Address density	No. of observations in category
1	>= 2,500 addresses / km^2	560
2	1,500–2,500 addresses / km^2	529
3	1,000–1,500 addresses / km^2	436
4	500–1000 addresses / km^2	495
5	<500 addresses / km^2	418

Figure 12 shows the results of univariate analysis of the adjusted (ADJ) model, stratified for each level of urbanicity. Larger effect sizes and more significant results are observed in areas of lower urbanicity, most notably in the lowest-urbanicity areas (see bottom row of figure). **Figure 13** shows volcano plots for urbanicity-stratified analysis of the 374 transcripts selected as significant in full-cohort univariate analysis. Again, more significant results are observed in areas of lower urbanicity, most notably in the lowest-urbanicity stratum.

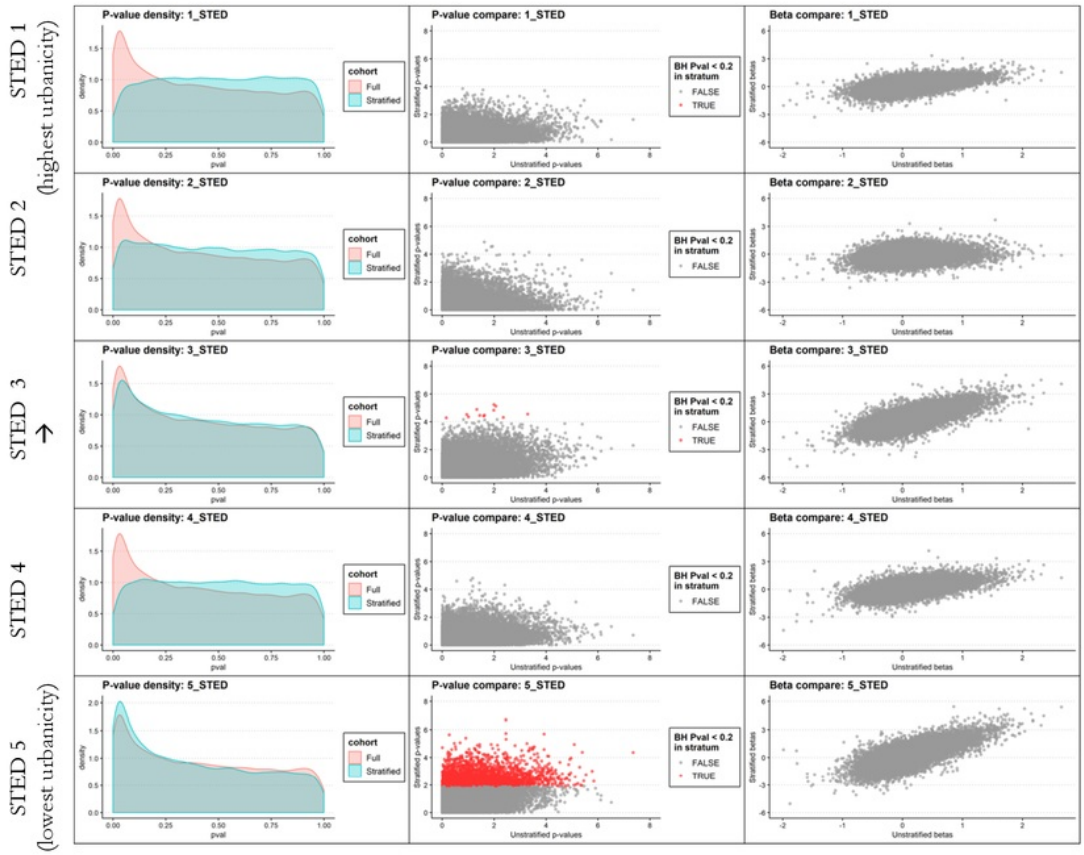

Figure 12 Plots comparing the distribution of p-values and beta coefficients in urbanicity-stratified results versus unstratified, investigating the associations between PM2.5 and transcripts. Significant associations (BH p-value <0.2) are highlighted in red.

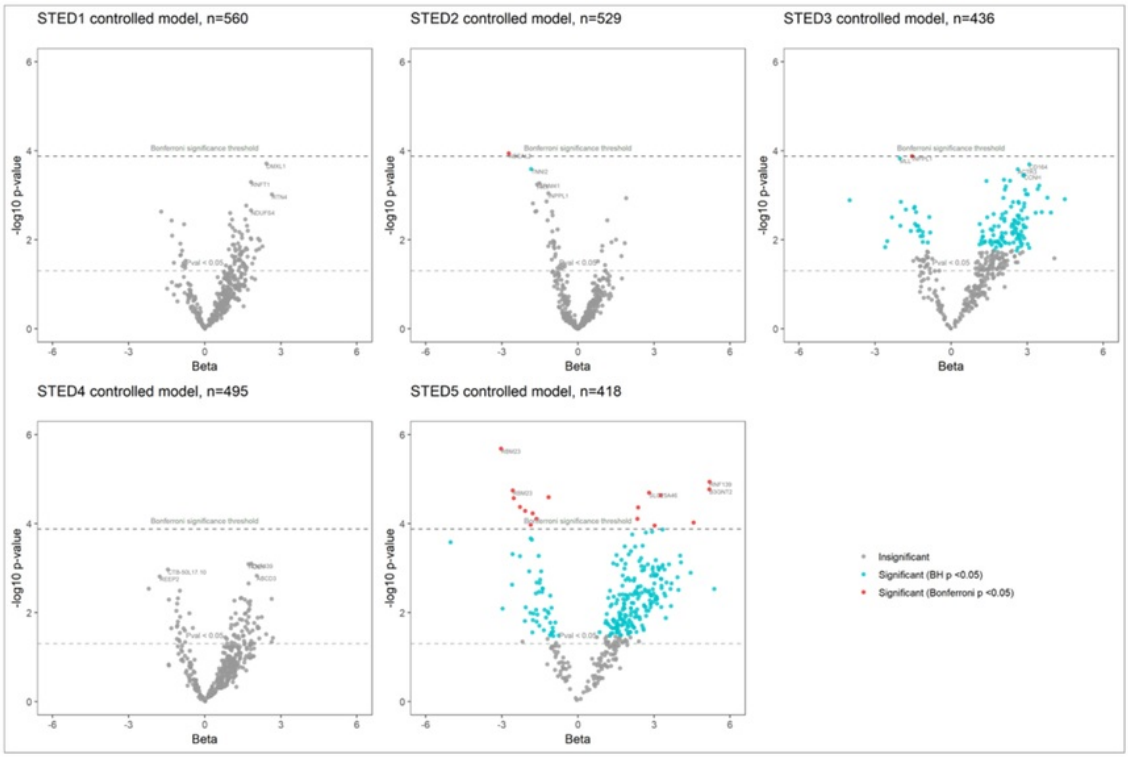

Figure 13 Volcano plots showing p-values and beta coefficients for significant transcripts in a controlled univariate model, investigating the associations between PM2.5 and transcripts, stratified by urbanicity. Only 374 selected transcripts are shown.
Stronger associations are seen in areas of lower urbanicity, especially STED 3 and STED 5.

The effect of elemental particulate matter

UNIVARIATE ANALYSIS OF ELEMENTAL PM2.5[1]

Figure 14 shows volcano plots for univariate analysis of elemental particulate matter PM2.5, for the 374 transcripts sound to be associated with total PM2.5. The most significant results can be seen in copper (Cu), Sulphur (S), Silicon (Si) and Iron (Fe).

Figure 15 and **Figure 16** show the relative distributions of p-values and beta coefficients for PM2.5 elemental particles compared with total PM2.5. A high degree of correlation (ranging from 0.90 to 0.98) is observed in the beta coefficients, indicating consistent transcriptomic effects across specific elemental particles and with total PM2.5. The effect sizes for elemental particles are much smaller than total PM2.5, as might be expected, but the p-values are of a similar order of magnitude.

[1] As no significant associations were observed between PM10 exposure and gene expression, results for analysis of elemental PM10 particles are not included here.

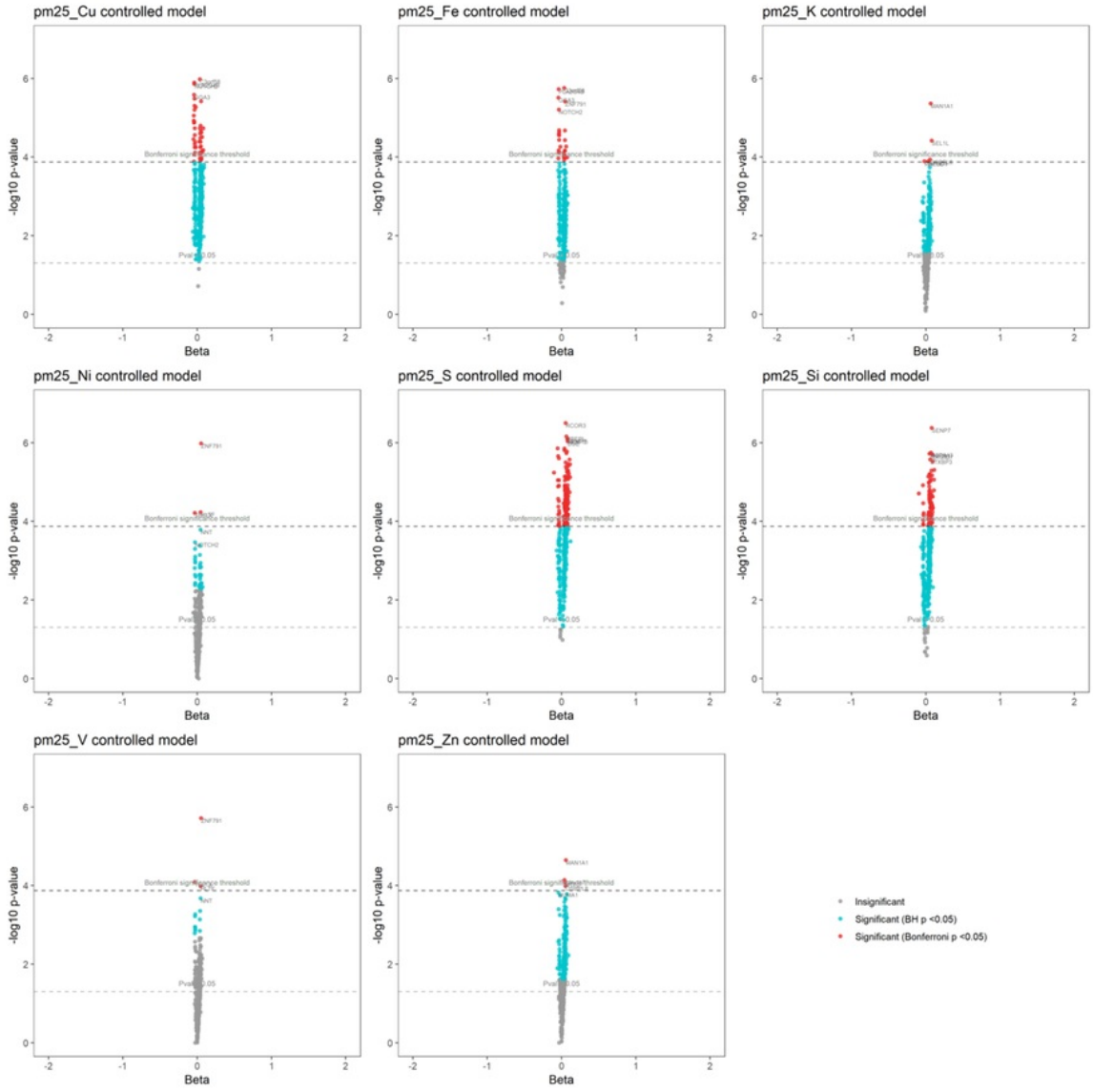

Figure 14 Volcano plots showing results from univariate analysis of 374 significant transcripts and eight elemental pm2.5 particles. X-axis is shown on the same scale as the volcano plots in the univariate analysis of total PM2.5 in **Figure 5**, to demonstrate the difference in effect size. The most significant results can be seen in copper (Cu), Sulphur (S), Silicon (Si) and Iron (Fe).

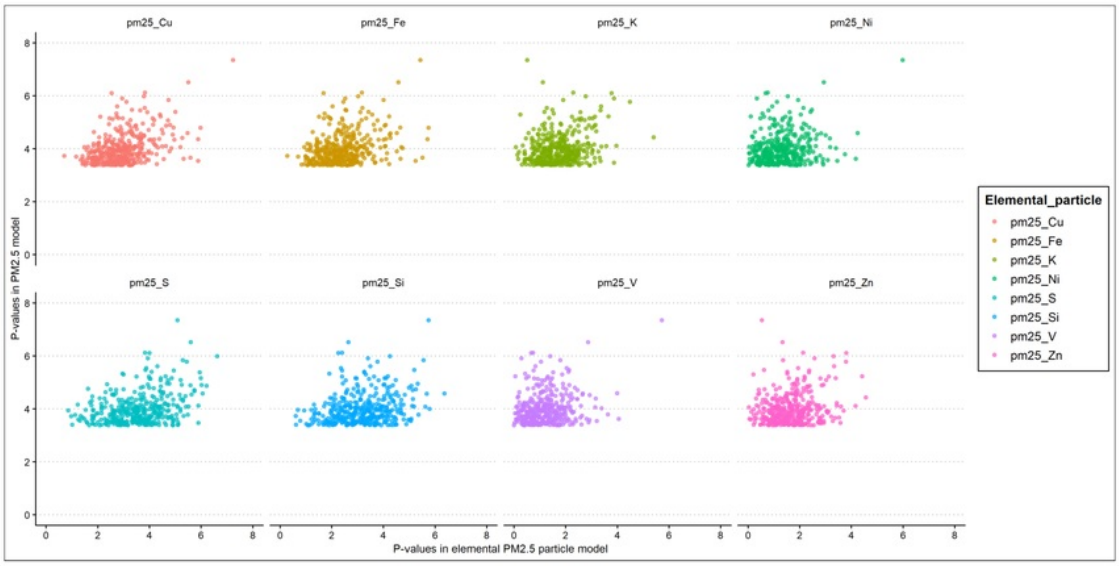

Figure 15 P-values for univariate regression models for elemental PM2.5 particles plotted against P-values for univariate regression of PM2.5
Only the 374 transcripts selected by univariate analysis are shown.

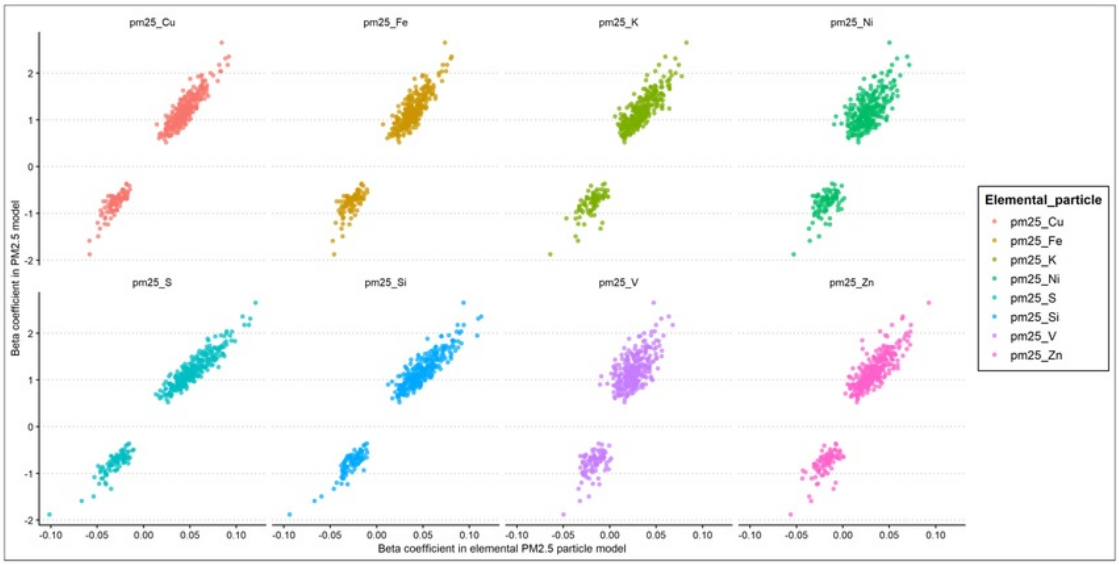

Figure 16 Beta coefficients for univariate regression models for elemental PM2.5 particles plotted against beta coefficients for univariate regression of PM2.5. Effects are highly consistent in direction, but smaller in size in elemental particles than in total PM2.5.
Only the 374 transcripts selected by univariate analysis are shown.

PENALISED REGRESSION: ELASTICNET

Figure 17 shows results from stability analysis of the four most significant transcripts identified in univariate analysis, regressed onto PM2.5 elemental particles in repeated ElasticNet models.

Copper, sulphur and zinc are stably selected in most of the models, with copper meeting the stability threshold of 0.6 in all four of the transcripts, and sulphur and zinc in three.

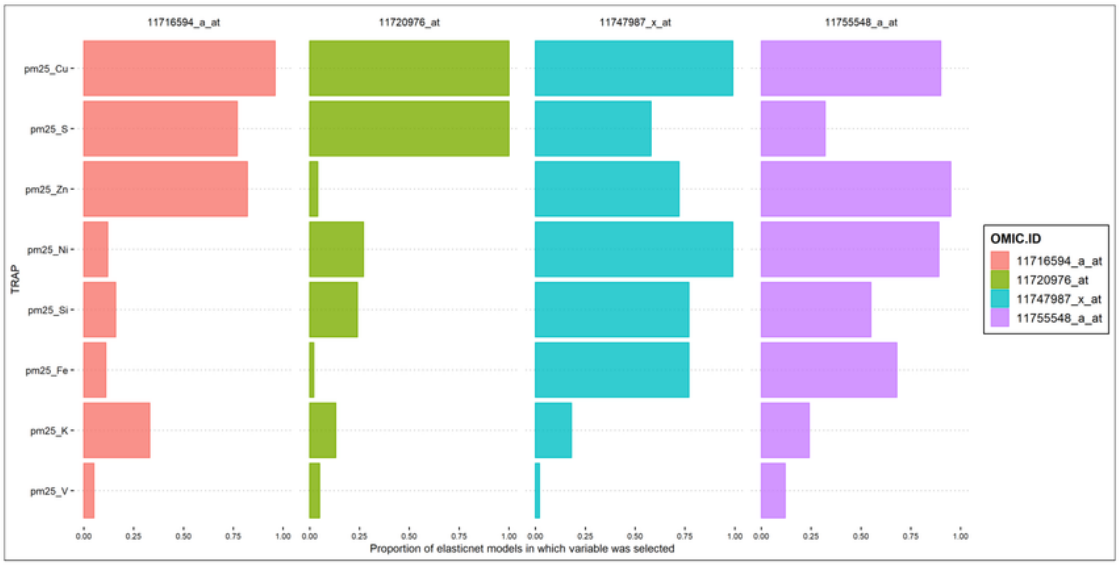

Figure 17 Results of stability analysis, regressing 4 significant transcripts separately onto PM25 elemental particles using repeated ElasticNet regression.

R2GUESS

R2GUESS results are shown below. Output plots from the R2GUESS model runs are included in Appendix B.

Figure 19 shows the Marginal Posterior Probability of Inclusion (MPPI) for elemental PM2.5 particles when used as predictor variables for a multivariate Y that included the four most significant of the selected transcripts. Only elemental copper reaches the threshold MPPI for significance of 0.1. MPPIs for other variables were very low, ranging from 0.0032 for sulphur to 0.000000078 for potassium.

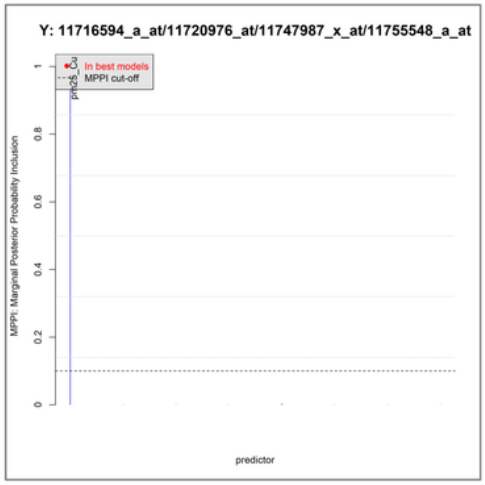

Figure 19 MPPI output from R2GUESS when eight elemental PM2.5 particles are used as predictors for four significant TRAPs. Elemental copper PM2.5 is the only selected predictor variable.

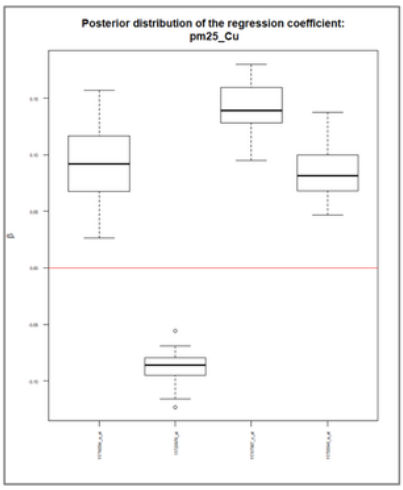

Figure 18 Posterior distribution of the regression coefficient for PM2.5_Cu (elemental copper particulate matter).
Plots show upregulation of transcripts coded by ZNF791, BTBD1 and OSBPL8, and downregulation of a transcript coded by WIPF2.

Figure 18 shows the posterior distribution for the regression coefficient of PM2.5_Cu (elementary copper particulate matter 2.5) on the four most significant of the selected transcripts. The effect direction is the same as detected by univariate analysis of elemental copper.

Air pollution and smoking: similar molecular signatures?

Of the 374 transcripts selected in univariate analysis, 80 were encoded by genes identified as associated with smoking by Beineke *et al.*[94] A hypergeometric test indicated 2.25-fold enrichment relative to expectations (there were 2.25 times as many smoking-related genes selected in the TRAP analysis as would be expected under a null hypothesis of there being no similarities between smoking and air pollution exposure), with a p-value of 4.33×10^{-12}. This is strong evidence for an overlap in gene expression impact between TRAP exposure and smoking, reinforcing previous research on this topic.[106,107]

A full univariate analysis to identify transcripts significantly associated with smoking status in this cohort was also conducted. Results are included in Appendix D.

Gene enrichment and functional analysis

Table 12 shows the results of functional analysis of genes associated with all transcripts identified as significantly associated with PM2.5 at a BH-adjusted p-value of P<0.1. The significance threshold was lowered to increase the number of genes included in the analysis and thereby increase the power of the test.

Several functional pathways are identified as being perturbed to a high level of significance, most notably phosphoprotein and alternative splicing.

Table 12 Functional annotation for all transcripts identified at BH pval < 0.1
Analysis carried out using DAVID Functional Annotation Tool[95]
Only 10 most significant functional pathways shown: full functional chart included in supplementary material

Category	Term	Count	%	PValue	Bonferroni	Benjamini	FDR
UP_KEYWORDS	Phosphoprotein	530	65.19	6.17E-50	2.55E-47	2.55E-47	8.65E-47
UP_KEYWORDS	Alternative splicing	580	71.34	5.84E-33	2.42E-30	1.21E-30	8.19E-30
UP_SEQ_FEATURE	splice variant	433	53.26	7.04E-19	1.60E-15	1.60E-15	1.23E-15
UP_KEYWORDS	Acetylation	228	28.04	3.50E-17	1.45E-14	4.82E-15	4.90E-14
UP_KEYWORDS	Nucleus	311	38.25	1.88E-16	9.19E-14	2.30E-14	3.11E-13
UP_KEYWORDS	Ubl conjugation	131	16.11	1.06E-13	4.38E-11	8.75E-12	1.48E-10
UP_KEYWORDS	Isopeptide bond	96	11.81	1.75E-12	7.23E-10	1.21E-10	2.45E-09
GOTERM_MF_DIRECT	GO:0005515~protein binding	485	59.66	2.11E-13	1.91E-10	1.91E-10	3.29E-10
UP_KEYWORDS	Zinc	157	19.31	1.91E-11	7.90E-09	1.13E-09	2.68E-08
GOTERM_CC_DIRECT	GO:0005829~cytosol	217	26.69	1.68E-11	9.62E-09	4.81E-09	2.47E-08

Figure 20 shows gene ontology over-representation analysis, visualised as a network, using the Bingo[96] gene ontology tool in Cytoscape. The most significantly over-represented biological functions in the selected gene set are cellular processes, cellular component biogenesis and pseudouridine synthesis.

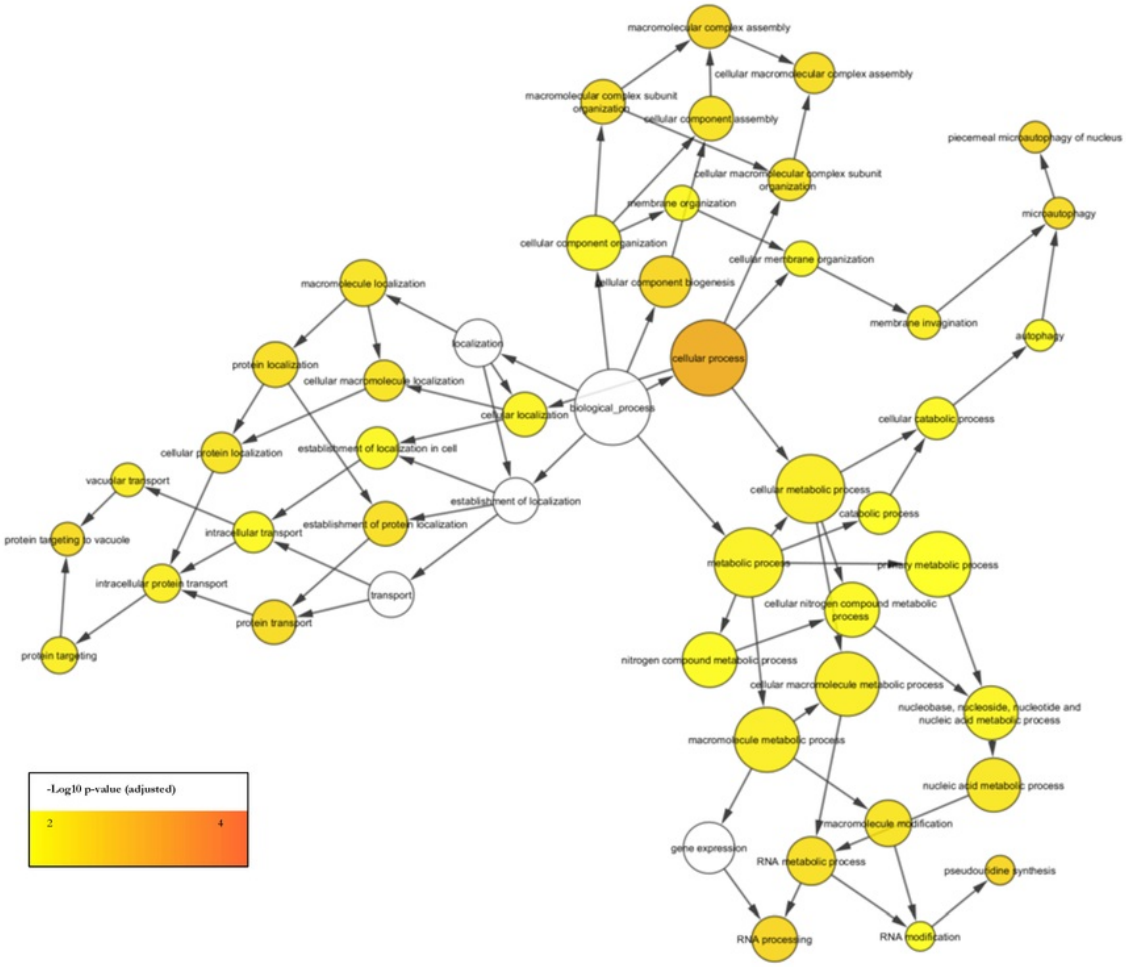

Figure 20 Gene ontology biological process analysis using BINGO[96]
The number of genes involved in the process is indicated by the size of the node.
The significance of the over-representation of the selected gene set in the relevant biological process is represented through the colour of the node (darker = lower adjusted p-value)

Gene functions and details for the eight most significant genes identified in univariate analysis are described in **Table 13**, along with relevant references from the research literature.

Table 13 Gene names, functions, and relevant literature for most significant genes

Gene symbol	Gene name	Biological process	Relevant literature
ZNF791	Zinc Finger Protein 791	May be involved in transcription regulation by RNA polymerase II[108]	Prognostic marker in lung cancer and urothelial cancer (high expression is favourable for both. Gene expression was elevated in response to exposure in this study)[109].
WIPF2	WAS/WASL Interacting Protein Family Member 2	Plays a role in forming cell surface protrusions after activation of PDGFB receptors[110]	Role in mediating internalisation of pathogen in airway epithelial cells[111] Amplified in HER2+ gastric cancer[112] Regulator of breast cancer cell migration and invasion[113]
BTBD1	BTB Domain Containing 1	Binding topoisomerase I; involved in protein-protein interactions[114]	Prognostic marker in renal cancer (high expression is favourable. Gene expression was elevated in response to exposure in this study [109]
OSBPL8	Oxysterol Binding Protein Like 8	Lipid transporter[115]	Associated with lipoprotein cholesterol levels (risk factor for cardiovascular disease)[116]
PGGT1B	Protein geranylgeranyltransferase type I subunit beta	Post-translational protein modifier by adding prenyl group to target proteins.	Elevated levels found in motor neurons of early vs late ALS patients[117]
RCOR3	REST corepressor 3	Protein-encoding; chromatin binding[118]	Prognostic marker in glioma and lung cancer (high expression is favourable for both. Gene expression was elevated in response to exposure in this study)[109]
NEDD1	NEDD1 Gamma-Tubulin Ring Complex Targeting Factor	Protein coding gene, plays a role in mitosis[119]	Prognostic marker in liver cancer and lung cancer (high expression is unfavourable for both. Gene expression was elevated in response to exposure in this study) [109]
SEL1L	SEL1L ERAD E3 ligase adaptor subunit	Involved in protein degradation[120]	Plays a role in malignant gliomas[121]

Unsupervised machine learning results

Table 14 shows the clusters that were found in the data set, with the number of observations allocated to each cluster, the stability of the cluster, and the demographic and exposomic (age, urbanicity, exposure) characteristics of the cluster. Transcriptomic clustering is not included as it was not found to be informative. All clusters are stable (stability score of >0.75) or highly stable (stability score of >0.85), according to the thresholds defined by Hennig.[99] Visualisations of the clusters through PCA plotting are available in Appendix C.

Table 14 Clusters identified in the data set, with exposomic characteristics and stability scores. The final column shows the averaged, normalised exposure level among all cluster members, across the seven principal TRAPs. A score of +1 indicates that the average cluster member was 1 standard deviation above the average exposure level, averaged over all seven TRAPs.

Cluster		Number of observations in cluster	Stability (0–1)	Mean age	Urbanicity (percentage in class) 1=highest urbanicity, 5=lowest urbanicity					Mean centred and scaled exposure level across all 7 principal TRAPs
					1	2	3	4	5	
TRAPs (2 clusters)										
	1	1906	0.98093	37.7	11	21	21	25	22	-0.35
	2	532	0.93877	33.9	78	23	7	2	2	+1.27
Geography (3 clusters)										
	1	1060	0.929931	38.7	0	4	17	40	39	-0.57
	2	278	0.800027	32.8	80	13	5	1	1	+1.76
	3	1100	0.880317	36.1	31	41	22	6	0	+0.10
Traffic (2 clusters)										
	1	2180	0.99342	37.2	19	22	19	22	19	-0.20
	2	258	0.94712	33.9	59	22	9	4	5	+1.65
Demographics (2 clusters)										
	1	671	0.974738	33.1	61	28	8	1	1	+0.63
	2	1766	0.990145	38.3	8	19	22	29	23	-0.24
Biology (3 clusters)										
	1	792	0.824251	36.1	20	22	19	21	18	-0.05
	2	665	0.844351	40.1	20	22	19	23	16	-0.07
	3	981	0.844924	34.7	28	21	16	19	17	+0.09

With the exception of the biological dimension, all cluster dimensions are picking up on a similar structure in the data: a smaller subset of younger, urban residents who have much higher levels of exposure, and a larger cluster or clusters of older residents who live in less densely populated

locations and have average or lower levels of exposure. This is visualised in **Figure 21**, a Sankey plot showing how study participants group within clusters across four dimensions (biological and transcriptomic dimensions are excluded as they are not informative; a full plot showing all dimensions is reproduced in Appendix C).

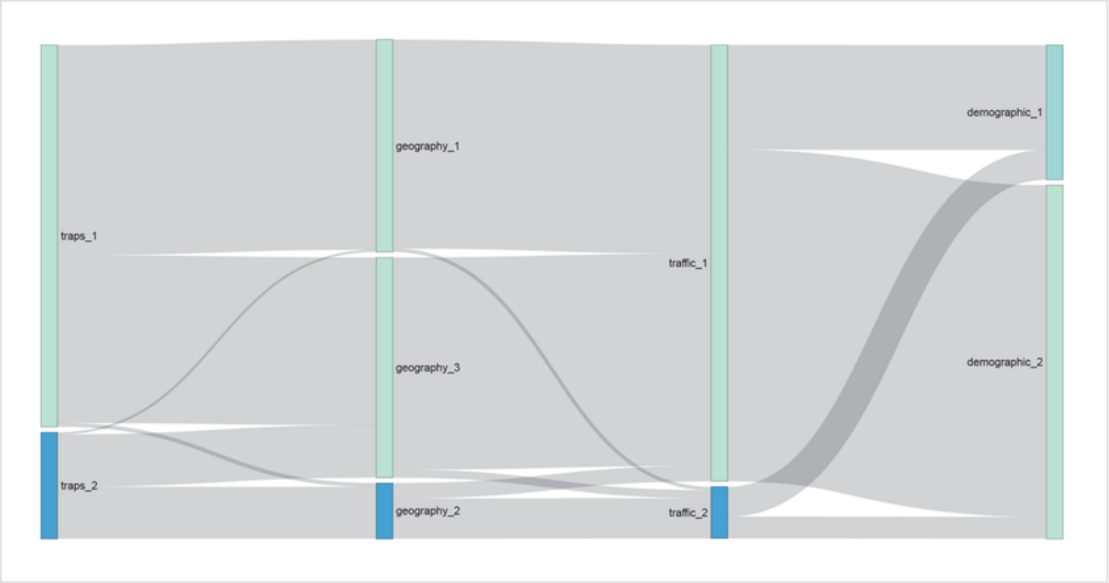

Figure 21 A Sankey diagram showing how study participants are grouped within clusters across four dimensions From left to right, dimensions are: TRAPs, Geography, Traffic, Demographic. The diagram clearly shows how a very similar group of participants in clustered together across three dimensions (the second cluster in TRAPs, geography and traffic).

The diagram shows how a very similar group of participants is clustered together across three dimensions (the second clusters in TRAPs, geography and traffic).

EXPOSURE PROFILES WITHIN CLUSTERS

Figure 22 and **Figure 23** show TRAP exposure levels in each of the three geographical clusters,[2] for the seven principal TRAPs and for elemental PM2.5 particles, respectively.

The distribution of TRAP exposures in the clusters is fairly consistent across TRAPs: cluster 1 is the least highly exposed; cluster 2 the most. There is less variation in exposure levels of PM2.5 between clusters than the other TRAPs.

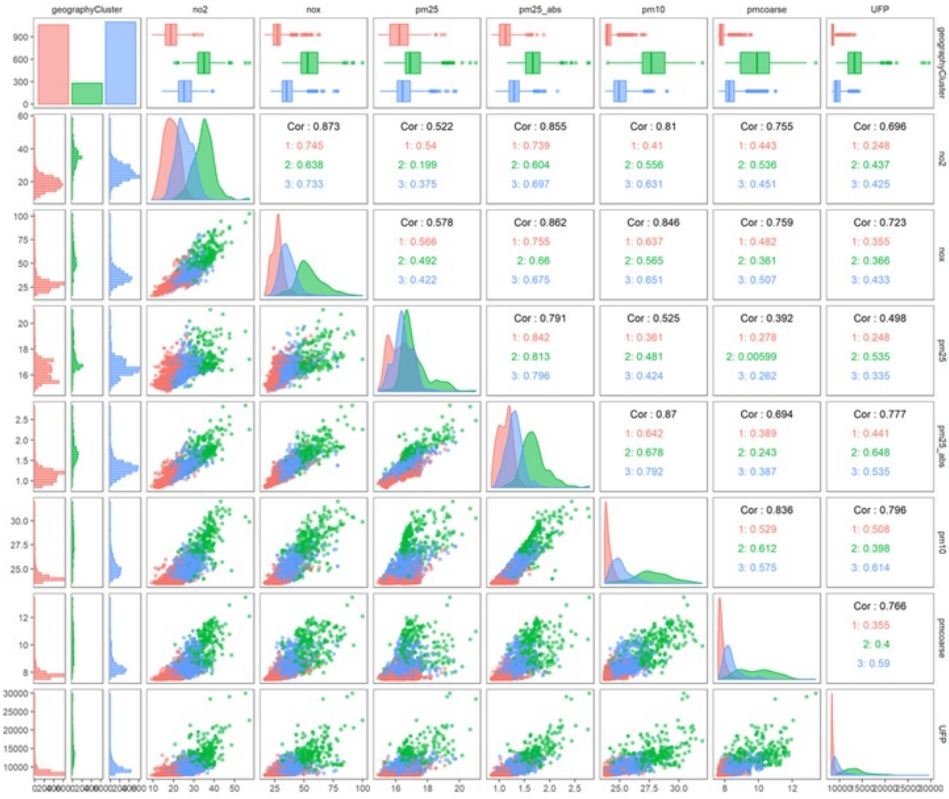

Figure 22 Joint distribution of exposure levels in geographical clusters, for seven principal TRAPs.
For all TRAPs, exposure is highest in cluster 2, and lowest in cluster 1, although there is less variation in PM2.5 levels between clusters than the other TRAPs.
Plots along the top and left margins show the marginal distributions of each exposure, grouped by geographical cluster. Sub-diagonal plots show joint distributions for all exposures, with the colour of each data point indicating which geographical cluster the point belonged to. Numbers represent correlations between exposures, broken down within geographical clusters.

---

[2] Only geographical clusters are shown, for reasons of space. The TRAP exposure profiles within the traffic clusters and the TRAP clusters are very similar to the geographical clusters.

The pattern is similar for elemental particles, with the exception of potassium and zinc, levels of which are lower in cluster 3, and higher in cluster 1. This reflects the fact that zinc and potassium correlate weakly or negatively with other elemental PM2.5 particles.

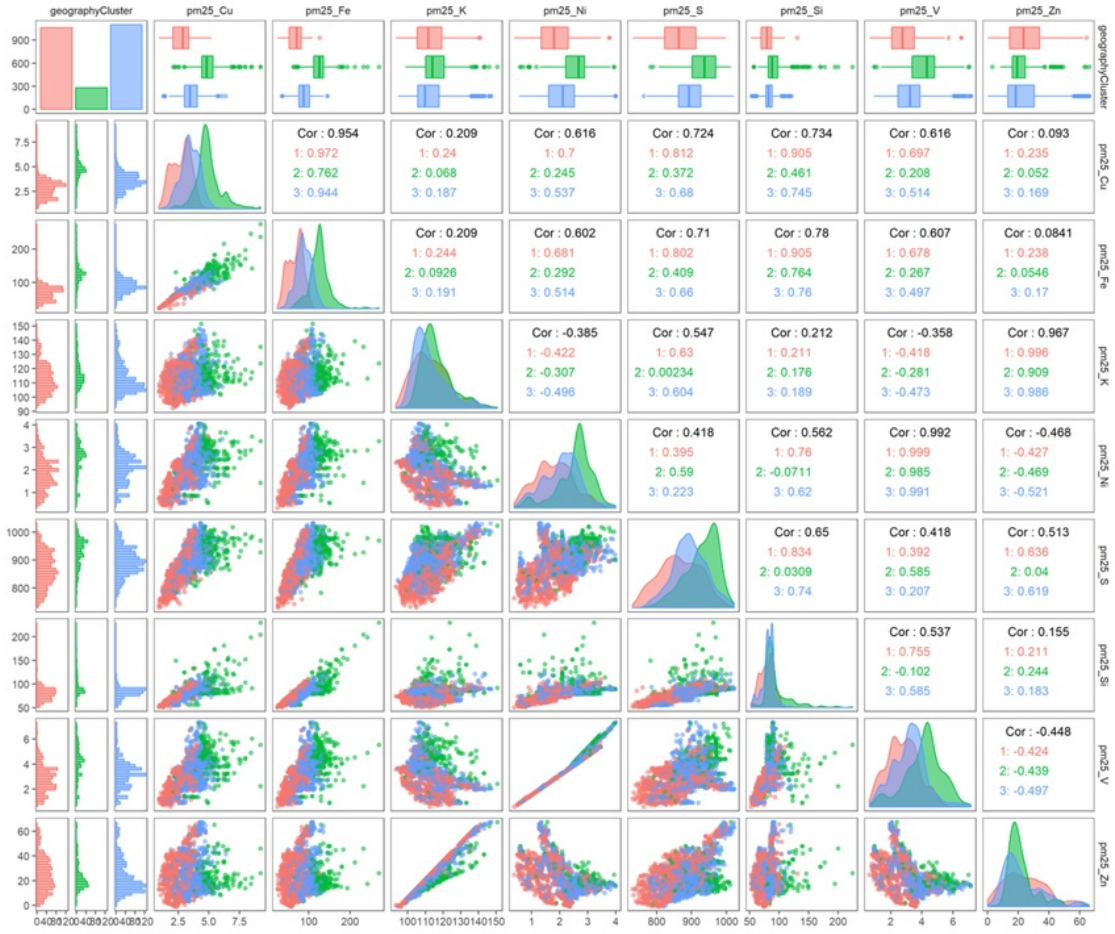

Figure 23 Joint distribution of exposure levels in geographical clusters, for elemental PM2.5 particles. Exposure levels follow a similar pattern to the principal TRAPs – with cluster 2 being the highest exposed and cluster 1 the lowest – with the exception of Zinc and Potassium.
Plots along the top and left margins show the marginal distributions of each elemental particle exposure, grouped by geographical cluster. Sub-diagonal plots show joint distributions for all exposures, with the colour of each data point indicating which geographical cluster the point belonged to. Numbers represent correlations between elemental particle levels, broken down within geographical clusters.

RANDOM FORESTS TO IDENTIFY VARIABLE IMPORTANCE

Figure 24 shows the variable importance of a random forest model using geographical variables to predict which TRAP cluster a person belongs to.

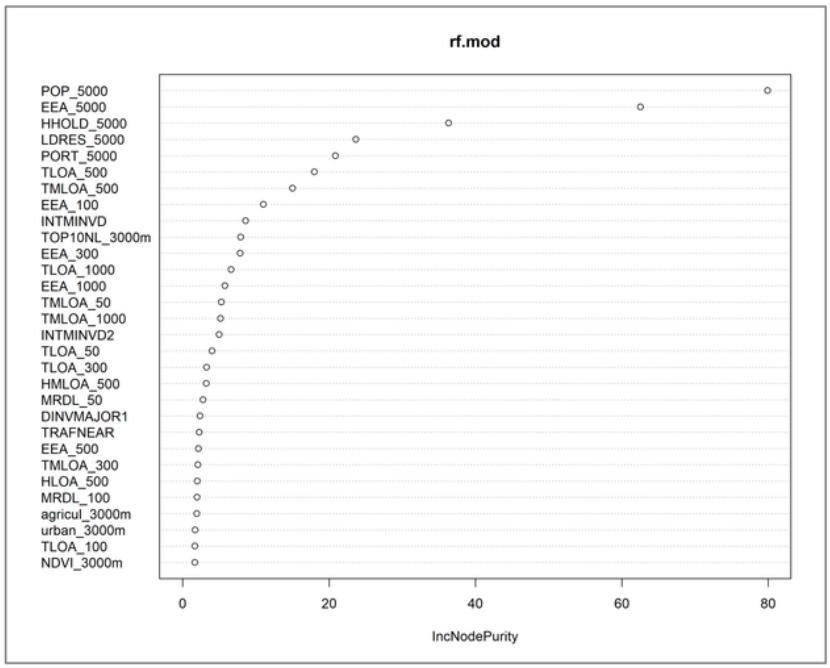

Figure 24 Variable importance in random forest regression models, using geographical variables to predict which TRAP cluster a person will be assigned to. The model notably selects the variables which operate at the broadest geographical scale.

The model strongly favours the lowest-fidelity, largest-scale geographical variables to predict TRAP cluster membership. This reflects that in the regression equations used to model exposure levels, the effect sizes assigned to regional background measurements are much larger than the effects of, for example, local traffic.

The dominance of the 5000m-scale variables reveals the level of geographical fidelity at which the LUR models are functioning most effectively. The implications of this are addressed in the discussion.

UNIVARIATE REGRESSION ANALYSIS, STRATIFIED BY CLUSTER

Table 15 shows the results of univariate analysis of seven principal TRAPs onto the 374 selected transcripts, stratified by cluster. Significantly more results are identified in the lower-urbanicity, lower-exposure, older clusters, across all dimensions.

The most significant associations are found with PM2.5, but running the analysis on the 'rural' clusters – Traps 1, Geography 1 and Traffic 1, reveals associations with some of the other principal TRAPs, associations that are not revealed when conducting univariate analysis of the full cohort. **Figure 25** visualises the same results, showing log10 p-values for univariate regressions.

Notably, while very few significant associations for NO_2 and NO_x are found in a full-cohort analysis, more associations emerge in a few clusters, especially geography cluster 1. The exposomic characteristics of geography cluster 1 are low levels of exposure (the lowest of any individual cluster), older age and lower urbanicity.

Table 15 Results of univariate controlled regression of all 374 interesting transcripts onto seven principal TRAPs, stratified by cluster. Significance threshold $p<0.05$ after BH multiple testing correction.

Cluster	Number in cluster	no2	nox	pm10	pm25	pm25_abs	pmcoarse	UFP
Traps Cluster_1	1906	96	16	0	373	325	3	0
Traps Cluster_2	532	0	0	0	0	0	30	0
Geography Cluster_1	1060	204	64	0	343	324	0	0
Geography Cluster_2	278	0	0	0	0	0	19	0
Geography Cluster_3	1100	0	0	0	10	4	0	0
Traffic Cluster_1	2180	23	12	0	374	310	0	1
Traffic Cluster_2	258	0	0	0	0	0	0	0
Demographic Cluster_1	672	0	0	0	2	0	0	0
Demographic Cluster_2	1766	15	5	9	374	292	12	29
Biology Cluster_1	792	1	0	5	110	12	0	1
Biology Cluster_2	665	0	0	0	0	0	0	0
Biology Cluster_3	981	0	0	0	314	4	0	0
Full cohort	**2438**	**14**	**3**	**0**	**374**	**296**	**0**	**1**

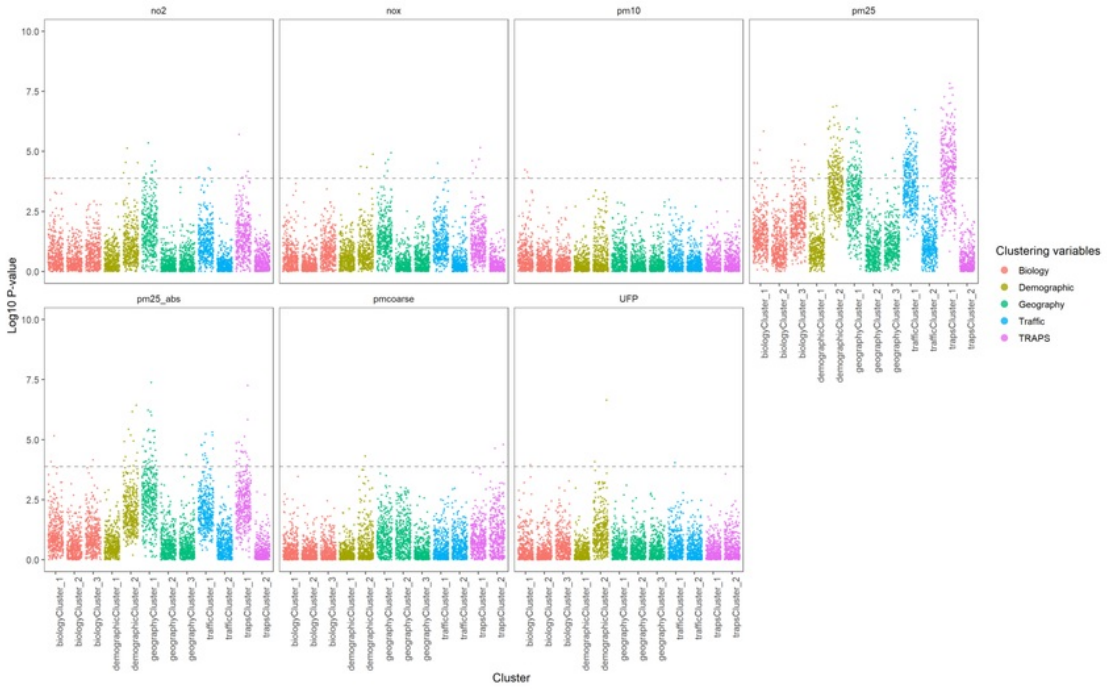

Figure 25 Scatter plots showing log10 p-values for adjusted univariate analysis of 374 transcripts against 7 principal TRAPs, stratified by cluster. Results show strong associations in demographic cluster 2, geographical cluster 1, traffic cluster 1 and TRAP cluster 1.

DISCUSSION

Findings

The results of the analysis in this study suggest a distinct molecular signature of TRAP exposure at a transcriptomic level, driven by PM2.5. The results were robust to truncation of the exposure data and to changes in the specification of the models through adding additional environmental covariates.

Further univariate and multivariate modelling showed that some elemental PM2.5 particles were especially strong contributors to the transcriptomic signature of total particulate matter. Both ElasticNet and R2GUESS regressions of elemental particles strongly selected elemental PM2.5 copper and sulphur as having the most significant contributory effect on gene expression. Elemental copper and sulphur are generally characterised as 'non tailpipe emissions',[67] originating from tyre or break wear on vehicles rather than from exhausts. Previous research has indicated exposure to elemental particulate copper and sulphur in oxidative damage[122,123] and tumour growth, as well as increasing risk from all-cause mortality.[124]

FUNCTIONAL ANALYSIS

Functional analysis of the significant genes revealed a number of significantly affected functional pathways. These pathways have been associated with oncogenic processes and multiple diseases in the literature.

Protein phosphorylation is one of the key mechanisms by which protein function is regulated, and dysregulation of the process has been linked with many diseases, including several types of cancer.[125,126] Phosphoproteins have therefore become widely investigated as biomarkers useful in predicting and detecting various types of cancer,[127–129] as well as predictors for cancer therapy.[125,130]

Alternative splicing is a mechanism first described in 1978 by Gilbert[131] by which multiple proteins may be encoded by single genes, and is essential for creating protein diversity[132]. Disruption of the complex alternative splicing system is linked to multiple diseases[133] and oncogenic processes and is widely discussed in the literature as a factor in tumour development.[134–140]

A **splice variant**, also called a splice-site variant or splice-site mutation, is a related process in alterative splicing which can cause an altered protein coding sequence and introduce variation into the proteome. Splice variants are also studied as cancer biomarkers.[141–143]

REPLICATION OF PREVIOUS RESEARCH

The findings support previous research which has found that PM2.5 exposure has a significant effect on gene expression. However, there is limited overlap in the genes identified as significant in this study and those identified in previous studies. **Table 16** summarises previous literature in which transcriptomic responses to PM2.5 were identified and highlights any overlaps with significant findings in this study.

The only study that identifies a gene expression response that is replicated in the results of this study is Villarreal-Calderon *et al*,[56] who found upregulation of the PRNP gene – a protein encoding gene expressed predominantly in the nervous system and associated with several cognitive disorders and neurodegenerative conditions[144,145] – in residents of urban areas in Mexico who were highly exposed to PM2.5, relative to the control group of residents of less urban areas. Our results also identified upregulation of PRNP, at a significance level of p=0.0003 (significant at FDR 0.05).

Table 16 Summary of previous research identifying transcriptomic responses to PM2.5 exposure

Paper	Study design	Gene expression response to PM2.5 exposure	Findings in this study
Saenen *et* al. In *Utero* Fine Particle Air Pollution and Placental Expression of Genes in the Brain-Derived Neurotrophic Factor Signaling Pathway: An ENVIR*ONAGE* Birth Cohort Study[53]	N=90 birth cohort study, measurements taken from placental tissue after birth. Long-term PM2.5 modelled using spatiotemporal models	BDNF upregulated SYN1 downregulated	No significant effect in BDNF or SYN1
Villarreal-Calderon *et al.* Up-Regulation of mRNA Ventricular *PRNP* Prion Protein Gene Expression in Air Pollution Highly Exposed Young Urbanites: Endoplasmic Reticulum Stress, Glucose Regulated Protein 78, and Nanosized Particles[56]	N=21 children and young adults, differentially exposed in long term. Candidate gene expression in left vs right ventricular tissue analysed.	PRNP upregulated GRP78 upregulated	PRNP upregulated (p=0.00037. Significant at FDR = 0.05)
Wittkop *et al.* Nrf2-related gene expression and exposure to traffic-related air pollution in elderly subjects with cardiovascular disease: An exploratory panel study[50]	N=43, cohort panel, repeated measures of elderly CAD patients. Short term exposure measured daily using sensors. Gene expression in whole blood analysed.	CYP1B1 upregulated	CYP1B1 upregulated (p=0.002. Not significant at FDR = 0.05)
Tripathi *et al.* Variation in doses and duration of particulate matter exposure in bronchial epithelial cells results in upregulation of different genes associated with airway disorders[58]	In vitro response of human bronchial epithelial cells to short-term PM2.5 exposure	IL6, TNF, TSLP, CSF2, PTGS2, IL4R, SPINK5, ADAM33, ORMDL3, DPP10, CYP1A1, IL13 and TGFB1 upregulated	PTGS2 upregulated (p=0.003. Not significant at FDR = 0.05)
Petit *et al.* Alteration of peripheral blood monocyte gene expression in humans following diesel exhaust inhalation[146]	N=14. Peripheral blood transcriptome analysed before and after short term exposure to diesel fumes (or clean filtered air, in control)	F2R, USP10, UBR2, NOS2A and NOS3 upregulated. CDH1, PLAU and CBL downregulated	CBL downregulated (p=0.002. Not significant at FDR = 0.05) UBR2 downregulated (p=0.01. Not significant at FDR = 0.05)
Chu *et al.* Gene expression network analyses in response to air pollution exposures in the trucking industry[147]	N=63. Repeated measures analysis of truck drivers over 3 weeks. Whole blood samples analysed; PM2.5 exposure measured using sensors in truck cabs and offices.	49 genes identified as significantly associated with PM2.5 exposure. Results reported as network / GSEA analysis.	Unclear.

Wittkop *et al*[50] identified upregulation of CYP1B1 in response to PM2.5 exposure in elderly patients with cardiovascular disease. CYP1B1 – a gene associated with oxidative stress pathways[148] – was also upregulated in the findings in this study, although the results were not significant after FDR correction.

The designs of the studies listed in **Table 16** are very different to this one. They generally use small numbers of participants and focus on short-term exposure, and some use gene expression data from different body tissues, rather than peripheral blood. Therefore the absence of significant replication should be seen partly as a function of how comparatively new this type of study design is, rather than a crisis of generalisability.

Interestingly, a review of collective results from studies of air pollution and human health identifies the univariate approach to studying air pollution exposures as a methodological weakness. On the basis of their meta-analysis of 29 studies, in which they note the lack of replication, Staneck *et al* suggest that the chemical composition of particulate matter may be a better predictor of health impact than the mass or size of particulate matter itself.[149] This is both a good potential explanation for some of the lack of replication in studies of air pollution, and a strong case for complementing univariate approaches with multivariate analysis and in-depth investigation of elemental particulate matter.

THE URBAN–RURAL DIVIDE

Both the stratified analysis and the clustering analysis revealed that stronger associations were present between TRAP exposure and gene expression in individuals who lived in less urban areas and were less highly exposed. There are a number of possible explanations for this. In decreasing order of likely significance:

1. *Measurement inaccuracy*

Land-use regression is by definition an inexact science. Modelled exposure measurements are highly sensitive to the design of the model and the available input variables.[150] A review of LUR models found R^2 statistics – the proportion of variation explained by the model – ranging from .54 to .81 in six different models.[151] The LUR used to model TRAP exposure in this study, which was built on training data from multiple locations in Europe, reported R^2 statistics of between .25 and .79 depending on the country to which it was applied (the figure was .61 in the Netherlands). Although within-country accuracy data is not available, it can reasonably be assumed that accuracy varies by a similar amount within-country as between-country.

Further: analysis of accuracy in other published LUR models shows that variance in measured exposure levels for TRAPs increases as exposure levels increase (see **Figure 26**).[152] Therefore,

accuracy of LUR models is at its lowest in areas of very high exposure. This is a compelling explanation for the comparative lack of significant findings among residents of the highest-exposure locations in this study.

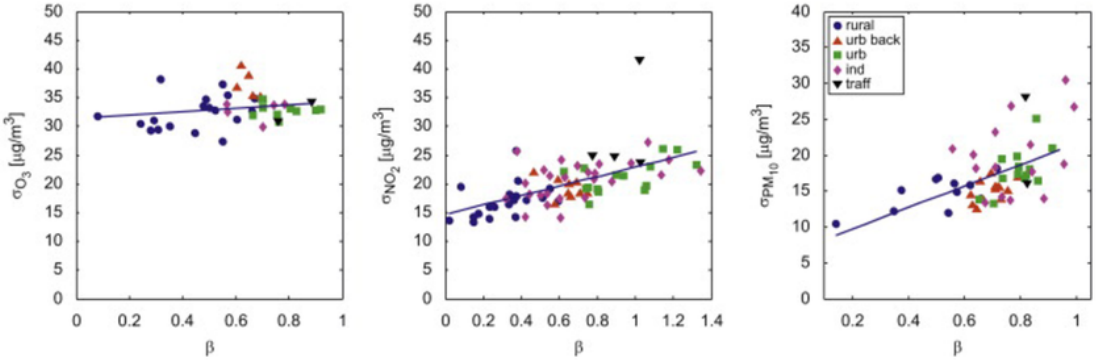

Figure 26 Variance of measurements of O_3 (left) NO_2 (middle) and PM10 (right), plotted against Beta-parameters from land use regression models. Variance increases as TRAP exposure increases, and rural areas are shown to have the lowest variance.
Reproduced, with permission, from: Janssen *et al.* Spatial interpolation of air pollution measurements using CORINE land cover data. Atmospheric Environment; Volume 42, Issue 20, June 2008, Pages 4884-4903.[152]

There is also the question of temporal stability in LUR models. TRAP levels vary significantly from season to season,[153] and while the univariate analysis adjusted for the month of sampling, this will not have captured the full nature of the variation. Similarly, of course, settlements develop over time and therefore exposure levels change substantially across time – a study in Canada found a .87 correlation between measured exposure levels across a seven-year interval[154] – and in this study there was a gap of several years between the blood samples and the modelling of exposure.

2. *Unmeasured variables*

The individual-specific data collected in this study by questionnaire were basic: smoking status, gender, age, etc. A wealth of environmental data was available, but no lifestyle or behavioural data. So while the available data may accurately model *average* levels of exposure on individuals in given environments, the variation in *actual* exposure levels of individuals within those environments may be large, owing to the way those individuals interact with their environment. The amount of variation owing to behavioural and lifestyle factors may be greater in urban areas than in rural areas, reducing the strength of signal in urban areas. Previous studies have identified this absence of

lifestyle-related data as a source of potential bias and confounding in epidemiolocal air pollution studies,[155] and have recommended adding data on human time-activity patterns and other personal exposure factors to improve the accuracy of models.[156,157]

3. *A non-linear exposure-response curve*

Univariate regression analysis assumed linearity in relationships between dependent and independent variables (transcripts and TRAPs). It is possible that gene expression responds differently to TRAP exposure at much higher levels of exposure than at lower levels, reducing the power of the univariate analysis. However, interaction analysis of all TRAPS and urbanicity did not reveal any significant interactions (see Appendix B). Furthermore, while no literature exists specifically examining exposure-response curves at a transcriptomic level, the response curve of general health outcomes and diseases to air pollution has been widely investigated, and a linear relationship is found to be suitable.[158–161]

AGE

The age-stratified analysis revealed strikingly different results for the 28–33-year-old cohort – there is very little correlation in significance levels or effect direction and size with the full-cohort findings. While it is possible that age is an effect modifier for some biological reason, it is much more likely that the anomalous findings are a result of imperfect study design. The set-up of this study, with land-use regression models used to model exposure, is calibrated to pick up longer-term impacts of TRAP exposure, so the same exposure measurement will be modelled for someone who has lived in a city for 20 years as for someone who has just moved there from a rural area, which is a significant weakness in the methodology. Evidence shows[162] that address mobility decreases with age: younger people are likely to have lived at their address for less time than older people, so, as hypothesised in the Methods section for Stratified analysis, this study design may be more effective at picking up results in older people. This hypothesis in undermined, however, by the fact that the youngest cohort (17–27) exhibit more consistency with the full-cohort results. It seems likely that there may be unmeasured variables associated with lifestyle that our model is not accounting for, which may dilute the results in the 28–33 stratum.

One possible unmeasured variable that may be a contributory factor in both the age anomaly and the variation in results between urban and rural areas is car ownership and time spent driving in cars: 30 year olds are much more likely to own a car in the Netherlands than 20 year olds (see **Figure**

27)[163] and therefore highly likely to spend more time driving. This effect may be in reverse relation to the urbanicity of where a person lives, as inhabitants of rural areas are much more likely to travel in and own a car:[164] research shows rural residents take more than twice as many trips by car per years as urban residents,[165] and that 50.1% of transport miles travelled in the Dutch countryside are by car, compared with 30.6% of transport miles in large cities[166]. UK data shows that rural dwellers travel nearly three times as many miles per year than urbanites (**Figure 28**).[167]

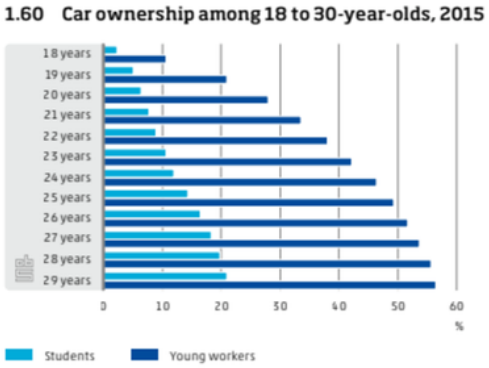

Figure 27 Car ownership in the Netherlands in 2015
Reproduced from *Trends in the Netherlands 2018*, Central Bureau of Statistics Netherlands, 2018[163]

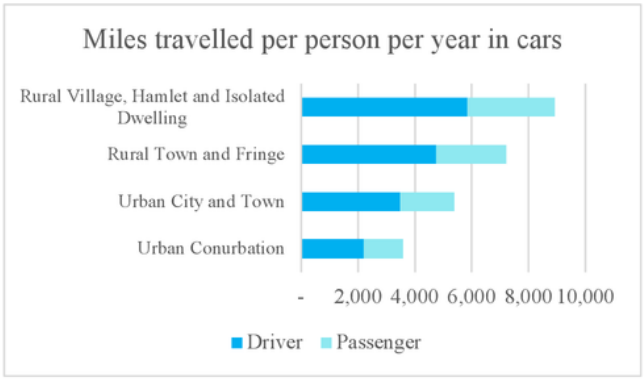

Figure 28 Chart showing miles travelled annually in cars, per person per year, in urban and rural areas in the UK.
Data taken from UK government land management statistics: *Transport and Travel in Rural Areas*, 2016.[167]

Without person-specific data on car ownership and use in this cohort, it is only speculation as to whether this variable contributes to the anomalous stratified results. The addition of this data would therefore significantly enhance the analysis.

GENDER

The results of the stability analysis within the gender-stratified results, combined with existing results in the literature indicating differential health impacts of air pollution in women and men,[84–87] provide a fairly strong evidence base that gender may play a significant role in the way gene expression is affected by air pollution. The scope for any more conclusive statements within this study is limited: without individual-specific lifestyle data it is impossible to make any claims about whether this is a true biological gender difference, or a result of different behavioural and lifestyle factors between the genders. Nonetheless, the results are sufficiently interesting to merit further investigation.

Study design and technical comments

LIMITATIONS

Some limitations of this study design have already been discussed in relation to the findings: the inherent level of inaccuracy in LUR as a measure of exposure; the likelihood that unmeasured variables are reducing the strength of the signal; the lack of time-at-address data; the assumptions of linearity and a univariate relationship that underpin the initial univariate analysis. Other limitations are addressed here.

Unmeasured confounding

The potential for unmeasured confounding – for air pollution to be in part a proxy for other variables relating to traffic and urban living – has been controlled and investigated as much as possible in this analysis but can never be eliminated. To take one example: exposure to noise has been reported to have an impact on cardiovascular health that shares similarities with TRAP exposure,[168,169] although the impact at a transcriptomic level has not been investigated. In an epidemiological study such as this, where noise levels would be modelled using LUR in the same way that air pollution is, it would be very difficult to separate the effects even if modelled noise measurements were available, but it is worth acknowledging the potential effects nonetheless.

Unrepresentative cohort

The gender-imbalanced cohort also presented an analytical challenge in disambiguating the effects of statistical power from any real gender differences in the results, and showed the value of a having a well-balanced, representative study population.

Lack of fidelity

The regression models used to predict exposure levels are highly weighted towards regional background exposure levels. The clustering analysis in this study shows that the largest-scale geographical variables are the most significant in predicting exposure clusters (even though the LUR models may be using smaller-scale variables in the regression). This indicates an innate lack of spatial fidelity in the models – they cannot accurately capture very localised variations in exposure level.

With this fact in mind, it is worth noting that all significant findings in this study were driven by PM2.5. PM2.5 is the only exposure that does not correlate strongly with urbanicity (see **Figure 29**).

This lack of correlation is explained by the fact that PM2.5 levels are less strongly driven by local sources and more by regional sources.[105] It seems likely, therefore, that LUR-based analysis may be more effective at modelling PM2.5, and therefore better at picking up the effects of PM2.5 exposure, than other exposures, because PM2.5 levels are less variable over small geographical areas than other exposures. This is a potential source of bias that could be addressed in further research and analysis.

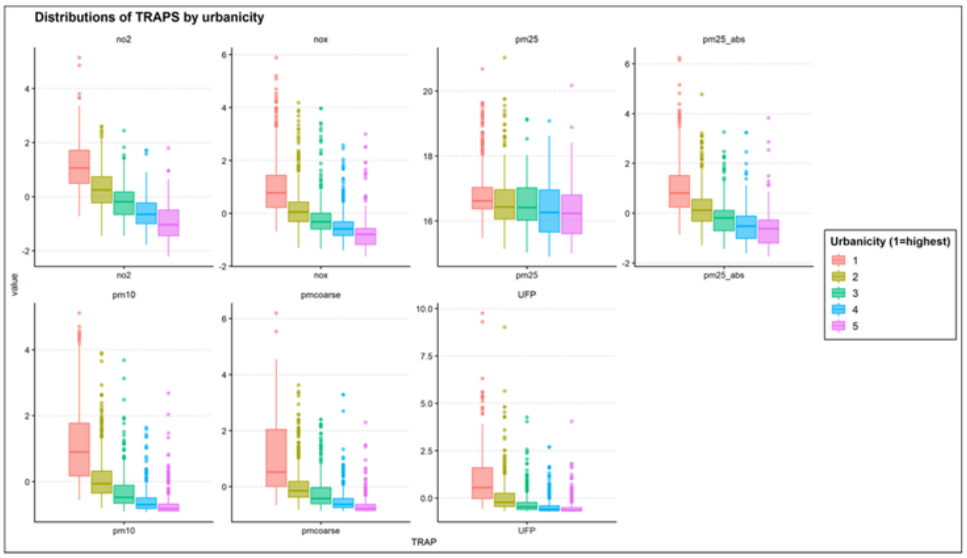

Figure 29 Boxplots showing distribution of modelled main TRAP exposure levels, by urbanicity. PM2.5 is the only exposure that does not clearly and substantially correlate with urbanicity.

FURTHER RESEARCH

The results in this study identify a molecular signature of TRAP exposure, but linking this signature to health outcomes through functional analysis and biological pathway analysis is somewhat speculative. A far better approach to joining the dots between exposure and health outcomes via the internal exposome would be to combine the analysis in this study with real health outcomes data. This could be done either by acquiring and analysing health data for the study participants (this data does exist as part of the NTR twin cohort study, it was just not made available for this analysis), or done at an epidemiological scale by analysing geographical health and disease patterns and how they relate to land use and environment, and linking this analysis to the results of this study.

The slightly surprising results relating to urbanicity and the proposed explanations relating to modelling accuracy and fidelity naturally raise the question of what can be done to improve the models. Land Use Regression modelling focuses on building models that are the best predictors of exposure levels at stationary measurement points in urban or rural locations – and the variable selection and calibration methods to improve these models are a field of study in themselves. But to date efforts to examine how modelled background exposure relates to actual individual exposure (as commonly measured by PEMs in small scale studies) have been limited, and have produced results that indicate a high level of variation between the two. The Research Triangle Park Particulate Matter Panel Study, a 1-year study run by the US Environmental Protection Agency, compared particulate matter measurements from PEM devices with ambient measurements, and found only a moderate correlation between mean personal exposures and ambient measurements ($r=0.39$).[170]

Cohen *et al*,[107] outline an 'ensemble' method of combining multiple LUR prediction measurements at different levels of fidelity to improve modelling accuracy. Extending this approach to a 'hybrid' model, as advocated by Baxter *et al*[56], by combining information from PEM studies and LUR studies, using LUR to model background base levels and personal lifestyle data (commute times, length of time spent outside, preferred modes of transport, etc) to add an individual component to the exposure estimates, would improve accuracy of measurements and reduce bias in studies such as this, and over time hopefully lead to more replicable results.

REFERENCES

1. Seaton, A., Godden, D., MacNee, W. & Donaldson, K. Particulate air pollution and acute health effects. *Lancet* **345**, 176–178 (1995).
2. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525**, 367–371 (2015).
3. Holgate, S. T. 'Every breath we take: the lifelong impact of air pollution' - a call for action. *Clin. Med.* **17**, 8 (2017).
4. Hime, N. J., Marks, G. B. & Cowie, C. T. A Comparison of the Health Effects of Ambient Particulate Matter Air Pollution from Five Emission Sources. *Int. J. Environ. Res. Public Health* **15**, (2018).
5. Fusco, D. *et al.* Air pollution and hospital admissions for respiratory conditions in Rome, Italy. *Eur. Respir. J.* **17**, 1143–50 (2001).
6. Cohen, A. J. *et al.* The Global Burden of Disease Due to Outdoor Air Pollution. *J. Toxicol. Environ. Heal. Part A* **68**, 1301–1307 (2005).
7. Air Pollution and Hospital Admissions for Respiratory Disease on JSTOR.
8. Oberdörster, G. & Utell, M. J. Ultrafine particles in the urban air: to the respiratory tract--and beyond? *Environ. Health Perspect.* **110**, A440-1 (2002).
9. Sinharay, R. *et al.* Respiratory and cardiovascular responses to walking down a traffic-polluted road compared with walking in a traffic-free area in participants aged 60 years and older with chronic lung or heart disease and age-matched healthy controls: a randomised, crossover study. *Lancet* **391**, 339–349 (2018).
10. Dominici, F. *et al.* Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. *JAMA* **295**, 1127 (2006).
11. Pope, C. A. *et al.* Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution. *Circulation* **109**, 71–77 (2004).
12. Brook, R. D. *et al.* Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation* **109**, 2655–71 (2004).
13. Brook, R. D. *et al.* Particulate Matter Air Pollution and Cardiovascular Disease. *Circulation* **121**, 2331–2378 (2010).
14. Brook, R. D. *et al.* Particulate Matter Air Pollution and Cardiovascular Disease. *Circulation* (2010). doi:10.1161/CIR.0B013E3181DBECE1
15. Tagliabue, G. *et al.* Atmospheric fine particulate matter and breast cancer mortality: a population-based cohort study. *BMJ Open* **6**, e012580 (2016).
16. Krewski, D. *et al.* Overview of the Reanalysis of the Harvard Six Cities Study and American Cancer Society Study of Particulate Air Pollution and Mortality. *J. Toxicol. Environ. Heal. Part A* **66**, 1507–1552 (2003).
17. Cohen, A. J. & Pope, C. A. Lung cancer and air pollution. *Environ. Health Perspect.* **103**, 219–224 (1995).
18. Trédaniel, J. *et al.* Pollution atmosphérique et cancer bronchique : données épidémiologiques. *Rev. Mal. Respir.* **26**, 437–445 (2009).
19. Shima, M. & Yoda, Y. An Ecological Study of Lung Cancer Mortality and Severe Air Pollution in the 1960s in an Industrial City in Japan. *Asian J. Atmos. Environ.* **3**, 9–18 (2009).
20. Visser, O., van Wijnen, J. H. & van Leeuwen, F. E. Incidence of cancer in the area around Amsterdam Airport Schiphol in 1988–2003: a population-based ecological study. *BMC Public Health* **5**, 127 (2005).
21. Beeson, W. L., Abbey, D. E. & Knutsen, S. F. Long-term concentrations of ambient air pollutants and incident lung cancer in California adults: results from the AHSMOG study.Adventist Health Study on Smog. *Environ.*

Health Perspect. **106**, 813–823 (1998).

22. Hemminki, K. & Pershagen, G. Cancer risk of air pollution: epidemiological evidence. *Environ. Health Perspect.* **102**, 187–192 (1994).

23. Carnow, B. W. The "urban factor" and lung cancer: cigarette smoking or air pollution? *Environ. Health Perspect.* **22**, 17–21 (1978).

24. Samet, J. M. Fine Particulate Air Pollution and Mortality in 20 U.S. Cities, 1987-1994. *N. Engl. J. Med.* **343**, 1742–1749

25. Dockery, D. W. *et al.* An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.* **329**, 1753–9 (1993).

26. WHO | Air pollution. *WHO* (2019).

27. 68% of the world population projected to live in urban areas by 2050, says UN | UN DESA | United Nations Department of Economic and Social Affairs. Available at: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. (Accessed: 27th June 2019)

28. Needham, J., Wang, L., Métailie, G. & Huang, H. T. *Science and civilisation in China.*

29. World Energy Resources: 2013 Survey. Available at: https://www.worldenergy.org/publications/2013/world-energy-resources-2013-survey/. (Accessed: 27th June 2019)

30. *Global Sources of Local Pollution.* (National Academies Press, 2009). doi:10.17226/12743

31. Anderson, H. R., Atkinson, R. W., Peacock, J. L., Marston, L. & Konstantinou, K. *META-ANALYSIS OF TIME-SERIES STUDIES AND PANEL STUDIES OF PARTICULATE MATTER (PM) AND OZONE (O 3).*

32. Energy Agency, I. *Energy and Air Pollution - World Energy Outlook 2016 Special Report.* (2016).

33. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).

34. Vineis, P., van Veldhoven, K., Chadeau-Hyam, M. & Athersuch, T. J. Advancing the application of omics-based biomarkers in environmental epidemiology. *Environ. Mol. Mutagen.* **54**, 461–467 (2013).

35. Chadeau-Hyam, M. *et al.* Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers. *Environ. Mol. Mutagen.* **54**, 542–557 (2013).

36. Wild, C. P. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol. Biomarkers Prev.* **14**, 1847–1850 (2005).

37. Wild, C. P. The exposome: from concept to utility. *Int. J. Epidemiol.* **41**, 24–32 (2012).

38. Vrijheid, M. The exposome: a new paradigm to study the impact of environment on health. *Thorax* **69**, 876–8 (2014).

39. Chadeau-Hyam, M. *et al.* Meeting-in-the-middle using metabolic profiling – a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers* **16**, 83–88 (2011).

40. Chadeau-Hyam, M. *et al.* Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers. *Environ. Mol. Mutagen.* **54**, 542–557 (2013).

41. Rice, T. K., Schork, N. J. & Rao, D. C. Methods for Handling Multiple Testing. *Adv. Genet.* **60**, 293–308 (2008).

42. Panasevich, S. *et al.* Interaction between air pollution exposure and genes in relation to levels of inflammatory markers and risk of myocardial infarction. *BMJ Open* **3**, e003058 (2013).

43. Goodson, J. M., MacDonald, J. W., Bammler, T. K., Chien, W.-M. & Chin, M. T. In utero exposure to diesel exhaust is associated with alterations in neonatal cardiomyocyte transcription, DNA methylation and metabolic perturbation. *Part. Fibre Toxicol.* **16**, (2019).

44. Bind, M.-A. *et al.* Air pollution and gene-specific methylation in the Normative Aging Study: Association, effect modification, and mediation analysis. *Epigenetics* **9**, 448 (2014).

45. Ljubimova, J. Y. *et al.* Coarse particulate matter (PM2.5-10) in Los Angeles Basin air induces expression of inflammation and cancer biomarkers in rat brains. *Sci. Rep.* **8**, 5708 (2018).

46. de F C Lichtenfels, A. J. *et al.* Long-term Air Pollution Exposure, Genome-wide DNA Methylation and Lung Function in the LifeLines Cohort Study. *Environ. Health Perspect.* **126**, 027004 (2018).

47. Plusquin, M. *et al.* DNA methylation and exposure to ambient air pollution in two prospective cohorts. *Environ. Int.* **108**, 127–136 (2017).

48. Jiang, C.-L. *et al.* Air pollution and DNA methylation alterations in lung cancer: A systematic and comparative study. *Oncotarget* **8**, 1369–1391 (2017).

49. Li, H. *et al.* Short-term exposure to fine particulate air pollution and genome-wide DNA methylation: A randomized, double-blind, crossover trial. *Environ. Int.* **120**, 130–136 (2018).

50. Wittkopp, S. *et al.* Nrf2-related gene expression and exposure to traffic-related air pollution in elderly subjects with cardiovascular disease: An exploratory panel study. *J. Expo. Sci. Environ. Epidemiol.* **26**, 141–9 (2016).

51. Zhu, J. *et al.* Effects of Different Components of PM2.5 on the Expression Levels of NF-κB Family Gene mRNA and Inflammatory Molecules in Human Macrophage. *Int. J. Environ. Res. Public Health* **16**, (2019).

52. Mostafavi, N. *et al.* Associations Between Genome-wide Gene Expression and Ambient Nitrogen Oxides. *Epidemiology* **28**, 320–328 (2017).

53. Saenen, N. D. *et al.* In Utero Fine Particle Air Pollution and Placental Expression of Genes in the Brain-Derived Neurotrophic Factor Signaling Pathway: An ENVIRONAGE Birth Cohort Study. *Environ. Health Perspect.* (2015). doi:10.1289/EHP.1408549

54. Chu, J. *et al.* Gene expression network analyses in response to air pollution exposures in the trucking industry. *Environ. Heal.* **15**, 101 (2016).

55. Ouyang, Y. *et al.* Changes in gene expression in chronic allergy mouse model exposed to natural environmental PM2.5-rich ambient air pollution. *Sci. Rep.* **8**, 6326 (2018).

56. Villarreal-Calderon, R. *et al.* Up-regulation of mRNA ventricular PRNP prion protein gene expression in air pollution highly exposed young urbanites: endoplasmic reticulum stress, glucose regulated protein 78, and nanosized particles. *Int. J. Mol. Sci.* **14**, 23471–91 (2013).

57. Wang, T. W. *et al.* Gene-expression profiling of buccal epithelium among non-smoking women exposed to household air pollution from smoky coal. *Carcinogenesis* **36**, 1494 (2015).

58. Tripathi, P., Deng, F., Scruggs, A. M., Chen, Y. & Huang, S. K. Variation in doses and duration of particulate matter exposure in bronchial epithelial cells results in upregulation of different genes associated with airway disorders. *Toxicol. In Vitro* **51**, 95–105 (2018).

59. Sancini, G. *et al.* Health risk assessment for air pollutants: alterations in lung and cardiac gene expression in mice exposed to Milano winter fine particulate matter (PM2.5). *PLoS One* **9**, e109685 (2014).

60. Mostafavi, N. *et al.* Inflammatory markers in relation to long-term air pollution. *Environ. Int.* **81**, 1–7 (2015).

61. Liang, D. *et al.* Use of high-resolution metabolomics for the identification of metabolic signals associated with traffic-related air pollution. *Environ. Int.* **120**, 145 (2018).

62. Eeftens, M. *et al.* Elemental Composition of Particulate Matter and the Association with Lung Function. *Epidemiology* **25**, 648–657 (2014).

63. Willemsen, G. *et al.* The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies. *Twin Res. Hum. Genet.* **13**, 231–245 (2010).

64. Eeftens, M. *et al.* Development of Land Use Regression Models for $PM_{2.5}$, $PM_{2.5}$ Absorbance, PM_{10} and PM_{coarse} in 20 European Study Areas; Results of the ESCAPE Project. *Environ. Sci. Technol.* **46**, 11195–11205 (2012).

65. Gramsch, E., Ormeño, I., Palma, G., Cereceda-Balic, F. & Oyola, P. Use of the light absorption coefficient to monitor elemental carbon and PM2.5--example of Santiago de Chile. *J. Air Waste Manag. Assoc.* **54**, 799–808 (2004).

66. Kelly, F. J. & Fussell, J. C. Size, source and chemical composition as determinants of toxicity attributable to

ambient particulate matter. *Atmos. Environ.* **60**, 504–526 (2012).

67. Shirmohammadi, F. *et al.* The relative importance of tailpipe and non-tailpipe emissions on the oxidative potential of ambient particles in Los Angeles, CA. *Faraday Discuss.* **189**, 361–80 (2016).

68. Weinmayr, G. *et al.* Particulate matter air pollution components and incidence of cancers of the stomach and the upper aerodigestive tract in the European Study of Cohorts of Air Pollution Effects (ESCAPE). *Environ. Int.* **120**, 163–171 (2018).

69. Beelen, R. *et al.* Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmos. Environ.* **72**, 10–23 (2013).

70. Eeftens, M. *et al.* Development of Land Use Regression Models for $PM_{2.5}$, $PM_{2.5}$ Absorbance, PM_{10} and PM_{coarse} in 20 European Study Areas; Results of the ESCAPE Project. *Environ. Sci. Technol.* **46**, 11195–11205 (2012).

71. van Nunen, E. *et al.* Land Use Regression Models for Ultrafine Particles in Six European Areas. *Environ. Sci. Technol.* **51**, 3336–3345 (2017).

72. de Hoogh, K. *et al.* Development of Land Use Regression Models for Particle Composition in Twenty Study Areas in Europe. *Environ. Sci. Technol.* **47**, 5778–5786 (2013).

73. Agency, E. E. CORINE Land Cover database. *CORINE Land Cover Database* (1995). Available at: https://www.eea.europa.eu/publications/COR0-landcover.

74. Nationaal Wegenbestand en WEGGEG | Rijkswaterstaat. Available at: https://www.rijkswaterstaat.nl/zakelijk/zakendoen-met-rijkswaterstaat/werkwijzen/werkwijze-in-gww/data-eisen-rijkswaterstaatcontracten/nationaal-wegenbestand.aspx. (Accessed: 9th July 2019)

75. Theune, C. J. *Structuurvisie Buisleidingen Inschatting Groepsrisico.* (2010).

76. Municipal size and urbanity. Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/overig/gemeentegrootte-en-stedelijkheid. (Accessed: 9th July 2019)

77. Normalized Difference Vegetation Index - an overview | ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/normalized-difference-vegetation-index. (Accessed: 9th July 2019)

78. Buuren, S. van & Groothuis-Oudshoorn, K. **mice** : Multivariate Imputation by Chained Equations in *R. J. Stat. Softw.* **45**, (2011).

79. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30**, 377–399 (2011).

80. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).

81. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

82. Winsorize. in *Encyclopedia of Research Design* (SAGE Publications, Inc.). doi:10.4135/9781412961288.n502

83. TOP10NL - Productinformatie - Kadaster zakelijk. Available at: https://zakelijk.kadaster.nl/en/-/top10nl. (Accessed: 9th July 2019)

84. Demoulin-Alexikova, S. *et al.* Impact of Air Pollution on Age and Gender Related Increase in Cough Reflex Sensitivity of Healthy Children in Slovakia. *Front. Physiol.* **7**, 54 (2016).

85. Clougherty, J. E. A growing role for gender analysis in air pollution epidemiology. *Environ. Health Perspect.* **118**, 167–76 (2010).

86. Oiamo, T. H. & Luginaah, I. N. Extricating Sex and Gender in Air Pollution Research: A Community-Based Study on Cardinal Symptoms of Exposure. *Int. J. Environ. Res. Public Heal.* **10**, 3801–3817 (2013).

87. Butter, M. E. Are Women More Vulnerable to Environmental Pollution? *J. Hum. Ecol.* **20**, 221–226 (2006).

88. Granados-Canal, D. J., Chardon, B., Lefranc, A. & Gremy, I. Air Pollution and Respiratory Hospital Admissions in Greater Paris: Exploring Sex Differences. *Arch. Environ. Occup. Health* **60**, 307–313 (2005).

89. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **72**, 417–473 (2010).

90. Zou, H., Hastie, T. & Zou, M. H. *Package 'elasticnet' Title Elastic-Net for Sparse Estimation and Sparse PCA.* (2018).

91. Bottolo, L., Chadeau-Hyam, M., Liquet, B. & Richardson, S. GUESS: GPU-based C++ software for Bayesian variable selection regression of multiple responses. *Australas. Appl. Stat. Conf.* (2012).

92. Liquet, B., Bottolo, L., Campanella, G., Richardson, S. & Chadeau-Hyam, M. **R2GUESS** : A Graphics Processing Unit-Based *R* Package for Bayesian Variable Selection Regression of Multivariate Responses. *J. Stat. Softw.* **69**, 1–32 (2016).

93. Bottolo, L. *et al.* ESS++: a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics* **27**, 587–8 (2011).

94. Beineke, P. *et al.* A whole blood gene expression-based signature for smoking status. *BMC Med. Genomics* **5**, 58 (2012).

95. DAVID Functional Annotation Bioinformatics Microarray Analysis. Available at: https://david.ncifcrf.gov/. (Accessed: 7th August 2019)

96. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**, 3448–3449 (2005).

97. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **28**, 100 (1979).

98. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. CRAN - Package NbClust. (2015).

99. Hennig, C. Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* **52**, 258–271 (2007).

100. clusterboot function | R Documentation. Available at: https://www.rdocumentation.org/packages/fpc/versions/2.2-3/topics/clusterboot. (Accessed: 11th August 2019)

101. fpc package | R Documentation. Available at: https://www.rdocumentation.org/packages/fpc/versions/2.2-3. (Accessed: 11th August 2019)

102. Max Kuhn Contributions from Jed Wing, A. *et al.* Package 'caret' Title Classification and Regression Training Description Misc functions for training and plotting classification and regression models. 2 (2019).

103. London Councils. *Demystifying Air Pollution in London.* (2018).

104. DEFRA. *UK Air Pollution.* (2003).

105. Air Quality Expert Group. *Fine Particulate Matter (PM2.5) in the United Kingdom.* (2012).

106. Mostafavi, N. *et al.* Inflammatory markers in relation to long-term air pollution. *Environ. Int.* **81**, 1–7 (1978).

107. Cohen, G. *et al.* Cancer and mortality in relation to traffic-related air pollution among coronary patients: Using an ensemble of exposure estimates to identify high-risk individuals. *Environ. Res.* **176**, 108560 (2019).

108. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36**, 40–45 (2004).

109. The Human Protein Atlas. *Human Protein Atlas* Available at: https://www.proteinatlas.org. (Accessed: 17th August 2019)

110. WikiGenes - WIPF2 - WAS/WASL interacting protein family, member 2. Available at: https://www.wikigenes.org/e/gene/e/147179.html. (Accessed: 15th August 2019)

111. Culibrk, L. *et al.* Phagocytosis of Aspergillus fumigatus by Human Bronchial Epithelial Cells Is Mediated by the Arp2/3 Complex and WIPF2. *Front. Cell. Infect. Microbiol.* **9**, 16 (2019).

112. Zhou, C. *et al.* Difference of molecular alterations in HER2-positive and HER2-negative gastric cancers by whole-genome sequencing analysis. *Cancer Manag. Res.* **Volume 10**, 3945–3954 (2018).

113. Lehtinen, L. *et al.* High-throughput RNAi screening for novel modulators of vimentin expression identifies MTHFD2 as a regulator of breast cancer cell migration and invasion. *Oncotarget* **4**, 48–63 (2013).

114. Carim-Todd, L., Sumoy, L., Andreu, N., Estivill, X. & Escarceller, M. Identification and characterization of

BTBD1, a novel BTB domain containing gene on human chromosome 15q24. *Gene* **262**, 275–281 (2001).

115. No Title. *GeneCards* Available at: https://www.genecards.org/cgi-bin/carddisp.pl?gene=TRAT1.

116. Ma, L. *et al.* Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. *BMC Med. Genet.* **11**, 55 (2010).

117. Li, H. *et al.* Protein Prenylation Constitutes an Endogenous Brake on Axonal Growth. *Cell Rep.* **16**, 545–558 (2016).

118. RCOR2. *GeneCards* Available at: https://www.genecards.org/cgi-bin/carddisp.pl?gene=RCOR3.

119. NEDD1. *WikiGenes* Available at: https://en.wikipedia.org/wiki/NEDD1. (Accessed: 17th August 2019)

120. SEL1L gene - Semantic Scholar. Available at: https://www.semanticscholar.org/topic/SEL1L-gene/468769. (Accessed: 18th August 2019)

121. Mellai, M. *et al.* SEL1L Plays a Major Role in Human Malignant Gliomas. *J. Pathol. Clin. Res.* cjp2.134 (2019). doi:10.1002/cjp2.134

122. Ishida, S., Andreux, P., Poitry-Yamate, C., Auwerx, J. & Hanahan, D. Bioavailable copper modulates oxidative phosphorylation and growth of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19507–12 (2013).

123. Lui, K. H. *et al.* The effects of particle-induced oxidative damage from exposure to airborne fine particulate matter components in the vicinity of landfill sites on Hong Kong. *Chemosphere* **230**, 578–586 (2019).

124. Badaloni, C. *et al.* Effects of long-term exposure to particulate matter and metal components on mortality in the Rome longitudinal study. *Environ. Int.* **109**, 146–154 (2017).

125. M. Frawley Cass, S. & J. Tepe, J. Identification of Phosphoproteins and their Impact as Biomarkers in Cancer Therapeutics. *Curr. Signal Transduct. Ther.* **6**, 113–140 (2011).

126. *Current signal transduction therapy.* (Bentham Science Publishers, 2006).

127. Takano, S. *et al.* Increased circulating cell signalling phosphoproteins in sera are useful for the detection of pancreatic cancer. *Br. J. Cancer* **103**, 223–231 (2010).

128. Olszewski, U., Deally, A., Tacke, M. & Hamilton, G. Alterations of Phosphoproteins in NCI-H526 Small Cell Lung Cancer Cells Involved in Cytotoxicity of Cisplatin and Titanocene Y. *Neoplasia* **14**, 813–822 (2012).

129. Baker, A. F. *et al.* Stability of Phosphoprotein as a Biological Marker of Tumor Signaling. *Clin. Cancer Res.* **11**, 4338–4340 (2005).

130. Carter, A. M. *et al.* Phosphoprotein-based Biomarkers as Predictors for Cancer Therapy. *bioRxiv* 675637 (2019). doi:10.1101/675637

131. Gilbert, W. Why genes in pieces? *Nature* **271**, 501–501 (1978).

132. Wang, Y. *et al.* Mechanism of alternative splicing and its regulation. *Biomed. reports* **3**, 152–158 (2015).

133. Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1792**, 14–26 (2009).

134. Escobar-Hoyos, L., Knorr, K. & Abdel-Wahab, O. Aberrant RNA Splicing in Cancer. *Annu. Rev. Cancer Biol.* **3**, 167–185 (2019).

135. Srebrow, A. & Kornblihtt, A. R. The connection between splicing and cancer. *J. Cell Sci.* **119**, 2635–41 (2006).

136. El Marabti, E. & Younis, I. The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer. *Front. Mol. Biosci.* **5**, 80 (2018).

137. Shapiro, I. M. *et al.* An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet.* **7**, e1002218 (2011).

138. Lu, Z. *et al.* Transcriptome-wide Landscape of Pre-mRNA Alternative Splicing Associated with Metastatic Colonization. *Mol. Cancer Res.* **13**, 305–318 (2015).

139. Tang, J.-Y. *et al.* Alternative splicing for diseases, cancers, drugs, and databases. *ScientificWorldJournal.* **2013**, 703568 (2013).

140. Singh, B. & Eyras, E. The role of alternative splicing in cancer. *Transcription* **8**, 91–98 (2017).

141. Brinkman, B. M. N. Splice variants as cancer biomarkers. *Clin. Biochem.* **37**, 584–594 (2004).

142. Gough, N. R. Cancerous splice variants. *Sci. Signal.* **9**, ec115–ec115 (2016).

143. Li, X. *et al.* A splicing switch from ketohexokinase-C to ketohexokinase-A drives hepatocellular carcinoma formation. *Nat. Cell Biol.* **18**, 561–571 (2016).

144. Bagyinszky, E. *et al.* Early-onset Alzheimer's disease patient with prion (PRNP) p.Val180Ile mutation. *Neuropsychiatr. Dis. Treat.* **Volume 15**, 2003–2013 (2019).

145. Cousyn, L. *et al.* First European case of Creutzfeldt-Jakob disease with a PRNP G114V mutation. *Cortex* **117**, 407–413 (2019).

146. Pettit, A. P. *et al.* Alteration of peripheral blood monocyte gene expression in humans following diesel exhaust inhalation. *Inhal. Toxicol.* **24**, 172–81 (2012).

147. Chu, J.-H. *et al.* Gene expression network analyses in response to air pollution exposures in the trucking industry. doi:10.1186/s12940-016-0187-z

148. Tang, Y. *et al.* CYP1B1 expression promotes the proangiogenic phenotype of endothelium through decreased intracellular oxidative stress and thrombospondin-2 expression. *Blood* **113**, 744–754 (2009).

149. Stanek, L. W., Sacks, J. D., Dutton, S. J. & Dubois, J.-J. B. Attributing health effects to apportioned components and sources of particulate matter: An evaluation of collective results. *Atmos. Environ.* **45**, 5655–5663 (2011).

150. Proxies: the example of traffic-related air pollution | Integrated Environmental Health Impact Assessment System. Available at: http://www.integrated-assessment.eu/eu/guidebook/proxies_example_traffic_related_air_pollution.html. (Accessed: 17th August 2019)

151. Ryan, P. H. & LeMasters, G. K. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhal. Toxicol.* **19 Suppl 1**, 127–33 (2007).

152. Janssen, S., Dumont, G., Fierens, F. & Mensink, C. Spatial interpolation of air pollution measurements using CORINE land cover data. *Atmos. Environ.* **42**, 4884–4903 (2008).

153. Mukerjee, S. *et al.* Seasonal effects in land use regression models for nitrogen dioxide, coarse particulate matter, and gaseous ammonia in Cleveland, Ohio. *Atmos. Pollut. Res.* **3**, 352–361 (2012).

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

155. Strak, M. *et al.* Associations between lifestyle and air pollution exposure: Potential for confounding in large administrative data cohorts. *Environ. Res.* **156**, 364–373 (2017).

156. Baxter, L. K. *et al.* Exposure prediction approaches used in air pollution epidemiology studies: Key findings and future recommendations. *J. Expo. Sci. Environ. Epidemiol.* **23**, 654–659 (2013).

157. Breen, M. S. *et al.* Air Pollution Exposure Model for Individuals (EMI) in Health Studies: Evaluation for Ambient $PM_{2.5}$ in Central North Carolina. *Environ. Sci. Technol.* **49**, 14184–14194 (2015).

158. Samoli, E. *et al.* Investigating the dose-response relation between air pollution and total mortality in the APHEA-2 multicity project. *Occup. Environ. Med.* **60**, 977–82 (2003).

159. Samoli, E. *et al.* Estimating the exposure-response relationships between particulate matter and mortality within the APHEA multicity project. *Environ. Health Perspect.* **113**, 88–95 (2005).

160. Li Liu, Li-Ya Yu, Hui-Juan Mu, Li-Ying Xing, Yan-Xia Li, G.-W. P. Shape of concentration-response curves between long-term particulate matter exposure and morbidities of chronic bronchitis: a review of epidemiological evidence - Liu - Journal of Thoracic Disease. *J. Thorac. Dis.* **6**, (2014).

161. Daniels, M. J., Dominici, F., Samet, J. M. & Zeger, S. L. Estimating Particulate Matter-Mortality Dose-Response Curves and Threshold Levels: An Analysis of Daily Time-Series for the 20 Largest US Cities. *Am. J. Epidemiol.* **152**, 397–406 (2000).

162. Hansen, K. A. *Current Population Reports Seasonality of Moves and Duration of Residence.* **1**, (1998).

163. Central Bureau of Statistics, N. *Trends in the Netherlands 2018*. (2018).

164. Statistics Netherlands. *Transport and Mobility 2016*. (2016).

165. Department for Transport. *Travel in Urban and Rural areas.* (2010).

166. De Vos, J. The influence of land use and mobility policy on travel behavior: A comparative case study of Flanders and the Netherlands. *J. Transp. Land Use* **8**, 171 (2015).

167. Rural transport, travel and accessibility statistics - GOV.UK. Available at: https://www.gov.uk/government/statistics/rural-transport-travel-and-accessibility-statistics. (Accessed: 22nd August 2019)

168. Babisch, W., Beule, B., Schust, M., Kersten, N. & Ising, H. Traffic noise and risk of myocardial infarction. *Epidemiology* **16**, 33–40 (2005).

169. Davies, H. W., Vlaanderen, J. J., Henderson, S. B. & Brauer, M. Correlation between co-exposures to noise and air pollution from traffic sources. *Occup. Environ. Med.* **66**, 347–350 (2009).

170. Williams, R. *et al.* The Research Triangle Park particulate matter panel study: PM mass concentration relationships. *Atmos. Environ.* **37**, 5349–5363 (2003).

171. Huan, T. *et al.* A systematic heritability analysis of the human whole blood transcriptome. *Hum. Genet.* **134**, 343–58 (2015).

172. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).

173. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).

174. Price, A. L. *et al.* Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. *PLoS Genet.* **7**, e1001317 (2011).

175. Göring, H. H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39**, 1208–1216 (2007).

APPENDICES

Appendix A

FURTHER DESCRIPTIVE STATISTICS

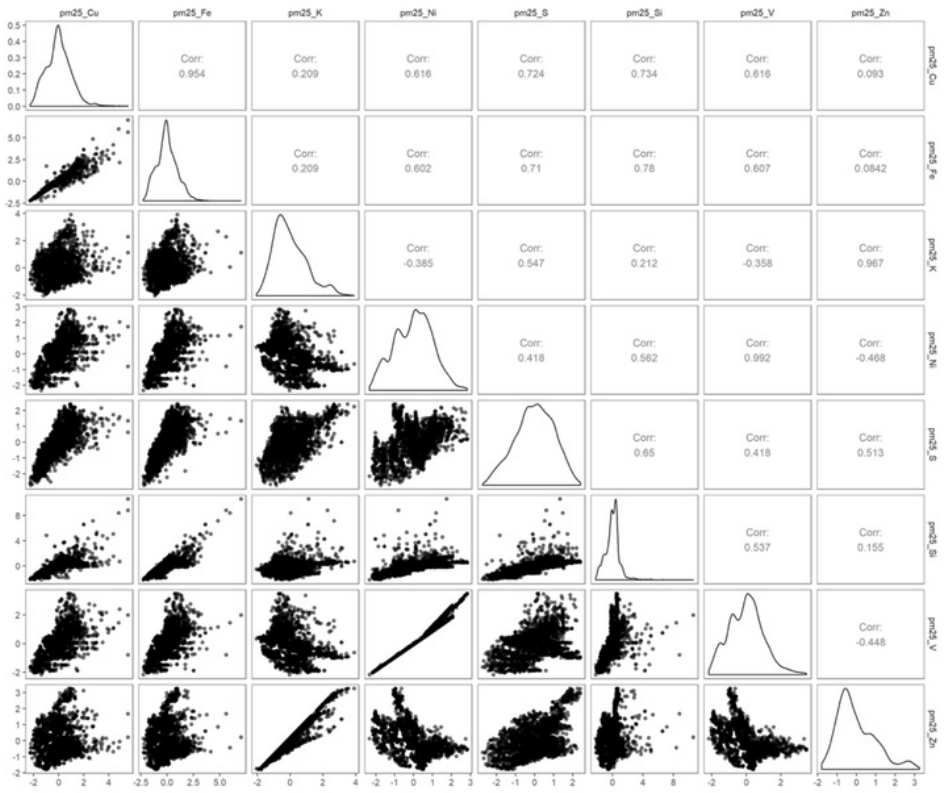

Figure 30 Comparative distributions for all elemental PM2.5 particles

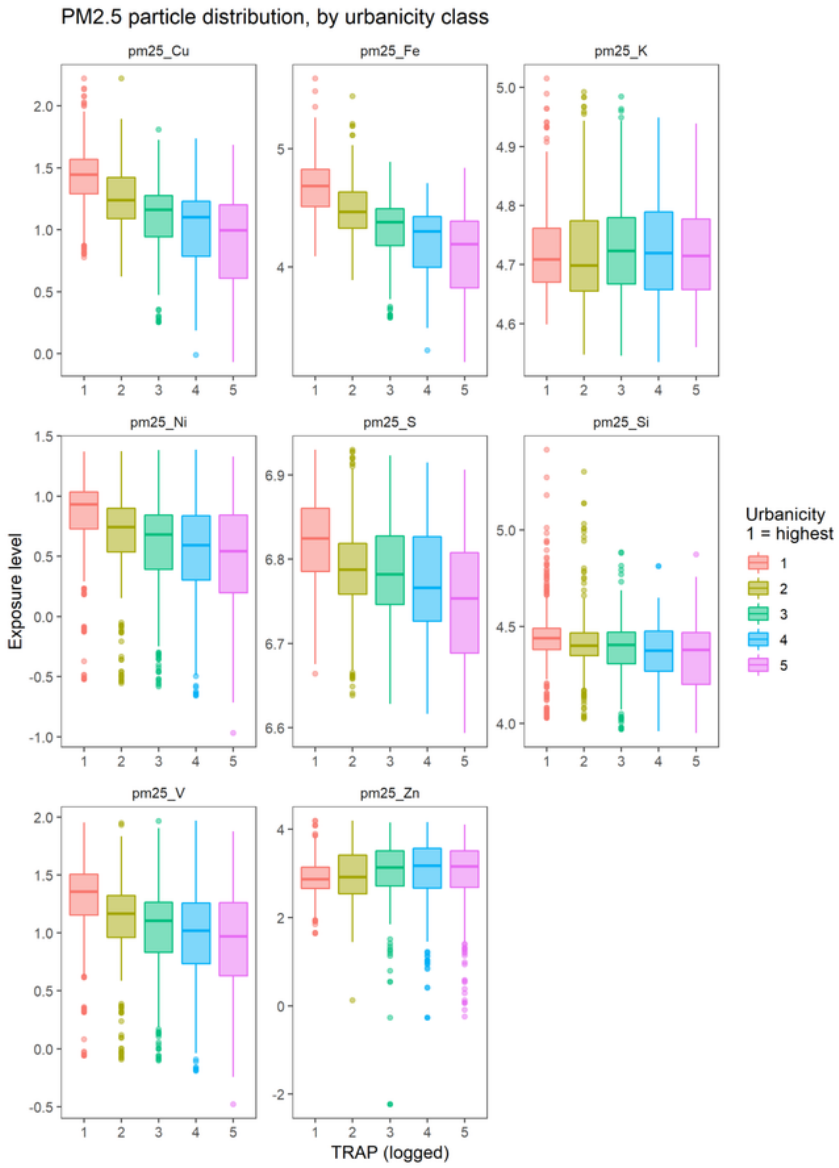

Figure 31 Distribution of PM2.5 elemental particle exposure levels, stratified by urbanicity

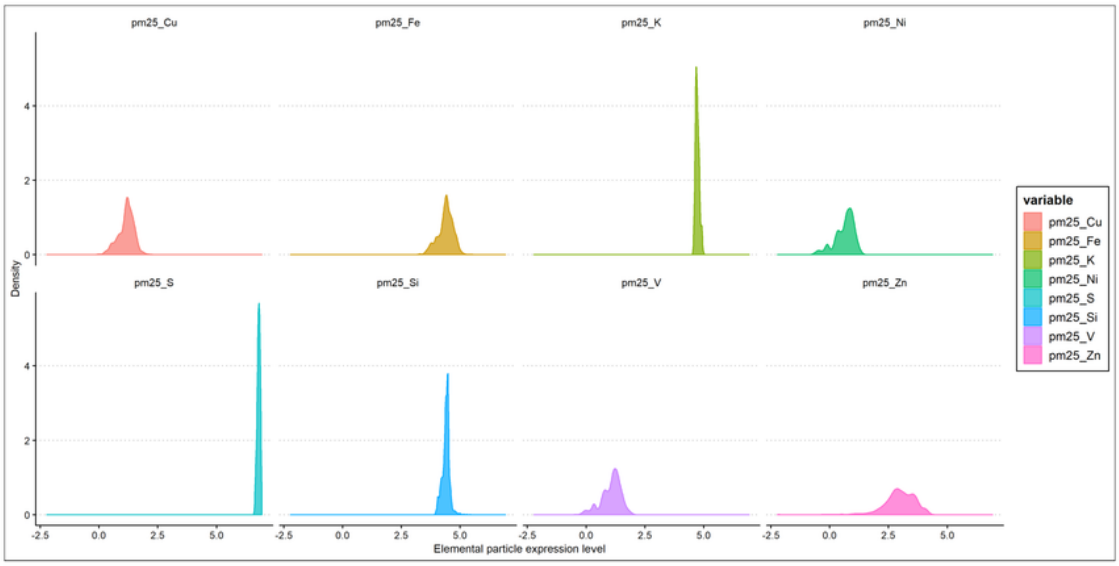

Figure 32 Elemental PM2.5 particle distribution (logged)

Appendix B

UNIVARIATE AND MULTIVARIATE APPROACHES

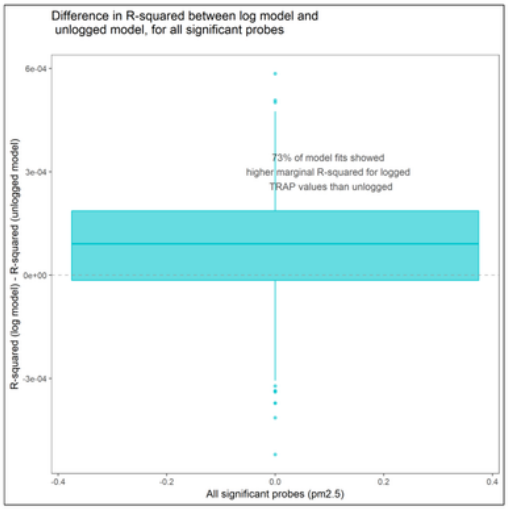

Figure 33 Distribution of deltas in R^2 figures between logged and unlogged models

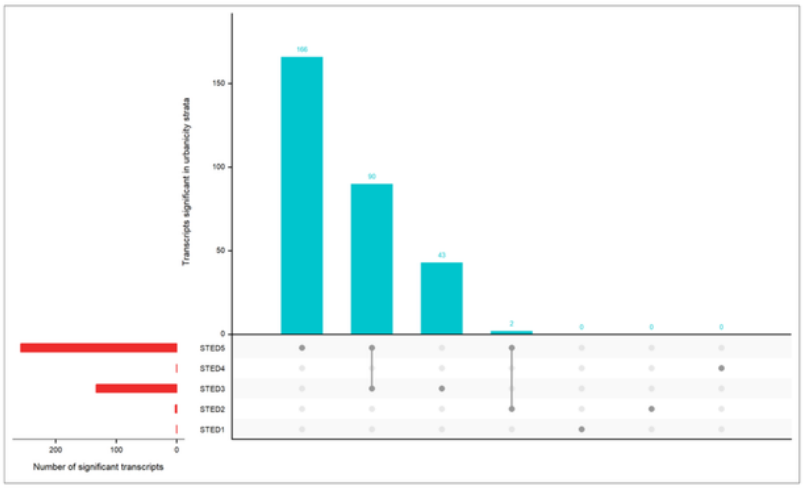

Figure 34 Upset plot showing significant findings by stratum in a stratified urbanicity model Connected dots in the lower right plot indicate where significant transcript are common to more than one urbanicity stratum

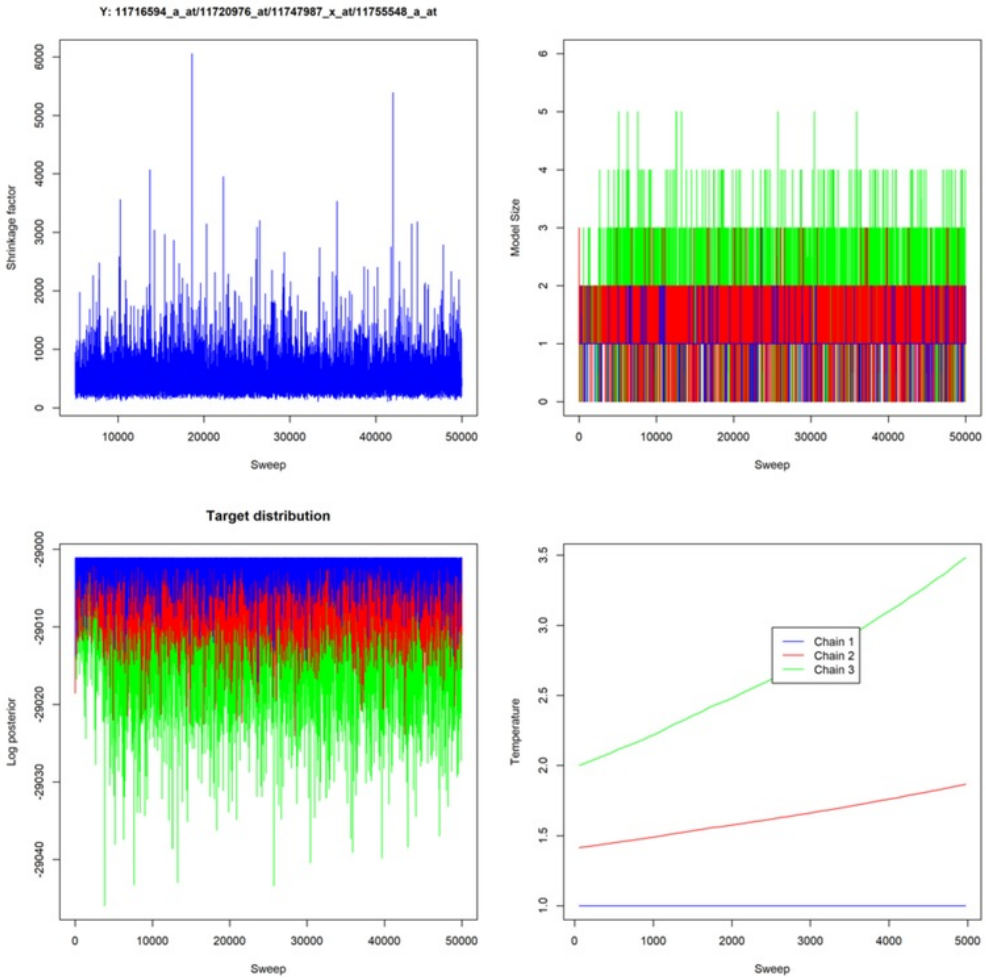

Figure 35 R2GUESS output plots showing convergence across three chains

INTERACTION ANALYSIS

All transcripts identified as being significantly associated with TRAP exposure in the initial screening process were investigated for interaction effects. Five variables were chosen: Age, Smoking status, BMI, Sex and Urbanicity. **Table 17** shows the number of significant ($p<0.05$) interaction effects after correction for multiple testing using the Benjamini-Hochberg procedure. **Figure 36** shows the distribution of p-values for each interaction tested.

Table 17 Significant interactions between TRAPs and five selected covariates
Significance level set at $p < 0.05$ after correction for multiple testing

TRAP	Age	Smoking	BMI	Sex	Urbanicity (STED)
no2	0	0	0	0	0
nox	0	0	0	0	0
pm25_abs	0	0	0	4	0
pm25	0	0	0	0	0
pm10	0	0	0	0	0
pmcoarse	0	0	0	0	0
UFP	0	0	0	0	0

The results show an absence of significant interaction effects in most of the variables, with only sex showing significant interaction effects with TRAPs after BH correction. This offers some support for the earlier analysis suggesting that gender is a factor in biological responses to air pollution, but does not shed any further light on the findings from the urbanicity-stratified analysis showing that effect size and strength of association vary with degree of urbanicity.

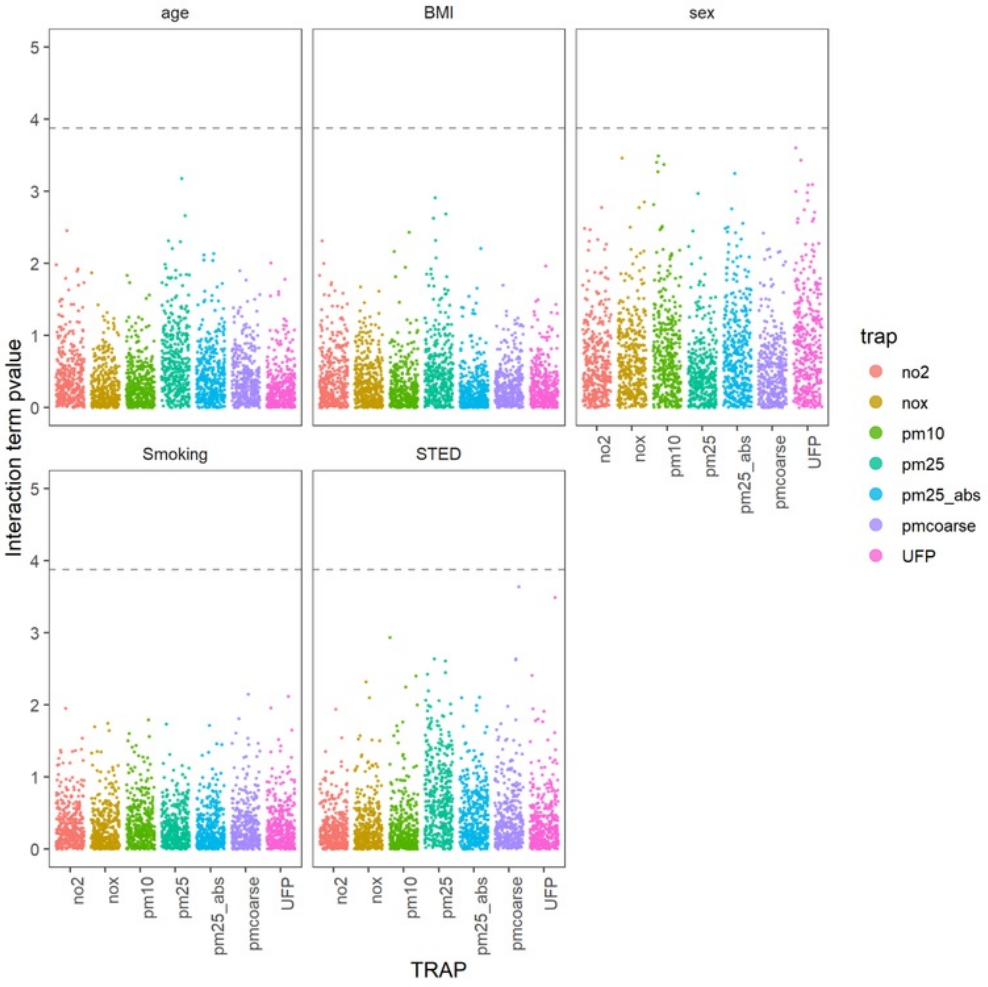

Figure 36 Plot showing p-values for interaction terms between TRAPs and other covariates: age, BMI, sex, smoking status and urbanicity.

FUNCTIONAL CLUSTER ANALYSIS USING DAVID

Table 18 Functional cluster annotation for all transcripts identified at BH pval < 0.1
Analysis carried out using DAVID Functional Annotation Tool[95]
Only 3 most significantly enriched clusters shown: full functional cluster chart included in supplementary material

Annotation Cluster 1							**Enrichment Score: 8.857965627318231**	
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**Bonferroni**	**BH**	**FDR**
UP_KEYWORDS	Zinc	157	19.38	1.40E-11	1.705	0.000	0.000	0.000
UP_KEYWORDS	Metal-binding	215	26.54	1.09E-10	1.506	0.000	0.000	0.000
UP_KEYWORDS	Zinc-finger	124	15.31	3.02E-10	1.776	0.000	0.000	0.000
GOTERM_MF_DIRECT	GO:0046872~metal ion binding	132	16.30	8.02E-06	1.450	0.007	0.002	0.013

Annotation Cluster 2							**Enrichment Score: 4.520132027250937**	
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**Bonferroni**	**BH**	**FDR**
UP_KEYWORDS	Transcription regulation	140	17.28	2.23E-07	1.531	0.000	0.000	0.000
UP_KEYWORDS	Transcription	141	17.41	6.62E-07	1.500	0.000	0.000	0.001
UP_KEYWORDS	DNA-binding	119	14.69	1.16E-05	1.480	0.005	0.000	0.016
GOTERM_MF_DIRECT	GO:0003677~DNA binding	107	13.21	6.78E-05	1.452	0.060	0.012	0.106
GOTERM_BP_DIRECT	GO:0006351~transcription, DNA-templated	117	14.44	3.22E-04	1.369	0.590	0.257	0.573
GOTERM_BP_DIRECT	GO:0006355~regulation of transcription, DNA-templated	83	10.25	0.0203	1.263	1.000	0.794	30.674

Annotation Cluster 3							**Enrichment Score: 4.462889761810322**	
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**Bonferroni**	**BH**	**FDR**
UP_KEYWORDS	Ubl conjugation pathway	62	7.65	1.46E-09	2.325	0.000	0.000	0.000
GOTERM_MF_DIRECT	GO:0004842~ubiquitin-protein transferase activity	31	3.83	1.33E-04	2.141	0.114	0.020	0.208
GOTERM_MF_DIRECT	GO:0016874~ligase activity	25	3.09	8.71E-04	2.104	0.547	0.084	1.351
GOTERM_BP_DIRECT	GO:0016567~protein ubiquitination	27	3.33	0.0083	1.721	1.000	0.685	13.830

Appendix C

UNSUPERVISED MACHINE LEARNING APPROACHES

Table 19 Modelled geographical and demographic variables, including descriptions, units, source and which cluster dimensions the variables were used for.

Variable name	Description	Unit	Source	Cluster dims
NATUR_500	Surface area of semi-natural and forested areas in buffers around address	m2	ESCAPE[64,69]	Geo
NATUR_1000	Surface area of semi-natural and forested areas in buffers around address	m2	ESCAPE[64,69]	Geo
NATUR_5000	Surface area of semi-natural and forested areas in buffers around address	m2	ESCAPE[64,69]	Geo
INDUS_100	Surface area assigned to industry in buffers around address	m2	ESCAPE[64,69]	Geo
INDUS_300	Surface area assigned to industry in buffers around address	m2	ESCAPE[64,69]	Geo
INDUS_500	Surface area assigned to industry in buffers around address	m2	ESCAPE[64,69]	Geo
INDUS_1000	Surface area assigned to industry in buffers around address	m2	ESCAPE[64,69]	Geo
INDUS_5000	Surface area assigned to industry in buffers around address	m2	ESCAPE[64,69]	Geo
LDRES_100	Surface area of low density residential land in buffers around address	m2	ESCAPE[64,69]	Geo
LDRES_300	Surface area of low density residential land in buffers around address	m2	ESCAPE[64,69]	Geo
LDRES_500	Surface area of low density residential land in buffers around address	m2	ESCAPE[64,69]	Geo
LDRES_1000	Surface area of low density residential land in buffers around address	m2	ESCAPE[64,69]	Geo
LDRES_5000	Surface area of low density residential land in buffers around address	m2	ESCAPE[64,69]	Geo
URBG_100	Surface area of urban green in buffers around address	m2	ESCAPE[64,69]	Geo
URBG_300	Surface area of urban green in buffers around address	m2	ESCAPE[64,69]	Geo
URBG_500	Surface area of urban green in buffers around address	m2	ESCAPE[64,69]	Geo
URBG_1000	Surface area of urban green in buffers around address	m2	ESCAPE[64,69]	Geo
URBG_5000	Surface area of urban green in buffers around address	m2	ESCAPE[64,69]	Geo
PORT_100	Surface area assigned to ports in buffers around address	m2	ESCAPE[64,69]	Geo
PORT_300	Surface area assigned to ports in buffers around address	m2	ESCAPE[64,69]	Geo
PORT_500	Surface area assigned to ports in buffers around address	m2	ESCAPE[64,69]	Geo
PORT_1000	Surface area assigned to ports in buffers around address	m2	ESCAPE[64,69]	Geo
PORT_5000	Surface area assigned to ports in buffers around address	m2	ESCAPE[64,69]	Geo
POP_100	Population density in buffers around address	n	ESCAPE[64,69]	Geo
POP_300	Population density in buffers around address	n	ESCAPE[64,69]	Geo
POP_500	Population density in buffers around address	n	ESCAPE[64,69]	Geo
POP_1000	Population density in buffers around address	n	ESCAPE[64,69]	Geo
POP_5000	Population density in buffers around address	n	ESCAPE[64,69]	Geo
HHOLD_100	Household density in buffers around address	n	ESCAPE[64,69]	Geo

HHOLD_300	Household density in buffers around address	n	ESCAPE[64,69]	Geo
HHOLD_500	Household density in buffers around address	n	ESCAPE[64,69]	Geo
HHOLD_1000	Household density in buffers around address	n	ESCAPE[64,69]	Geo
HHOLD_5000	Household density in buffers around address	n	ESCAPE[64,69]	Geo
INTINVD2	Product of traffic intensity on nearest road (TRAFNEAR) and inverse distance to nearest road squared (DISTINVNEAR2)	Vehicle-days/1m sq	ESCAPE[64,69]	Geo, Traffic
HTRAFNEAR	Heavy traffic intensity on nearest road	Vehicle-days	ESCAPE[64,69]	Geo, Traffic
HINTINVD	Product of heavy traffic intensity on nearest road (HTRAFNEAR) and inverse distance to nearest road (DISTINVNEAR1)	Vehicle-days	ESCAPE[64,69]	Geo, Traffic
DINVNEAR1	inverse distance to nearest road	m-1	ESCAPE[64,69]	Geo, Traffic
DINVNEAR2	inverse distance to nearest road squared	m-2	ESCAPE[64,69]	Geo, Traffic
HINTINVD2	Product of heavy traffic intensity on nearest road (HTRAFNEAR) and inverse distance to nearest road squared (DISTINVNEAR2)	Vehicle-days/1m sq	ESCAPE[64,69]	Geo, Traffic
TRAFNEAR	Traffic intensity on nearest road	Vehicle-days	ESCAPE[64,69]	Geo, Traffic
INTINVD	Product of traffic intensity on nearest road (TRAFNEAR) and inverse distance to nearest road (DISTINVNEAR)	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
DINVMAJOR2	Inverse squared distance to the nearest major road	m-2	ESCAPE[64,69]	Geo, Traffic
DINVMAJOR1	Inverse distance to the nearest major road	m-1	ESCAPE[64,69]	Geo, Traffic
INTMINVD	Product of traffic intensity on nearest major road (TRAFMAJOR) and inverse of distance to the nearest major road (DISTINVMAJOR1)	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TRAFMAJOR	Traffic intensity on nearest major road	Vehicle-days	ESCAPE[64,69]	Geo, Traffic
INTMINVD2	Product of traffic intensity on nearest major road (TRAFMAJOR) and inverse of distance to the nearest major road squared (DISTINVMAJOR2)	Vehicle-days/1m sq	ESCAPE[64,69]	Geo, Traffic
HTRAFMAJOR	Heavy-duty traffic intensity on nearest major road	Vehicle-days	ESCAPE[64,69]	Geo, Traffic
TMLOA_50	Total traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TMLOA_100	Total traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TMLOA_300	Total traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TMLOA_500	Total traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TMLOA_1000	Total traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HMLOA_50	Total heavy duty traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HMLOA_100	Total heavy duty traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HMLOA_300	Total heavy duty traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HMLOA_500	Total heavy duty traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HMLOA_1000	Total heavy duty traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TLOA_50	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TLOA_100	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TLOA_300	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
TLOA_500	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic

TLOA_1000	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HLOA_50	Total heavy-duty traffic load of all roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HLOA_100	Total heavy-duty traffic load of all roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HLOA_300	Total heavy-duty traffic load of all roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HLOA_500	Total heavy-duty traffic load of all roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
HLOA_1000	Total heavy-duty traffic load of all roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo, Traffic
RDL_50	Road length of all roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
RDL_100	Road length of all roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
RDL_300	Road length of all roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
RDL_500	Road length of all roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
RDL_1000	Road length of all roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
MRDL_50	Road length of major roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
MRDL_100	Road length of major roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
MRDL_300	Road length of major roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
MRDL_500	Road length of major roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
MRDL_1000	Road length of major roads in a buffer	m	ESCAPE[64,69]	Geo, Traffic
DINVNEARC1	Inverse distance to the nearest road	m-1	ESCAPE[64,69]	Geo, Traffic
DINVNEARC2	Inverse squared distance to the nearest road	m-2	ESCAPE[64,69]	Geo, Traffic
CMAJOCLASS	classification of the nearest major road based on the importance that the constituting road has in the total road network (NET2CLASS)	-	ESCAPE[64,69]	Geo, Traffic
DINVMAJOC1	Inverse distance to the nearest major road	m-1	ESCAPE[64,69]	Geo, Traffic
DINVMAJOC2	Inverse squared distance to the nearest major road	m-2	ESCAPE[64,69]	Geo, Traffic
EEA_100	Population density in buffers around address EEA database	n	ESCAPE[64,69]	Geo
EEA_300	Population density in buffers around address EEA database	n	ESCAPE[64,69]	Geo
EEA_500	Population density in buffers around address EEA database	n	ESCAPE[64,69]	Geo
EEA_1000	Population density in buffers around address EEA database	n	ESCAPE[64,69]	Geo
EEA_5000	Population density in buffers around address EEA database	n	ESCAPE[64,69]	Geo
TMLOA_25	Total traffic load of major roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo
HMLOA_25	Total heavy traffic load of major roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo
TLOA_25	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo
HLOA_25	Heavy traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Vehicle-days/1m	ESCAPE[64,69]	Geo
MRDL_25	Road length of major roads in a buffer	m	ESCAPE[64,69]	Geo
RDL_25	Road length of all roads in a buffer	m	ESCAPE[64,69]	Geo
NDVI_100m	Average normalized difference vegetation index in a buffer around the address	-	NDVI[77]	Geo
NDVI_300m	Average normalized difference vegetation index in a buffer around the address	-	NDVI[77]	Geo
NDVI_500m	Average normalized difference vegetation index in a buffer around the address	-	NDVI[77]	Geo

NDVI_1000m	Average normalized difference vegetation index in a buffer around the address	-	NDVI[77]	Geo
NDVI_3000m	Average normalized difference vegetation index in a buffer around the address	-	NDVI[77]	Geo
TOP10NL_100m	Proportion green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
TOP10NL_300m	Proportion green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
TOP10NL_500m	Proportion green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
TOP10NL_1000m	Proportion green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
TOP10NL_3000m	Proportion green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
agri_500m	Proportion agricultural green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
nature_500m	Proportion natural green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
urban_500m	proportion urban green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
agri_1000m	Proportion agricultural green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
nature_1000m	Proportion natural green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
urban_1000m	proportion urban green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
agricul_3000m	Proportion agricultural green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
nature_3000m	Proportion natural green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
urban_3000m	proportion urban green in a buffer around the address based on TOP10NL database	-	TOP10NL[83]	Geo
out_top10n	Adress lies within 3km of the border of the top10NL map	-	TOP10NL[83]	Geo
NDVI_out	Adress lies within 3km of the border of the NDVI map	-	NDVI[77]	Geo
P_ONGEHUWD	Unmarried	%	CBS[76]	Demo
P_GEHUWD	Married	%	CBS[76]	Demo
P_GESCHEID	Separated	%	CBS[76]	Demo
P_VERWEDUW	Widowed	%	CBS[76]	Demo
BEV_DICHTH	Population density	%	CBS[76]	Demo
P_WEST_AL	Western total	%	CBS[76]	Demo
P_N_W_AL	Non-western total	%	CBS[76]	Demo
P_MAROKKO	Percentage of immigrants from Morocco	%	CBS[76]	Demo
P_ANT_ARU	Percentage of immigrants from the Netherlands Antilles and Aruba	%	CBS[76]	Demo
P_SURINAM	Percentage of immigrants from Suriname	%	CBS[76]	Demo
P_TURKIJE	Percentage of immigrants from Turkey	%	CBS[76]	Demo
P_OVER_NW	Percentage of immigrants from other non-Western countries	%	CBS[76]	Demo
WOZ	Average house value	x 1 000 euro	CBS[76]	Demo
P_KOOPWON	Houses for sale	%	CBS[76]	Demo
P_HUURWON	Total rental properties	%	CBS[76]	Demo
P_LAAGINKH	Low-income households	%	CBS[76]	Demo
P_HOOGINKH	Households with a high income	%	CBS[76]	Demo

P_LKOOPKRH	Households with low purchasing power	%	CBS[76]	Demo
P_SOCMINH	Household below or around social minimum	%	CBS[76]	Demo
P_WWB_UIT	General assistance benefits relatively	per 1 000 households	CBS[76]	Demo
OPP_LAND	Land area	ha	CBS[76]	Geo
OPP_WATER	Surface water	ha	CBS[76]	Geo

CLUSTERING: FURTHER RESULTS AND VISUALISATION

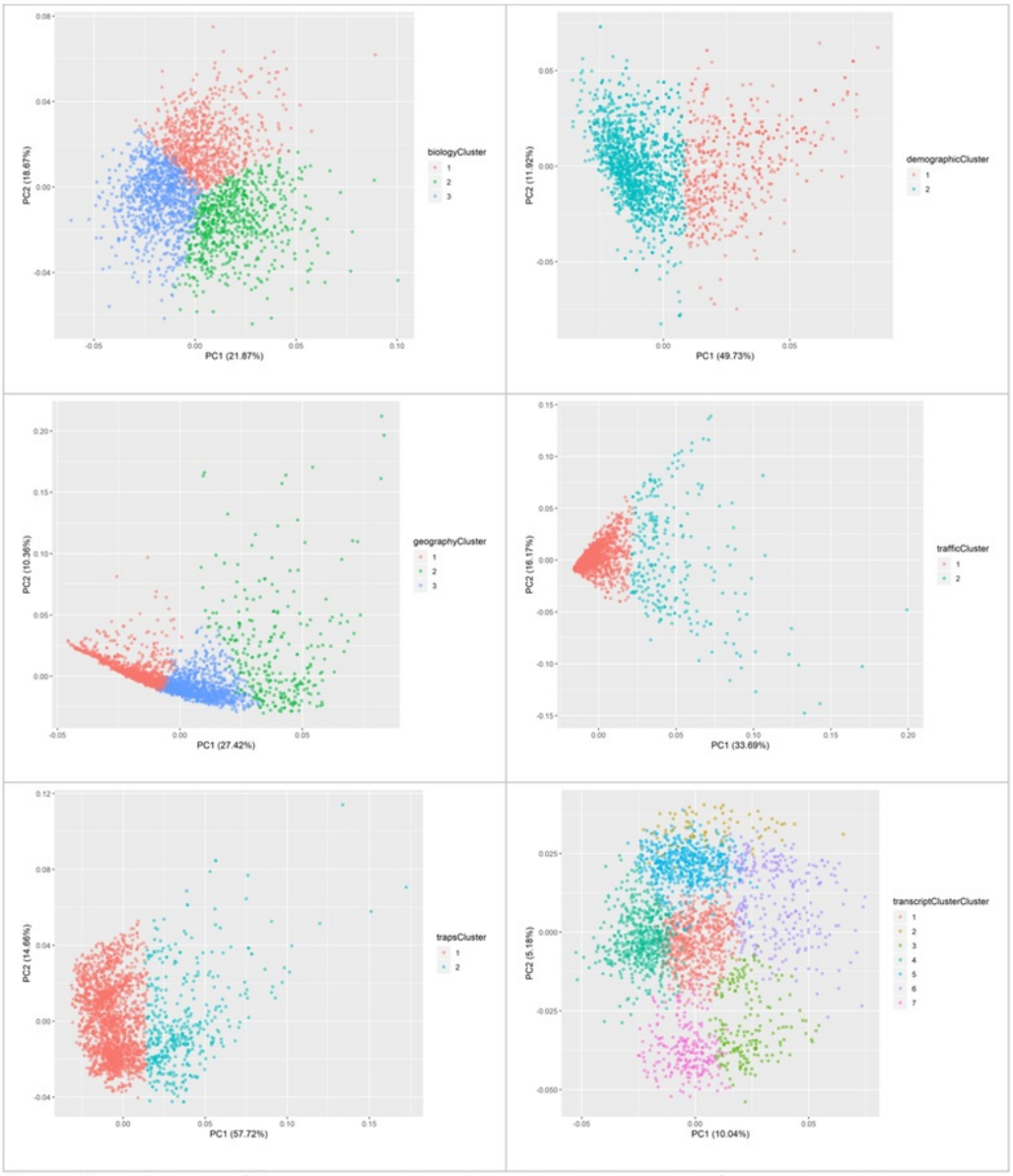

Figure 37 visualisations of clusters across the six clustering dimensions, using first two principal components of a principal component analysis. Clockwise from top left: biology, demography, geography, traffic, TRAPs, transcripts

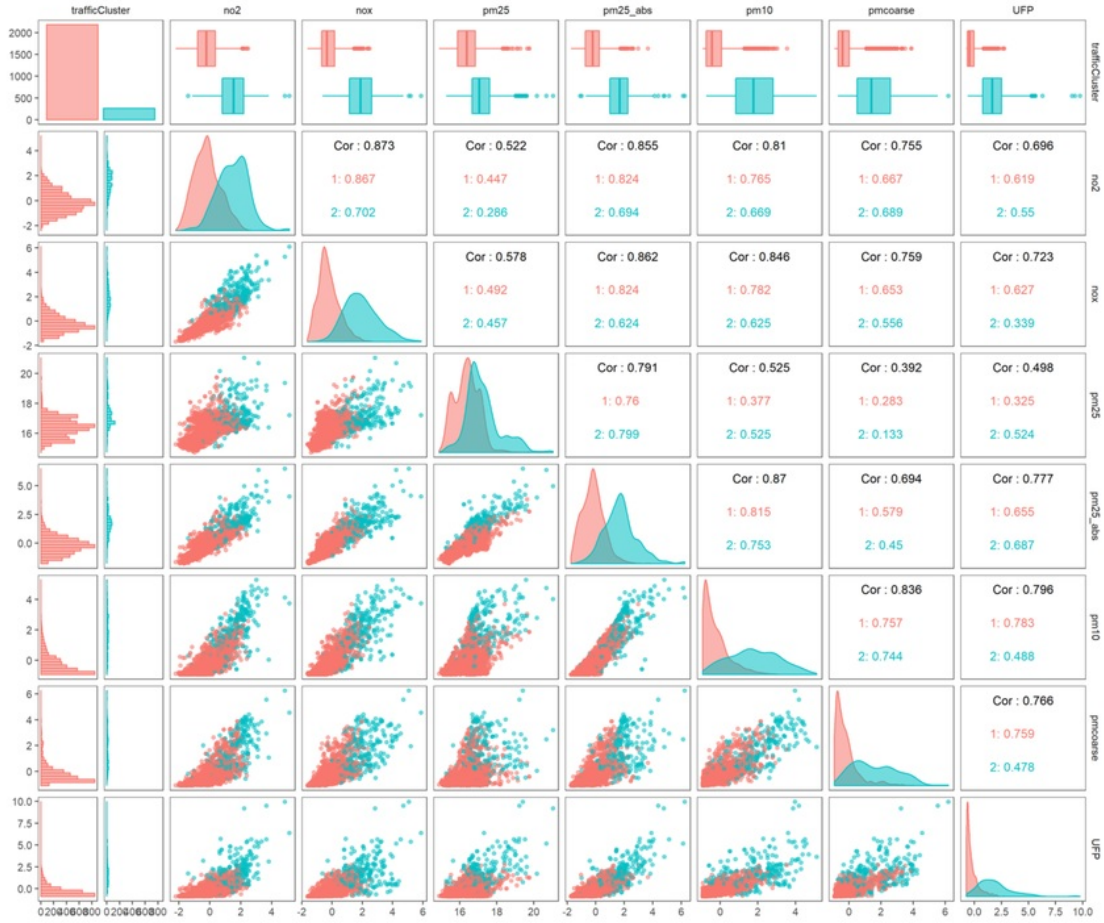

Figure 38 Comparison of exposure distributions for seven principal TRAPs, segmented by traffic cluster

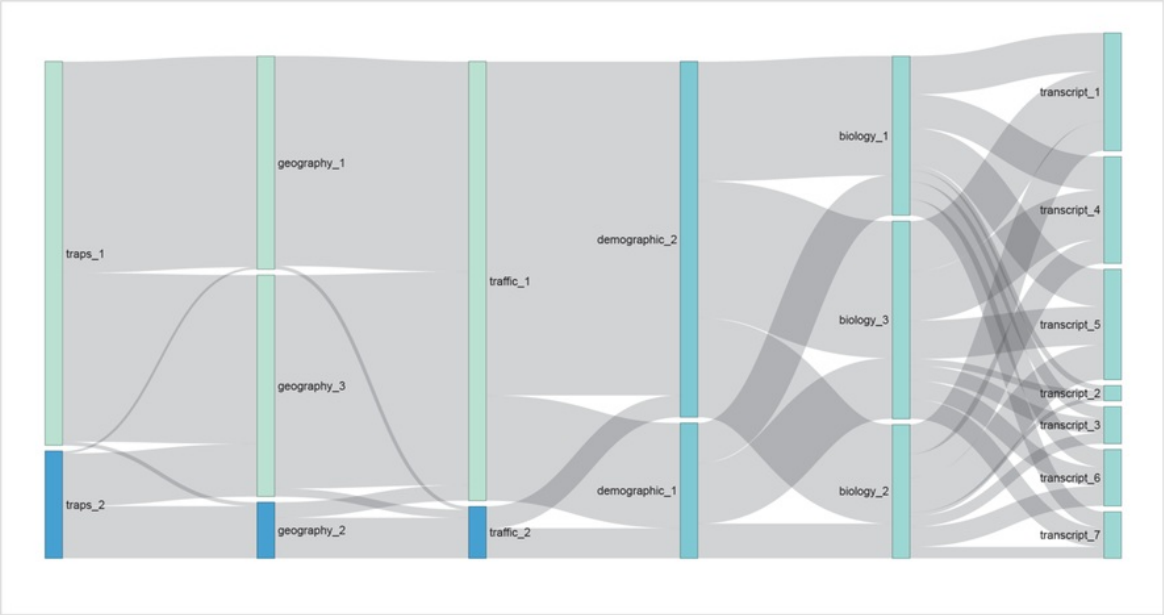

Figure 39 Full Sankey network showing how participants group within clusters across all dimensions. Dimensions, from left to right: TRAPS, Geography, Traffic, Demographic, Biology, Transcripts.

VARIABLE IMPORTANCE IN ALL CLUSTERING DIMENSIONS

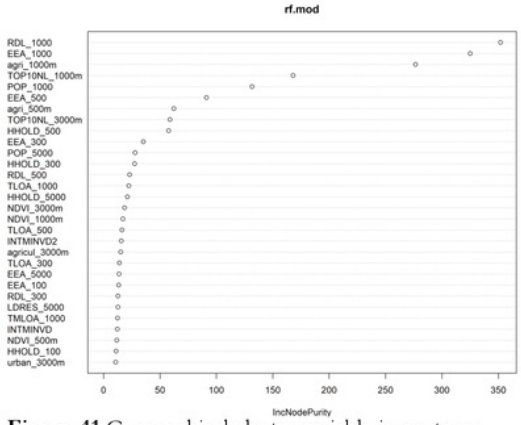

Figure 41 Geographical cluster variable importance

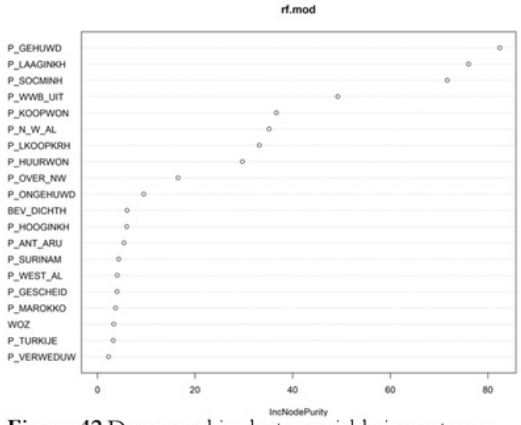

Figure 42 Demographic cluster variable importance

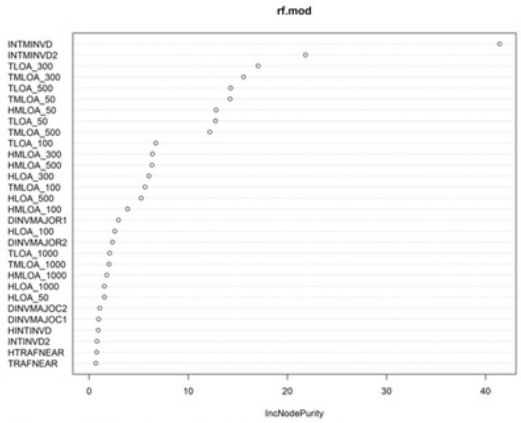

Figure 43 Traffic cluster variable importance

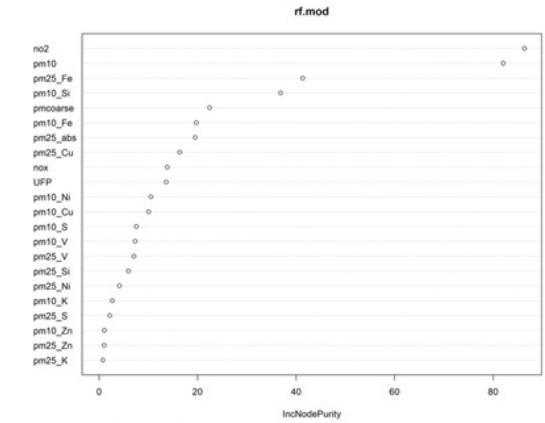

Figure 44 TRAP cluster variable importance

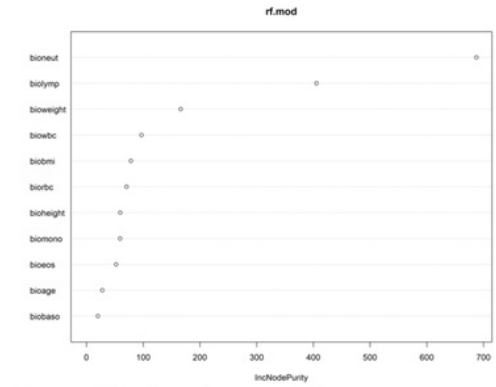

Figure 40 Biology cluster variable importance

Appendix D

MISCELLANEOUS

Heritability analysis of transcripts

Previous research estimates that around 40% of transcripts show some evidence of heritability, with the average proportion of heritability being 13%. A subset of 10% of transcripts showed evidence of being more than 20% heritable[171] (other research puts the estimated level of heritability higher[172–175]). This previous research deploys statistical methodology to separate directly genetic influences on gene expression from second-degree influences – where genetic factors may influence a person's choice of environment, and then their exposures in that environment influence their gene expression.

Computing the similarity in gene expression between twin pairs in this data set, and comparing the results for a denoised data set (with the effect of technical and demographic confounders removed) gives an indication of the level of 'direct' heritability in gene expression, relative to the second-degree influence.

To perform this analysis, the 842 monozygotic twin pairs in the data set were split into two groups of 421. For each transcript, the expression data for group A was regressed onto group B, and the p-value and R^2 value recorded. Table XXX shows the aggregated results for the same analysis performed on the raw transcript data and on the denoised data set. Figure XXX shows the QQ plots for the regressions.

The R^2 values are low – much lower than the proportions of heritability discovered in previous research, because this mode of analysis is comparatively quite a blunt instrument. However, the difference in the R^2 scores and number of significant transcripts in the raw and denoised data sets provides a neat illustration of the direct and indirect modes of heritability. In the raw data, the R^2 values are three times higher than in the denoised data – gene expression is being influenced by environmental factors that are common to the twin pairs and are at least in part attributable to common genetic profiles. In the denoised data set, the effect of several key environmental factors is removed, and the associations are much weaker.

Table 20 Results of monozygotic twin pair transcript regression

	Number of transcripts	Mean R^2	Number of significant transcripts (BH pval < 0.05)	Mean R^2 among significant transcripts
Denoised data	44,241	0.004396	676 (1.5%)	0.051905
Raw data	44,241	0.013512	10,101 (22.8%)	0.047007

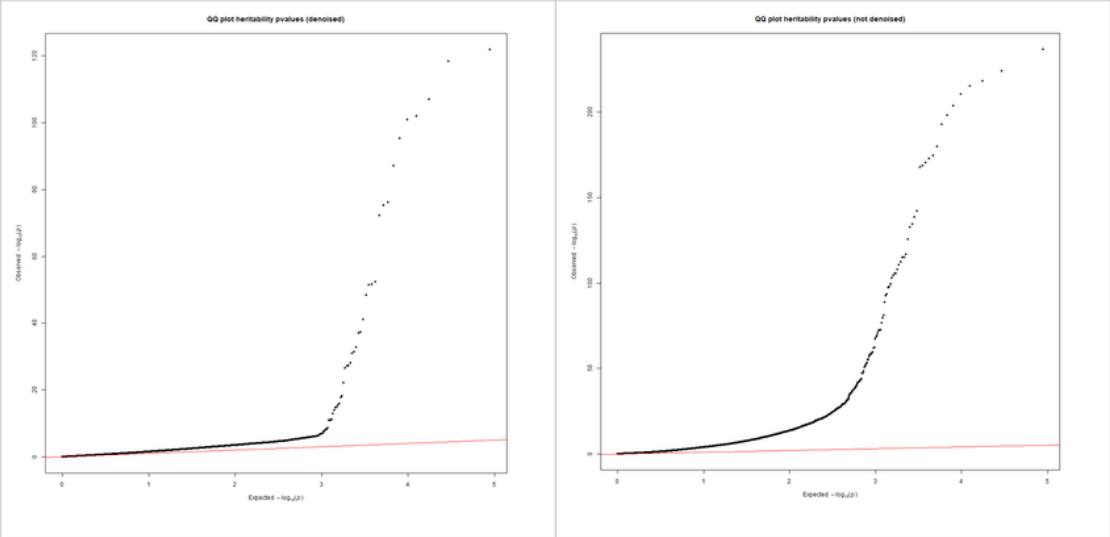

Figure 45 QQ plots for gene expression heritability regression. Denoised data set on the left.
Red lines indicate the expected plot path under the null assumption of no association (ie no heritability at all).

Smoking analysis

Univariate screening of the full dataset revealed 175 transcripts that were associated with smoking status (ever smoked / never smoked) at a BH-corrected p-value of 0.05. One of these transcripts was also significantly associated with pm2.5 exposure. The gene linked to the common transcript was TRAT-1, which has been associated with bone cancer[115]. Figure XXX shows volcano plots for the smoking univariate analysis.

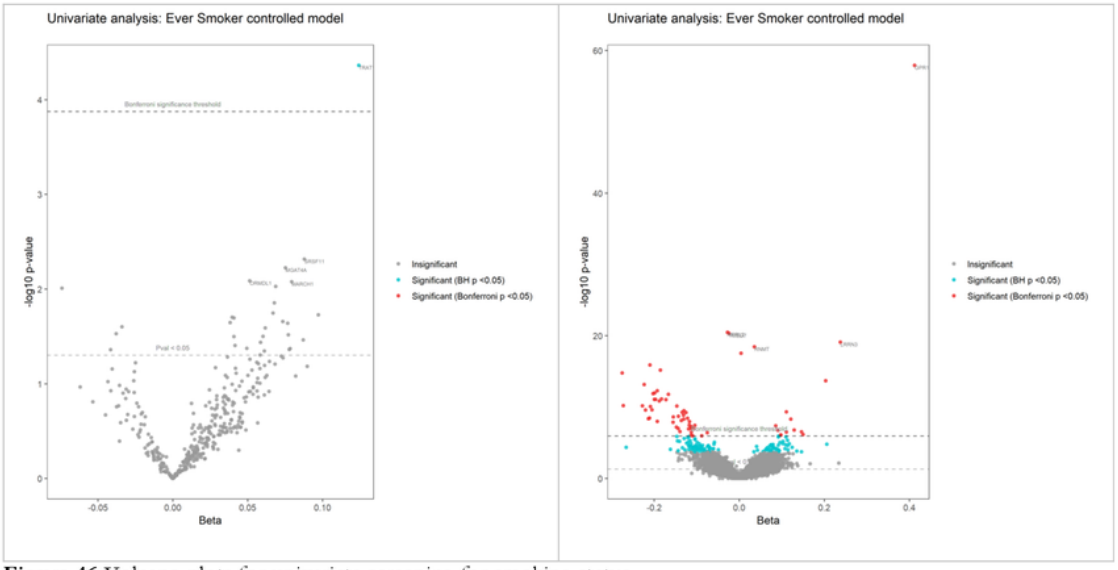

Figure 46 Volcano plots for univariate screening for smoking status
Left: only transcripts associated with pm2.5 shown; right: all transcripts shown.
The four transcripts with the lowest p-values are labelled with the appropriate gene name in each plot.

Figure XXX plots the beta coefficients for the univariate regression models for smoking status and pm2.5 exposure. The transcripts significantly associated with PM2.5 are highlighted. Of those transcripts associated with PM2.5, 81% had the same signed beta coefficient (χ^2 $p<0.01$), for smoking status as for PM2.5, indicating some significant overlap in the biological fingerprints for the two exposures.

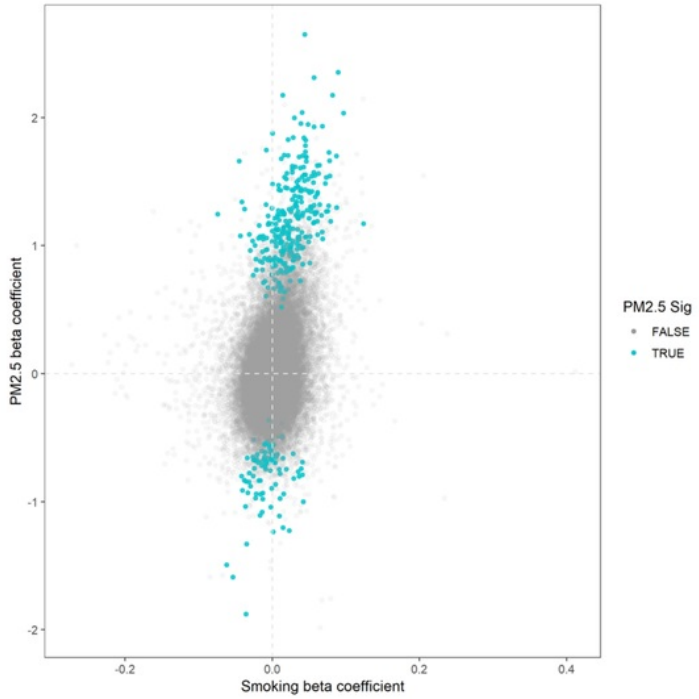

Figure 47 Regression coefficients for all transcripts. X-axis: smoking coefficients; Y-axis PM2.5 coefficients. Transcripts significantly associated with PM2.5 are highlighted in turquoise.

DIFFERENTIALLY EXPOSED TWIN REGRESSION ANALYSIS

Figure 48 shows QQ plots for the results of twin regression models in monozygotic and dizygotic twins, regressing transcript expression levels against PM2.5. **Figure 49** shows volcano plots for the same models. No significant results were found.

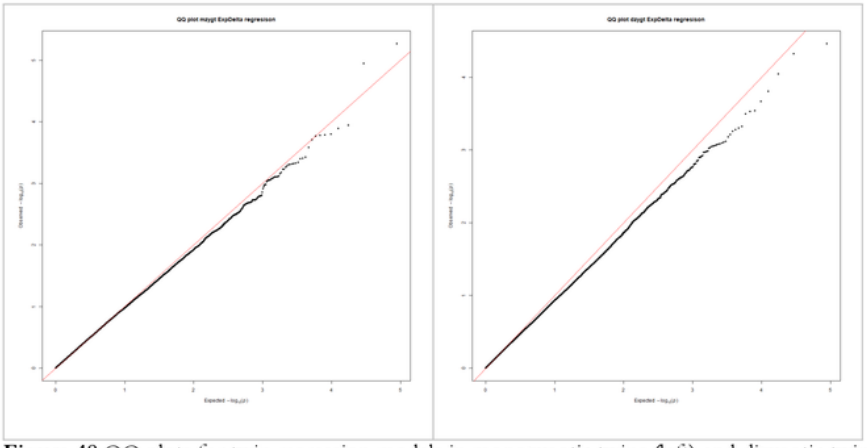

Figure 48 QQ plots for twin regression models in monozygotic twins (left) and dizygotic twins (right). No significant results were observed.

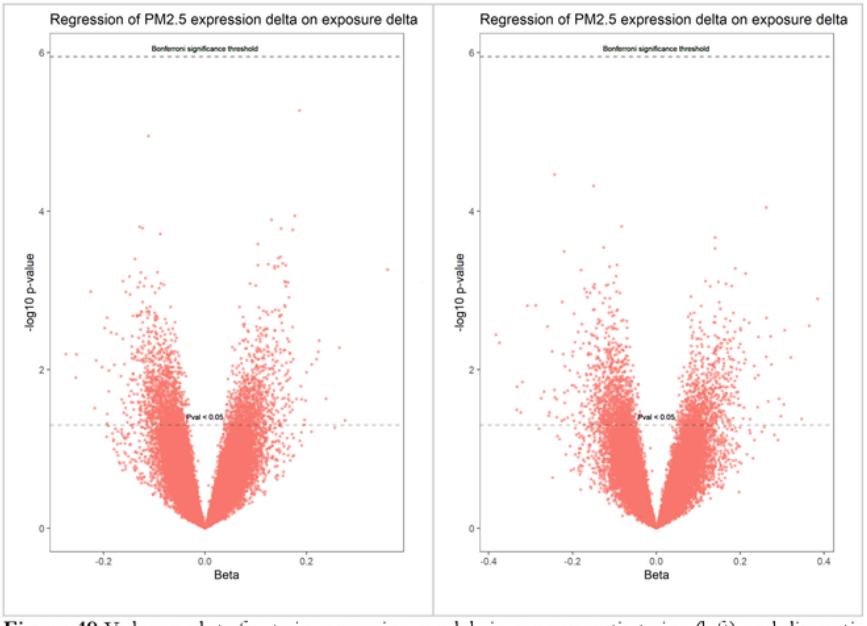

Figure 49 Volcano plots for twin regression models in monozygotic twins (left) and dizygotic twins (right). No significant results were observed.

DO TWINS 'CLUSTER'?

Monozygotic twin pairs were compared to dizygotic twin pairs, with the hypothesis that monozygotic twin pairs should be more likely to be grouped in the same cluster than dizygotic twin pairs.

A form of hypergeometric test for this hypothesis was conducted, with the null hypothesis that twins do not group within clusters. The expected distribution of twin-pairings within clusters under the null was modelled by repeatedly randomly shuffling the assignment of participants to clusters and counting the number of twin pairings within clusters. This was repeated 50,000 times for each clustering dimension and the resulting distributions plotted and tested for normality using the Shapiro-Wilk normality test. The Z-score and p-value for the true number of twin-pairings within clusters was calculated for each cluster dimension, giving a measurement of how the actual number of same-cluster twin pairs compared to the null prediction, and the Z-score for monozygotic twins compared to dizygotic twins.

Results

Table 21 shows how the 842 monozygotic (identical) and 424 dizygotic (non-identical) twins are paired or separated within the clusters. The Z-score indicates the number of standard deviations from the expected value.

Table 21 Hypergeometric test results for analysis of twin pairings within clusters. Results show that monozygotic twins are more likely to appear in the same geographical and exposure clusters than dizygotic twins

Grouping variable	No. clusters / categories	No. of twin pairs in same cluster		Expected no. of twin pairs in same cluster		Z-score		P-value	
		MZG	DZG	MZG	DZG	MZG	DZG	MZG	DZG
Traps Cluster	2	344	169	273.1	132.5	9.871	6.683	0	1.17E-11
Geography Cluster	3	245	135	167.3	82.5	8.126	7.737	2.22E-16	5.11E-15
Traffic Cluster	2	357	180	332.8	162.6	5.686	5.145	6.49E-09	1.34E-07
Demographic Cluster	2	319	162	246.7	119.2	8.467	6.676	0	1.23E-11
Biology Cluster	3	269	103	148.8	73.9	12.503	4.270	0	9.77E-06
Transcript cluster	7	133	61	75.6	37.7	7.563	4.339	1.98E-14	7.14E-06
Urbanicity	5	202	108	87.5	43.0	13.792	11.247	0	0
Smoking status	3	299	128	177.8	90.0	13.882	6.189	0	3.02E-10

Z-scores are higher in monozygotic twins than dizygotic twins in all dimensions. In other words, monozygotic twins are more likely to end up in the same cluster than dizygotic twins. This is to be expected in biology- or transcript-based clustering, for example. But for clusters based on geography, TRAP exposure, traffic and demographics, the increased grouping of monozygotic twin pairs relative to dizygotic is an indicator of the role that genetic profile has in shaping our choice of lifestyle and environment. The same effect is visible in urbanicity groupings and, especially, smoking status.